MACHINE LEARNING DROPOUT PREDICTIONS FOR PERSONALIZING DIGITAL MENTAL HEALTH INTERVENTIONS

Doctoral Thesis submitted to the Faculty of Management and Technology
of Leuphana University Lüneburg for the degree
Doctor of Natural Sciences
– Dr. rer. nat –

submitted by Kirsten Zantvoort
born 6th, February 1992 in Bremen, Germany

Submitted on 11th, September 2024
Oral defense (Disputation) on 10th, January 2025

First Supervisor: Prof. Dr. Burkhardt Funk, Leuphana Universität, Lüneburg, Germany
Second Supervisor: Prof. Dr. Viktor Kaldo, Karolinska Institutet, Stockholm, Sweden
Third Reviewer: Prof. Dr. Annet Kleiboer, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

The individual papers have been or will be published as follows:

Zantvoort, K., Nacke, B., Görlich, D., Hornstein, S., Jacobi, C., & Funk, B. (2024). Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. Npj Digital Medicine, 7(1), 1–10. https://doi.org/10.1038/s41746-024-01360-w

Zantvoort, K., Hentati Isacsson, N., Funk, B., & Kaldo, V. (2024). Data set size vs homogeneity – A Machine Learning study on pooling intervention data in E-Mental Health dropout predictions. SAGE Digital Health, 10, 1–11. https://doi.org/10.1177/20552076241248920

Zantvoort, K., Matthiesen, J. J., Bjurner, P., Bendix, M., Funk, B., & Kaldo, V. (Preprint). The Promise and Challenges of Computer Mouse Trajectories in DMHIs – A Feasibility Study on Pre-Treatment Dropout Predictions.

Zantvoort, K., Scharfenberger, J., Boß, L., Lehr, D., & Funk, B. (2023). Finding the Best Match—A Case Study on the (Text-)Feature and Model Choice in Digital Mental Health Interventions. Journal of Healthcare Informatics Research, 7(4), 447–479. https://doi.org/10.1007/s41666-023-00148-z

Zantvoort, K., Bjurner, P., Forsell, E., Wallert, J., Funk, B., & Kaldo, V. (working paper) Opening the black box – Effects of decision transparency on therapists' trust in and intended use of AI-based decision support systems in ICBTs

Hornstein, S., Zantvoort, K., Ulrike, L., Funk, B., & Kevin Hilbert. (2023). Personalization Strategies in Digital Mental Health Interventions: A Systematic Review and Conceptual Framework for Depressive Symptoms. Front. Digit. Health, 5. https://doi.org/10.3389/fdgth.2023.1170002

Year of Publication 2025

ABSTRACT

With the global need for psychological help long exceeding the supply, finding ways of increasing and better allocating mental health support is paramount. Digital Mental Health Interventions (DMHIs) are an effective way of complementing current mental healthcare measures. However, their effectiveness depends on engagement with the interventions and users prematurely dropping out are a significant problem across DMHIs. While it is known that measures such as human guidance can lower dropout, it remains unclear which user would most benefit from such measures. With health staff's time being scarce and costly, the question arises of how to best allocate it. Machine Learning (ML) models are a powerful tool for predicting individual behaviour and have the potential to identify those patients most in need of support. However, translating ML methods into the field of DMHIs has presented challenges, as 1) datasets tend to be small, 2) the gathered data is noisy, and 3) little insight into how to translate predictions into care exists. This thesis addresses these problems by systematically exploring these hurdles across six papers. In cooperation with clinical psychologists from three universities, a total of >11.000 individual users' data are analysed in four Machine Learning papers. Further, a randomised trial investigates therapists' opinions on explainable predictions, and a systematic review defines and quantifies the prevalence of personalization in DMHIs. The contributions 1) quantify and address the limitations of small data sets in DMHIs, 2) provide proofs-of-concept for common and innovative data types and model combinations, and 3) explore ways to increase trust and actionability of predictions to facilitate the adaption of ML-based Decision Support Tools in DMHIs. Beyond that, this PhD contributes to the very limited body of research on intervention dropout predictions by analysing a variety of large datasets and, hence, advancing a highly relevant academic and practical topic.

DECLARATION OF AUTHORSHIP

- This work was done wholly while in candidature for a research degree at this university and no part of this thesis has previously been submitted for a degree any other than the qualification at this university.
- Where published work of others has been consulted, it is clearly attributed. Where work of others has been cited, the source is always given. With the exception of such quotations, this work is entirely my own work.
- Where the thesis is based on work done in collaboration with others, it has been made clear what I have contributed myself.

ACKNOWLEDGEMENTS

To my husband, Nils: The last four years have challenged me in ways I could never have anticipated, and you were right there with me at every step – Thank you for being so amazing.

Now, I should probably thank my parents for raising me to believe I can do anything I set my mind to - But look at all the trouble that has gotten me into. I do, however, thank them for supporting me unconditionally through it all. I also want to thank my sisters, Stephanie, Eline, and Arlette, for, each in their own way, being a great inspiration to me. Further, I want to thank the rest of my (chosen) family and friends, especially the Feldmanns, Oma and Kristín — I am so fortunate to have you in my life.

Next, I want to express my deep gratitude to my first supervisor, Burkhardt Funk, for giving me the opportunity to pursue this thesis and for shaping my understanding of academic work. I have greatly enjoyed the possibility of following in your footsteps with leveraging Data Science to do something that really matters. Further, I want to thank my second supervisor, Viktor Kaldo, for his patience in explaining the world of psychological interventions and for allowing me to benefit from the incredible knowledge and data sets in his research group. I have greatly enjoyed my time in Stockholm and the ability to work with and learn from everyone there. I also extend my gratitude to Annet Kleiboer, my third evaluator, for the time and expertise she dedicated to reviewing my work. Further, I want to thank Madlen Schmaltz for her help with all the administrative questions – You are greatly appreciated.

Next, I want to express my deepest gratitude to all my coauthors – thank you for everything I was able to learn from and achieve with you. From this group, I want to specifically thank Silvan Hornstein: Your unbreakable optimism and advice pushed me through the most challenging moments in academia and most definitely made this PhD so much more fun. Likewise, I want to thank my fellow PhD students at Leuphana and Karolinska Institutet: Learning side by side has made all the difference. Thank you also to Jenny, for helping me design the cover of this PhD.

Finally, I would like to thank all the people and institutions involved in gathering the data sets I had the privilege to work with. Your time, effort, and resources made this thesis possible, and I am deeply appreciative of your work aiming at helping people to improve their mental health.

TABLE OF CONTENT

Abstract i

Declaration of Authorship ii

List of Figures viii

List of Tables ix

Acknowledgements iii

Chapter I: Introduction 1

1 Key Concepts 2

1.1 Intervention Dropout 2

1.2 Predictive Modelling 3

2 Related Work and Research Questions 4

2.1 RQ1: Small Data Set Sizes 5

2.2 RQ2: Feature and Model Combinations 7

2.3 RQ3: User Adaption Hurdles 8

3 Included Articles 9

4 Discussion 12

4.1 RQ 1: Small Data Set Sizes 12

4.2 RQ 2: Feature and Model Combinations 13

4.3 RQ 3: User Adaption Hurdles 16

4.4 Open Questions and Outlook 17

5 Conclusion 18

Chapter II: Learning Curves 21

1 Introduction 22

2 Results 23

2.1 Final Values 23

2.2 Predictive Power of Feature Groups 24

2.3 Overfitting on Small Data Set Sizes 25

2.4 Variance of Results 26

2.5 Performance Convergence per Model 27

2.6 Marginal Value and Convergence of Additional Features 29

2.7 Model and Feature Combinations 29

3 Discussion 30

4 Methods 33

4.1 Case Study Background – everyBody Study 33

4.2 Definition of Outcome 34
4.3 Feature Groups and Pre-Processing 34
4.4 Algorithms 37
4.5 Learning Curves and Training Set-up 37
4.6 Evaluation and Result Analysis 38
Chapter III: Intervention Data Pooling 39
1 Introduction 40
2 Methods 41
2.1 Interventions and Participants 41
2.2 Features 42
2.3 Exploration of Heterogeneity Between Interventions 43
2.4 Dropout Prediction 44
3 Results 45
3.1 Data Heterogeneity 45
3.2 Prediction 47
4 Discussion 50
5 Limitations 51
6 Conclusion 51
Chapter IV: Computer Mouse Trajectory 53
1 Introduction 54
2 Mouse Trajectory Data 55
2.1 Gathering Mouse Trajectories 55
2.2 Processing of Mouse Data 56
2.3 Data Science Methodology 57
3 Case Study 58
3.1 Interventions and Participants 58
3.2 Dropout Definition 58
3.3 Baseline Data 59
3.4 Mouse Data 59
3.5 Prediction Models and Experiment Set-up 59
4 Results 60
4.1 Final Data Set 60
4.2 Prediction Results 61
5 Discussion 61
Chapter V: Intervention Text Features 63
1 Introduction 64
2 Background 66

2.1 Pre-Intervention Text Data 66
2.2 Intervention Text Data 67
3 Study Set-Up 70
3.1 Data Description 70
3.2 Non-Text Data 72
4 Text Representation 73
4.1 Metadata 73
4.2 Bag-of-words 74
4.3 Embeddings 75
5 Machine Learning Models 75
5.1 ML Models for Non-Sequential Data 76
5.2 Recurrent Neural Network 77
5.3 BERT 78
6 Results 79
6.1 Intervention Failure 79
6.2 Dropout 80
6.3 Discussion of Clinical Usefulness 82
7 Conclusion 86
Chapter VI: Explainable AI and Trust 89
1 Introduction 90
2 Hypotheses 91
2.1 Understandability 92
2.2 Trust 92
2.3 Clinical Usefulness and Actionability 93
3 Methodology 93
3.1 Predictions 93
3.2 Patient Case Selection 94
3.3 Case Presentation 94
3.4 Randomisation 95
3.5 Participants Screening and Inclusion 96
3.6 Questionnaires 96
3.7 Statistical Analysis 98
4 Results 98
4.1 Final Cohort 98
4.2 Hypotheses Testing 99
4.3 Other Responses 101
5 Discussion 101
5.1 Evaluation of Hypotheses 102

5.2 Further Insights and Limitations 103
Ethical Approval 104
Chapter VII: Personalization Strategies 105
1 Introduction 106
2 Methods 107
2.1 Search Strategy 108
2.2 Selection Criteria 108
2.3 Selection Procedure 108
2.4 Development of the Conceptual Framework 108
2.5 Data Extraction 109
3 Results 110
3.1 Study Selection 110
3.2 Intervention and Study Characteristics 111
3.3 Conceptual Framework 111
3.4 Results on Personalization 113
3.5 Use of Automated Decisions for Personalization 114
3.6 Empirical Comparison of More and Less Personalized Interventions 114
4 Discussion 115
Bibliography 119
Appendix incl. published version of articles 143

LIST OF FIGURES

Figure 1: Related work's data set sizes and feature types per publication year 5
Figure 2: Histogram of 10.000 area under the curve scores per test data set size 6
Figure 3: Training and test learning curves per feature type 25
Figure 4: Result variance per feature type 27
Figure 5: Test learning curve and convergence points per model type 28
Figure 6: Random Forest training (CV) and test learning curve 29
Figure 7: Dropout curves per intervention arm with cutoff 34
Figure 8: Distribution of patients per intervention and cluster 46
Figure 9: Training and test data sample per pooled run 47
Figure 10: Balanced accuracy for training and test results 48
Figure 11: Example of mouse data representation of a single user on interface 56
Figure 12: Raw mouse trajectories to feature examples 57
Figure 13: Architecture of the 1-dimensional convolutional neural network 60
Figure 14: Task-specific LSTM-based model architecture 78
Figure 15: Histogram failure prediction output 85
Figure 16: Decision Support Tool interface example case for both conditions 95
Figure 17: Overview of trial set-up in dependence of the SHAP condition. 97
Figure 18: Overview test strategy 98
Figure 19: Histogram quiz errors 101
Figure 20: PRISMA flowchart of study inclusion 110
Figure 21: Delineation of personalization from other forms of adaptation. 111
Figure 22: Dimensions of personalization 112
Figure 23: Mechanisms of personalization 112
Figure 24: Number of studies per personalization type and mechanism 113

LIST OF TABLES

Table 1: Overview paper characteristics and contributions 10
Table 2: Overview area under the curve (AUC) per feature type 14
Table 3: Overview feature groups 36
Table 4: Descriptive statistics for a) baseline data and b) mouse data set 61
Table 5: Overview of the intervention studies included in this analysis. 71
Table 6: Result table intervention failure prediction 80
Table 7: Result table dropout prediction 82
Table 8: Cohort descriptives 99
Table 9: Analysis results with descriptive and p-values 100
Table 10: Personalization mechanisms per dimension in the intervention 114

Chapter I: Introduction

Mental disorders are a highly prevalent global health problem (Richter et al., 2019; Santomauro et al., 2021). Beyond their significant impact on the ones suffering from them, mental disorders put immense pressure on health and social care systems (Arias et al., 2022). While efficient treatments exist for many disorders, estimates are that in high-income countries, merely a third of those impacted receive the help they need (Rommel et al., 2017; Wang et al., 2005). With the demand for help continuously increasing, easing the tension on this overwhelmed system is paramount for individuals and societies as a whole (Andersson et al., 2019; Richter et al., 2019).

Digital Mental Health Interventions (DMHIs) complement current treatment options due to their accessibility, scalability, and personalizable design. DMHIs, including Internet-based Cognitive Behaviour Therapy (ICBT), follow similar goals as traditional face-to-face therapy but centre around time- and location-independent self-help via digital means. Many of them are supplemented by different levels of human guidance, such as via messages or phone calls (Andersson et al., 2019; Haller et al., 2023). Meta-analyses demonstrate their efficacy in treating mental health problems like depression (Karyotaki et al., 2021; Reins et al., 2019), anxiety (Andrews et al., 2018), and eating disorders (Linardon et al., 2020). At the same time, the amount of completed treatment is decisive for health outcome success (Donkin et al., 2013; Gan et al., 2021; Haller et al., 2023; Lipschitz et al., 2023). However, up to 50-80% of users drop out before completing the intervention (Lipschitz et al., 2023; Richards & Richardson, 2012). Even though dropout can be lowered through targeted measures such as human guidance (Baumeister et al., 2014; Gan et al., 2022; Hilvert-Bruce et al., 2012; Lipschitz et al., 2023), that comes at high marginal costs (Forsell et al., 2019; Lutz et al., 2022). In the current situation, adding or increasing human guidance for all is infeasible and inefficient (Forsell et al., 2020; Lutz et al., 2022). This trade-off raises the question of how participants needing additional measures can be identified to keep them in the intervention and maximize overall health outcomes.

The field of supervised Machine Learning (ML) has been dedicated to predicting future outcomes from various input features for more than 70 years (Breiman, 2001; Bzdok & Meyer-Lindenberg, 2018). Despite the continuous permeation of ML in many fields, psychological and psychiatric applications have been very limited (Eloranta & Boman, 2022). In contrast to face-to-face therapy, DMHIs' medium allows the efficient, scalable, and automated recording of treatment engagement and psychological data. Nevertheless, of the papers that apply ML in mental health settings, only 1% deal with the actual time during therapy, according to Shatte et al. (2018). One of the critical challenges proposed to explain these low numbers is the persistent lack of large datasets (Aafjes-van Doorn et al., 2021; Bzdok & Meyer-Lindenberg, 2018; DeMasi et al., 2017; Sajjadian et al., 2021; Squires et al., 2023). Shatte et al.'s (2018) analysis also showed that newer model and feature approaches, such as deep learning or text data, are strongly underrepresented. Although further studies have been published since, there is little doubt about the considerable remaining

research potential in this area (Chekroud et al., 2021; Lutz et al., 2023; Squires et al., 2023). Further, the studies that exist are proof-of-concepts and only very few discuss the implementation of ML-based Decision Support Tools (Cruz Rivera et al., 2023; Higgins et al., 2023; Triantafyllidis & Tsanas, 2019; Yin et al., 2021). Aiming to progress the effective use of dropout predictions in DMHIs, this thesis, therefore, investigates the following research questions:

RQ 1: How does a) data set size influence the results in intervention dropout prediction, and how can b) the resulting problems be addressed?

RQ 2: What combinations of features and models best predict dropout?

RQ 3: How can the user adaption of ML predictions in DMHIs be facilitated?

This thesis focuses on intervention dropout instead of health outcome predictions for two main reasons: Firstly, the absence of activity can unequivocally be categorised as dropout. In contrast, in outcome predictions, missing target variables have to be imputed or lead to the exclusion of patients (Prasad et al., 2023; Spineli et al., 2015). Thus, dropout predictions circumvent the challenge of missing target values and offer actionable insights into a user group suspected to be most in need of additional resources (Eysenbach, 2005; Lipschitz et al., 2023). Secondly, with first meta-studies published (Lee et al., 2018; Sajjadian et al., 2021; Vieira et al., 2022), research on ML outcome predictions is further advanced (Chekroud et al., 2021). On the contrary, despite Eysenbach's (2005) landmark paper soon passing its 20th anniversary, dropout – and its continuous inverse analogous, adherence – in DMHIs remain to be understudied and underreported (Beintner, Vollert, et al., 2019; Lipschitz et al., 2023).

Thus, this PhD aims to investigate how to best use Machine Learning models to predict and ultimately influence intervention dropout. To this end, the following section motivates the key concepts of intervention dropout and predictive modelling. Afterwards, the current state of research on dropout predictions in DMHIs is succinctly outlined to derive the research questions from the identified gaps. The last section discusses the contributions of this thesis and gives an outlook on the remaining open questions and possible next steps. The chapters following this introduction are the six individual studies contributing to this thesis.

1 Key Concepts

This section introduces the key concepts and working definitions used in this thesis, primarily based on Eysenbach (2005) and Beintner, Vollert et al. (2019), as well as Yarkoni & Westfall (2017), and Breiman (2001).

1.1 Intervention Dropout

In terms of the operationalisation of dropout in DMHIs, not starting is differentiated from discontinuing the intervention after start (intervention dropout), and not responding to study assessments (study dropout) (Beintner, Vollert, et al., 2019; Eysenbach, 2005; Lipschitz et al., 2023). This PhD focuses on intervention dropout, as the most promising concept for data-driven adaptation of ongoing treatment.

Commonly, intervention discontinuation is turned into binary dropout by considering the intervention completed as of a certain minimum (Bremer et al., 2020; Cote-Allard et al., 2022). As discussed by Beintner, Vollert et al. (2019), the proposal of minimal content considered necessary to benefit so far lacks standards, resulting in a research question itself. In this thesis, dropout operationalisations are based on intervention-specific clinical criteria combined with pragmatic concerns such as class balance. Grounded in the findings and subsequent suggestions of Donkin et al. (2013) and Gan et al. (2021), all minima are defined in the number of completed modules. The core modules deemed necessary to improve are defined by the clinical experts who developed and evaluated the respective intervention. At the same time, some patients may drop out because they require less than the average content – so-called *good leavers*. Allocating more resources to these patients would be detrimental to resource efficiency. Thus, even when focusing on intervention dropout and not outcome, neither concept can be assessed without the other (Gan et al., 2021). Outcomes will, therefore, be accounted for where available and applicable.

In conclusion, in the following, dropout refers to the start and subsequent discontinuation of an intervention before completing the best estimate of the minimal content considered necessary for users to benefit.

1.2 Predictive Modelling

Statistical modelling aims to derive generalisable rules from past problems to solve future ones. Within this field, the classical statistical (CS) approach differs from the algorithmic or Machine Learning (ML) approach in its goals, requirements, and outputs (Breiman, 2001).

Given a specific problem, the CS approach seeks to mechanistically mimic the data generation process. To this end, theory-driven hypotheses result in the postulation of a humanly understandable model whose goodness-of-fit is evaluated on a data sample often collected for that very purpose. CS approaches attempt to ensure generalisability through methodological rigour but seldom test for it explicitly. Ultimately, this process yields cohort-based insights to be interpreted; hence, it aims to explain (Yarkoni & Westfall, 2017).

In contrast, the ML approach places limited emphasis on causal hypotheses. Instead, the goal is to accurately predict the future by agnostically connecting the input to the output data (Eloranta & Boman, 2022; Yarkoni & Westfall, 2017). Instead of explaining, it provides the possibility to inform individual decisions. With no intent to depict causality, equifinality and pluralism are central to Data Science epistemology. Consequently, widely varying model types are typically tried and tested to determine which yields the best prediction performance.

Prediction performance varies depending on model complexity and the subsequent bias-variance trade-off, which is influenced by the form and number of parameters and features. For example, linear regression is simple and understandable but may miss non-linear relationships. More flexible models, however, can better approximate

realities' complexity but are prone to mimicking noise in the data, hence overfitting (Hastie et al., 2017; Yarkoni & Westfall, 2017).

Just as in CS, a larger number of features increases the risk of overfitting in ML. However, to reduce theory dependence and subsequent human bias, features are (semi-)automatically selected by, again, exploring and comparing different options (Yarkoni & Westfall, 2017). Deep Learning, a newer subcategory of ML, advances this even further as neural networks automatically extract features from unstructured data with limited to no human involvement (Piccialli et al., 2021; Shrestha & Mahmood, 2019).

To select the best model, including features, ML models are trained on part of the data and evaluated on a previously set-aside test set. If the model is overfitted, it is not expected to generalise on the test set. If it is underfitted, it is expected to be outperformed by more sophisticated models. While this approach allows an empirical evaluation of different models, it comes with the downside that only a fraction of the data set is available for evaluation. So-called cross validation (CV) splits the data set in k equal folds and increases power by evaluating a model on each. CV is often combined with grid-search, a systematic evaluation of hyperparameters, for example, for feature regularisation. However, choosing the hyperparameters in the same step as evaluating the model still leads to overfitting, thus necessitating separation (Bates et al., 2022; Cabitza & Campagner, 2021).

Predictions can be evaluated on balanced accuracy (BACC), where the accuracy of both classes is weighted equally regardless of the ratio in the data. However, this requires setting a threshold to decide as of when a prediction becomes a dropout or completer. While this makes sense when implementing a model, different thresholds come with context-dependent up- and downsides that cannot always be fully overseen at the time of the decision. The area under the curve (AUC), on the other hand, evaluates the model's ability to differentiate classes across all possible thresholds. Regarding the evaluation of AUCs, Kraemer et al.'s (2003) propose to differentiate between no (0.50-0-0.56 AUC), low (0.57-0.64), moderate (0.65-0.70), good (0.71-0.75) and very good (>0.75) predictive power.

In summary, the Data Science approach is inductive in nature and data- instead of theory-driven. It optimises predictive accuracy by iterating over options and testing their generalisability on an unseen test set. Evaluating on only a fraction of data and approximating complex relationships requires much larger data than CS approaches. In the end, ML infers an often-opaque decision logic to produce the outcome from a set of features and yields individual predictions.

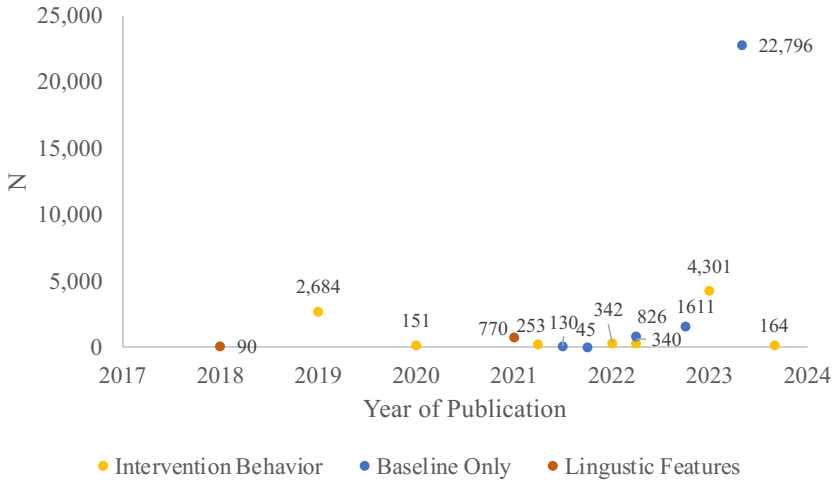
2 Related Work and Research Questions

Research papers using ML methods to predict dropout or adherence are limited but promising (Chekroud et al., 2021). To ensure an up-to-date and exhaustive related work section, the papers discovered during this PhD were supplemented through a Pub Med search on 26.06.24 (("Machine learning" OR "ML" OR "artificial intelligence" OR "AI" OR "predict*") AND ("dropout" OR "adherence" OR "attrition") AND ("mental" OR "psycho*" OR "psychia*") AND ("Internet-based" OR "ICBT" OR

"DMHI")). From the resulting papers, those were included that predict DMHI dropout or adherence using an ML approach according to the definitions introduced above.

The number of studies predicting dropout for DMHIs has greatly increased from three at the start of this PhD in 2020 (Bremer et al., 2020; Pedersen et al., 2019; Wallert et al., 2018) to twelve (Figure 1). Yet, as demonstrated in this section, the research gaps and challenges remained largely unsolved. Therefore, in each of the following subsections (Chapter I:2.1-2.3), the applicable characteristics, strengths, and limitations of these twelve related works are synthesised and expanded upon to motivate the respective research question (RQ1-3). A table with the overview of the discussed details per study is provided in the Appendix.

Figure 1: Related work's data set sizes and feature types per publication year



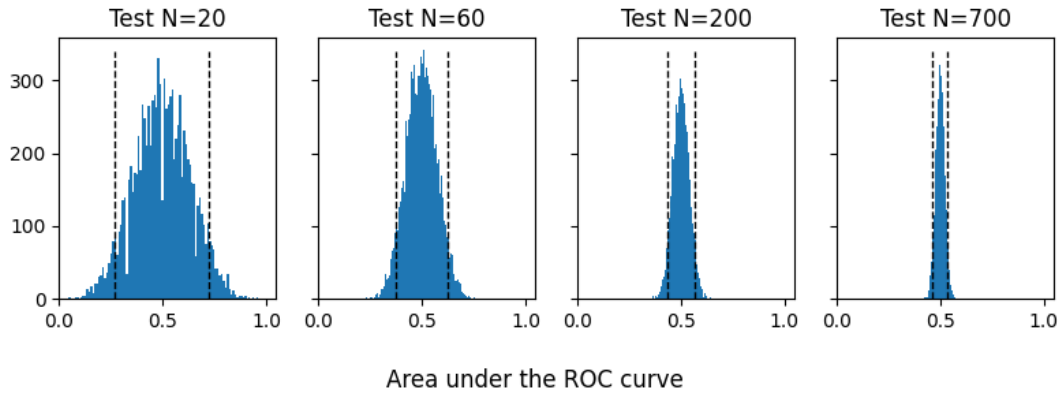
2.1 RQ1: Small Data Set Sizes

As explained in Chapter I:1.2, the Data Science paradigm proposes to shift from theory-rich and data-poor to theory-poor and data-rich scientific discoveries (Yarkoni & Westfall, 2017). However, large datasets are a prerequisite hard to satisfy in healthcare settings where data are costly (Pasini, 2015). Consequently, the median data set size of the twelve related works is 342, with a minimum of 45 (Kim et al., 2021) and a maximum of 22,796 (Günther et al., 2023) users. As Figure 1 indicates, large DMHI datasets remain uncommon (Cote-Allard et al., 2022; Ebert et al., 2019). Consequently, many related work studies specifically name their small data set sizes as problematic in the limitation section (Bremer et al., 2020; Kim et al., 2021; Linardon et al., 2022; Wallert et al., 2018). Unstable, overfitted, and thus not generalisable results are an inherent risk of small datasets that remain seldomly accounted for in clinical settings (Bates et al., 2022; Cabitza & Campagner, 2021; DeMasi et al., 2017).

To illustrate the problem of result instability on small data sets, the plots in Figure 2 (based on Regenthal, 2022) show the variation of prediction scores on balanced data sets sized 100, 300, 1000, and 3,500. Assuming a 20% test set, this corresponds to test set sizes of 20, 60, 200, and 700, respectively. The labels and predictions are generated randomly; hence, the supposed model has no predictive power and should produce an

AUC of 0.5. This simulation was repeated 10.000 times per test set size to infer the likelihood of producing any certain value.

Figure 2: Histogram of 10.000 area under the curve scores per test data set size



Having a test set of 20 data points (N=100) implied an immensely high variance with the 95%-confidence interval (CI) of 0.27 and 0.73 AUC. At a maximum AUC of 1 and almost a 4th of values above a good result of 0.6, accepting this model is not unlikely despite its uselessness. With a data set close to the median of related work (N=300), the maximum is still at 0.79, and almost 10% of AUCs are above 0.60 (CI: 0.37-0.62). At 200 test data points (N=1,000), this uncertainty drops significantly (CI=0.43-0.57) and is further narrowed down (CI=0.46-0.54) when increasing the test set size to 700 (N=3,500).

Regarding overfitting, single cross-validation (CV) or train-test split results are often reported when dealing with small data (Sajjadian et al., 2021). Such practices have repeatedly proven to lead to inflationary results (Bates et al., 2022; Hastie et al., 2017; Lateh et al., 2017). Insufficient data samples make it likely for models to focus on meaningless noise. As a result, especially more flexible models produce high training performances but fail to reproduce them on larger independent test sets (Atla et al., 2011; Rodriguez-Galiano & Chica-Rivas, 2014; Saseendran et al., 2019). Beyond overfitting, neural networks and many other ML methods require large data sets to perform well (Alzubaidi et al., 2023; Bzdok & Meyer-Lindenberg, 2018; Piccialli et al., 2021). Even for simpler models such as Logistic Regression, data set sizes can have a large influence on performance, dependent on the nature and number of predictors used (Smeden et al., 2019). In addition, prediction performance converges at a certain data set size, meaning that the marginal benefit of adding more data points is not constant, even for a single model (Perlich et al., 2004).

Despite their undebated relevance, which data set sizes or approaches are best suited to mitigate these problems remains unclear for ML research in DMHIs. Considering their decisive importance for the integrity and value contribution of this field of research, the first research question is, therefore:

RQ 1: How does a) data set size influence the results in intervention dropout prediction, and how can b) the resulting problems be addressed?

2.2 RQ2: Feature and Model Combinations

Beyond the question of how many data points are collected, the related works differ in the nature and number of features and models they are paired with. Regarding features, self-reported questionnaire answers are commonly gathered as baseline data before treatment start and often comprise socio-demographic and clinical variables. They can further be collected as treatment goes on, as often seen for symptom scores (Forsell et al., 2020). As the intervention starts, user intervention behaviour data in the form of log files become available. These automatically collected files usually comprise timestamps, user identifiers, and event types and can be transformed into a myriad of features (Bremer et al., 2020). Further, DMHIs regularly include asynchronous text-driven communication with participants, generally involving open-text intervention exercises and communication with e-coaches (Calvo et al., 2017).

Three of the twelve related works, including the one with the largest data set, only had pre-treatment questionnaire features available (Gonzalez Salas Duhne et al., 2022; Günther et al., 2023; Kim et al., 2021) as displayed in Figure 1. They all reported low (Kraemer et al., 2003) predictive power (AUC=0.57, R^2=0.42). One related work study focused on only intervention behaviour data with much higher results (BACC up to 0.79) (Cote-Allard et al., 2022). Most other related works (N=6) investigated socio-demographic, clinical, and user behaviour data in combinations with moderate to very good results (AUCs=0.60-0.94/ BACC=0.61-0.73) (Bremer et al., 2020; Bricker et al., 2023; Linardon et al., 2022; Linnet et al., 2023; Moshe et al., 2022; Pedersen et al., 2019). However, the three settings achieving an AUC of more than 0.90 had no or very little time between the point of prediction and dropout (Bremer et al., 2020; Bricker et al., 2023; Pedersen et al., 2019). Hence, the time left to intervene is an important aspect when evaluating the usefulness of features. The remaining two studies attempted to combine extracting information from text data through simple meta- and dictionary-based methods with limited success (BACC=0.54, accuracy=0.64 at 52% dropouts) (Smink et al., 2021; Wallert et al., 2018).

As explained in Chapter I:1.2, ML models can efficiently process a high number of features, and automated feature selection is a cornerstone of Data Science methodology (Bzdok & Meyer-Lindenberg, 2018; Yarkoni & Westfall, 2017). At the same time, data protection concerns require data scientists to limit themselves to the minimally necessary features, especially concerning sensitive health data (Cote-Allard et al., 2022). For the related works, the number of features used was not always reliably derivable from the manuscript but is estimated to be between seven (Bricker et al., 2023) and 401 (Bremer et al., 2020). The latter study concluded that their 25 hand-crafted features were the most predictive ones (Bremer et al., 2020), which is in line with findings on outcome predictions (Hentati Isacsson et al., 2023). Accordingly, authors such as Forsell et al. (2019), Côte-Allard et al. (2022), and Bricker et al. (2023) determinately argue for a focus on a limited number of features.

Whenever evaluating the predictive power of a group of features, the type of model used to extract it has to be considered. As introduced in section Chapter I:1.2, algorithms differ in their ability to handle feature characteristics such as non-linearity,

interaction effects, multicollinearity, sparseness, or noisiness (Atla et al., 2011; DeMasi et al., 2017; Kwon & Sim, 2013; Nettleton et al., 2010; Rodriguez-Galiano & Chica-Rivas, 2014; Saseendran et al., 2019). Mismatching, for example, a complex non-linear feature type with a too simple model or overfitting a flexible model type on noisy data will lead to inferior results.

In the related work, the by far most commonly implemented models were Logistic Regression (N=8 of 12), Random Forest models (N=5), followed by different tree-based boosting models (N=3) and Support Vector Machines (N=2). To feed data into these non-sequential models, the features were aggregated (e.g., means or sums) or all kept as independent features. Neural networks leveraging the time-dependence of features were only used twice; Cote-Allard et al. (2022) implemented a Self-attention network for sequential login data without benchmarking it to any other model type. Smink et al. (2021) compared a task-specific neural network fed with one email's features at a time with several non-sequential models and concluded that neither approach was superior.

In summary, so far, no successful DMHI predictions before treatment start have been published, the marginal value of adding further features is unclear, and the focus of related work is on simple feature and model types. These three prospects are gathered under the second research question:

RQ 2: What combinations of features and models best predict dropout?

2.3 RQ3: User Adaption Hurdles

So far, the focus has been on optimising dropout prediction accuracy, including generalisability. However, in order to add value, ML predictions must be implemented in practice, commonly done through so-called Decision Support Tools (DSTs) (Jacobs et al., 2021).

Yet, none of the twelve related works implemented, let alone prospectively evaluated, the clinical use of their ML model. One study argues that their RF model was too expensive to implement and provided their staff with a simple activity index instead (Pedersen et al., 2019). While the impact was not measured methodologically, they report dropout to have lowered from 54% (N=2,684) to 19% (N=6,402) after this addition. The other eleven studies discuss translating the predictions into care very briefly or not at all. The most named use case motivating the studies in the introduction is adapting human support levels (Bricker et al., 2023; Günther et al., 2023; Kim et al., 2021). However, most other mentions just refer to "*personalizing*" or "*tailoring*" treatment without further clarification of what this entails.

While other studies implement outcome predictions in DMHIs (Forsell et al., 2019) or face-to-face therapy (Lutz et al., 2022; Popescu et al., 2021), they are very few (Higgins et al., 2023; Triantafyllidis & Tsanas, 2019). This lack of adaption is attributed to the challenges of integrating predictions into the complex, time-constrained and user-driven healthcare processes (Devaraj et al., 2014; Jacobs et al., 2021; Khairat et al., 2018; Maslej et al., 2023). Ethical and legal constraints place human oversight and

accountability as a prerequisite for AI applications in health care, thus precluding automated decisions (Diprose et al., 2020; Duffourc & Gerke, 2023). As a result, predictions are offered to human decision-makers who retain the decision autonomy over if and how to use them. While good prediction accuracy is undoubtedly a prerequisite for confidence in such DSTs, it alone does not suffice to prompt usage. Instead, the users' lack of confidence is a critical hurdle in the adaption of ML-driven DSTs (Devaraj et al., 2014; Higgins et al., 2023; Jacobs et al., 2021; Maslej et al., 2023).

Several authors have postulated that decision intransparency causes a lack of trust in ML-based DSTs and subsequently lowers the behavioural intention to act on predictions (Devaraj et al., 2014; Gille et al., 2020; Higgins et al., 2023). As described in Chapter I:1.2, the strength of more sophisticated models is their ability to formalise complex non-linear relationships, often through advanced architecture and a myriad of parameters. It is precisely these characteristics that make it difficult, if not impossible, for humans to understand how they arrived at their prediction (Gille et al., 2020; Piccialli et al., 2021). But even if trust is established and clinicians want to use the DST, Jacobs et al. (2021) and Lutz et al. (2022) report that healthcare staff often do not know what to do with the predictions. In these cases, time constraints seem to force them to disregard the additional information from the DST and revert to previous decision patterns (Jacobs et al., 2021; Lutz et al., 2022). Further, Jacobs et al. (2021) and Tonekaboni et al. (2019) report that interviewed clinicians required evidence-based recommendations in order to be willing to act on them.

Consequently, even when finding a model with good prediction accuracy, user adaption through trust and actionability remains a decisive impediment in the adaption of DSTs in DMHIs. These aspects result in the last, more explorative and not only dropout-specific question:

RQ 3: How can the user adaption of ML predictions in DMHIs be facilitated?

3 Included Articles

This thesis investigates the three introduced research questions through six studies employing various methodologies. Four studies are ML use cases, each analysing data from one to five interventions. Their focus ranges from learning curves over data pooling to investigating newer feature types. The fifth study is a randomised trial examining how explainable AI affects therapists' trust in and intended use of an ML-based DST. The final study is a systematic review providing insights into the concept, use, and evaluation of personalization in DMHIs for depressive symptoms.

Table 1 offers an overview of the studies' content, characteristics, and findings. The first table column corresponds to the chapter number of this thesis, which is used for reference throughout this discussion and, hence, starts at II. Chapters VI and VII are not solely focused on dropout predictions but concern DSTs and ML applications in general.

Table 1: Overview paper characteristics and contributions

#	Paper and Status	JIF[1]/ H-index	Interventions	Content	Key Findings
II	Zantvoort et al., (Preprint). Predictive Power, Variance and Generalizability – A Machine Learning Case Study on Minimal Necessary Data Sets Sizes in Digital Mental Health Intervention Predictions In revision at *NPJ Digital Health* (submitted June 2024)	12.4 (2023) / 84	3,654 users from five eating disorder prevention DMHIs	Investigation of domain-specific learning curves depending on… • Dataset sizes between N=100-3,654 • Feature groups between 2-129 questionnaire and/ or intervention behaviour features • Model choice ranging from Naive Bayes to neural network models	RQ1 and RQ2: • Minimally necessary N=500-1,000 in DMHIs • N≤300 overestimate prediction performance • Complex features and models need more data to converge in performance • Intervention behaviour data from first week reliably predicts dropout (0.72-0.80 AUC)
III	Zantvoort et al., (2024). Dataset Size versus Homogeneity: A Machine Learning Study on Pooling Intervention Data in E-mental Health Dropout Predictions. *SAGE Digital Health*	3.9 (2022) / 30	6,418 patients from depression, social anxiety and panic disorder routine care ICBTs[2]	Investigation of pooling data from three different interventions to one, including… • Data exploration through clustering analysis • Investigation if and how pooling the interventions affects intervention dropout prediction results • Repetition across small, medium and large training data with evaluation on same test data (N=730)	RQ1: • Pooling interventions reduces overfitting and improves test results (avg. 0.77 AUC) • Results from N≤300 overestimate prediction performance, especially on single intervention data • Similar intervention behaviour patterns across depression, social anxiety and panic disorder patients
IV	Zantvoort et al., (Preprint). The Promise and Challenges of Computer Mouse Trajectories for Pre-Treatment Predictions in DMHIs. Under consideration at *Internet Interventions* (submitted Sep. 2024)	If accepted: 3.6 (2023) / 52	408 baseline/ 181 mouse data points from two depression, social anxiety and panic disorder ICBT RCTs	Introducing computer mouse trajectory data as novel data type in DMHI predictions, including… • Dissemination of tracker, code and methodology • Comparison of hand-crafted features with non-sequential models to time-spatial data in combination with a sequential neural network • Case study on pre-treatment dropout prediction	RQ2 (peripherally RQ1): • Computer mouse trajectories could be a novel data type available before and during treatment • Tracking and processing mouse trajectories is challenging but possible • In the exemplary case, little to no information value (0.50 and 0.58 AUC); however, possibly due to small data set size

V	Zantvoort et al., (2023), Finding the Best Match — a Case Study on the (Text-) Feature and Model Choice in Digital Mental Health Interventions. *Journal of Healthcare Informatics Research.*2024)	5.9 (2022) / 17	849 users from one stress DMHI across 6 RCTs	Analysing nearly 16,000 open-text answers to systematically compare the predictive power of… • Text representation methods (metadata, TF-IDF, sentiment and topic analysis, and word embeddings) in combination with supplementary numerical inputs (socio-demographic, evaluation, and closed-question data) • In combination with ML model types – ranging from linear to a task-specific LSTM-based deep learning and BERT model	RQ2 (peripherally RQ3): • Word Embeddings can produce competitive results (0.70AUC) if matched with adequate model, and BERT works as well (0.67 AUC) • However, performance parity of complex (0.70 AUC) and simple (0.69 AUC) setups for given data set size • Outcome and dropout predictions differ in findings • Detailed discussion of clinical value of the model
VI	Zantvoort et al., (Working Paper). Effects of SHAP Values for Decision Transparency on Therapists' Perception and Use of AI-based Decision Support Systems in ICBTs. (Preliminary results as of Aug. 2024)	Only final results will be published	35 therapists and 2,881 patients from Depression Routine Care ICBT[2]	• Randomised trial investigating influence of adding SHAP values to a Decision Support Tool predicting outcome • Measurements for understandability, trust, clinical usefulness, and actionability	RQ3: • Local SHAP values increase therapists' understanding of decisions, trust in and usefulness of DST • However, no change in intended treatment support level • Proposal and discussion of a user interface to present predictions to therapists
VII	Hornstein et al., (2023), Personalization Strategies in Digital Mental Health Interventions: A Systematic Review and Conceptual Framework for Depressive Symptoms. *Frontiers in Digital Health.*	3.2 (2023) / 21	138 studies on 94 distinct depressive symptom DMHIs serving 24k users	• Systematic review on DMHIs targeting depressive symptoms. Categorisation of personalization based on what is personalized (i.e., intervention content, content order, level of guidance or communication) and the underlying mechanism (i.e., user choice, provider choice, decision rules, and ML-based approaches). • Evaluation which of the 138 included studies personalize what/ how and if they evaluated the effects.	RQ3: • Framework to categorise personalization in DMHIs and delineate it from other adaption concepts • 66% of studied interventions personalize at least one dimension, most of them in a rule-based manner • Only three studies use ML, and only two comparatively evaluate the value of personalization.

[1] *Web of Science Journal Citation reports at time of submission or latest available score and Scimago Journal Ranking H-score*

4 Discussion

Assuming timely publication, 25% of studies on dropout predictions in DMHIs will originate from this thesis. Including these studies increases the median dataset size from 342 to 770, introduces three additional feature types, and doubles the number of studies utilising sequential neural networks. In total, the four ML studies analyse data from 11,329 distinct patients across ten DMHIs and include the second and fourth largest data sets among related work up to date. Additionally, the randomised trial contributes the first quantitative evidence for the role decision transparency plays for clinicians in DMHIs. Lastly, the comprehensive evaluation of 138 studies involving 92 DMHIs for depressive symptoms contributes reliable evidence on the use of personalization in DMHIs among one of the most prevalent mental disorders.

A key strength of this thesis is its stringent and systematic investigation of a singular, complex and relevant problem. Exploring the research questions with different foci across articles yielded complementary and corroborating insights. At the same time, a key limitation of this thesis is that it does not provide empirical evidence to what extent dropout predictions improve clinical results and outcomes in DMHIs. Given that this PhD thesis is written in the Department of Information Science, and considering that RCTs require extensive clinical resources, conducting such trials was beyond the scope of this work.

The investigated interventions cover the most common psychological disorders effectively addressed with DMHIs (i.e., depression, anxiety, and eating disorders). As such, the insights add up to a possible global target group of up to 595 million (W.H.O., 2022). Moreover, the data sets range from dissemination trials and RCTs of prevention measures to psychiatric routine care treatment for severe symptom levels. Despite the reported differences in dropout patterns (Lipschitz et al., 2023), this thesis suggests similar tendencies in terms of dropout predictions, thus providing insights across a range of settings.

In the following, the contributions made to each of the introduced research questions are synthesised and evaluated. Table 1 indicates which article is discussed in which section.

4.1 RQ 1: Small Data Set Sizes

Chapters II and III contribute to the knowledge on how data set sizes influence dropout prediction performance (RQ1.a) and what strategies can be used to mitigate the problem (RQ1.b). In terms of commonalities, both articles attest to the problems inherent in the commonly used small data set sizes ($N \leq 300$) in DMHI predictions. As such, the risk of mistaking validation results from small datasets for adequate performance measures is quantified for two unrelated data sets at up to 0.12 in AUC. Both articles further provide insights into when dataset sizes negatively correlate with prediction performance, a phenomenon previously reported for outcome predictions (Sajjadian et al., 2021). Chapter III proposes that this happens in single-intervention settings but can be prevented through pooling interventions. Chapter II shows that– despite pooling

interventions – this phenomenon may still occur in the case of simple and uninformative features.

Further, Chapter II extensively discusses domain-specific learning curves – not only considering performance levels and variance but also their convergence. The provided evidence that complex features and models require larger data sets to converge in performance contextualises the heterogeneity of previous findings. The findings on high result variance corroborate the theoretical and simulated assumptions from Chapter I:2.1, proving that reporting a single test or CV results carries an immense risk of result oscillation. Following the learning curve approach led to a proposed minimum dataset size of N=500-1,000, which is currently only met by 50-33% of related work. Chapter II is the first work comprehensively investigating sample size adequacy in DMHI predictions. Recommendations to mitigate overfitting include minimally necessary data set sizes (N≥500) but also measures such as nested CV and selecting appropriate model, hyperparameter and feature combinations.

Chapter III introduces a different and pragmatic approach to improve prediction results and mitigate overfitting when facing small data set sizes - Pooling different interventions to one data set. Despite three different target disorders and dropout rates, the results from pooled data sets were twice as likely to achieve clinically relevant prediction levels (Forsell et al., 2022) and more stable than those of single interventions. As it is not uncommon for providers to have different but similarly setup interventions, this paper proposes a simple way to double or multiply data set sizes and simultaneously lower costs for model training and maintenance.

In conclusion, these studies leverage two of the largest intervention behaviour data set sizes so far (N=3,654-6,418) to comprehensively discuss the role of data set sizes in DMHI dropout predictions. Their results challenge the current standards of (dropout) predictions in DMHIs and provide an empirical reference for authors, clinicians and reviewers alike. The strengths of the papers are that they not only flag and quantify problems but also offer actionable insights and recommendations to mitigate them. Further, they replicate similar results despite significant differences in set-ups (prevention dissemination trial vs routine care treatment), countries (Germany vs Sweden), and target disorders (eating vs affective disorders). However, they still only offer case-based insights into a heterogeneous and complex question. A variety of changes in methodological decisions, data types and settings could influence results. Moreover, the fact that despite these insights, this thesis presents a study on only 181 users (Chapter IV) attests to the difficulties of implementing the minimal data set requirements. At the same time, Chapter IV also serves as evidence that using nested CV as proposed validation method mitigates overfitting.

In summary, the discussed articles provide the first empirical reference for the role of data set sizes in producing and interpreting ML dropout predictions in DMHIs.

4.2 RQ 2: Feature and Model Combinations

The four ML case studies of this thesis contribute insights into a variety of feature and model combinations of different sophistication levels. Features range from socio-

demographic, symptom scores, psychological measures, intervention behaviour, evaluation, and linguistic features to computer mouse trajectories. The ML models range from simple non-sequential Naïve Bayes and Logistic Regression over Support Vector Machines to tree-based models such as Random Forest, AdaBoost, and XGBoost. The second category comprises sequential models, i.e., recurrent and convolutional neural networks and a pre-trained BERT model. Table 2 summarises the AUC per feature type and paper. The results of the sequential neural networks are reported in the "*other*" column.

Table 2: Overview area under the curve (AUC) per feature type

Short Title	Dropout and Prediction Time	Baseline Questionnaire	Intervention Behaviour	Other
Chapter II: Learning Curves	63% dropouts with <4 modules After first week	Simple: 0.53 Extended: 0.66	Simple: 0.72 Selected: 0.80 Extended: 0.77	Mixed behaviour and questionnaire: 0.81
Chapter III: Intervention Data Pooling	44% dropouts with <7/10 modules After four weeks	-	-	Mixed user behaviour and questionnaire: 0.77
Chapter IV: Computer Mouse Trajectories	48% dropouts with <7 or <6/10 modules Preintervention	0.56	-	Hand-crafted mouse features: extended: 0.56 simple: 0.58 1D-CNN + temporal mouse features: 0.50
Chapter V: Intervention text Features	25% dropouts with <6/8 modules After two modules	0.60	Text metadata: 0.65	Evaluation + simple text metadata: 0.69 Word2Vec + baseline data: 0.70 BERT + evaluation data: 0.67

Notes: All metrics are reported as AUCs. In case of several options, only the most promising set-up (e.g., the largest data set size investigated) is reported; where applicable different interventions are averaged.

A key commonality across findings is that theory-driven feature and simpler model combinations are preferable in their cost-benefit ratio. In Chapter II, the twelve selected behaviour features outperformed the seven simple and 129 extended ones. In Chapters IV and V, neither simple feature-model combination was significantly outperformed by neural networks trained on sequential data. In Chapters II and V, adding features partly deteriorated results. This finding starkly contrasts what would be expected in the field of Data Science (Chapter I:1.2).

However, exhaustive relational data is the cornerstone of the Data Science epistemology (Kitchin, 2014; Yarkoni & Westfall, 2017). In contrast, data sets in DMHI tend to be small, static and poorly relational and, thus, seem to not yet meet the requirements to fully move to a theory-poor, data-rich approach (Yarkoni & Westfall, 2017). Further, features within a specific setting are often highly multicollinear (Patel et al., 2008; Sander et al., 2021; Tomitaka & Furukawa, 2021) as – contrary to multimodal big data – larger numbers of features in DMHIs often stem from few underlying processes (e.g., log files Bremer et al. (2020)). Consequently, the decreasing marginal benefit of

additional features may be easier outweighed by the added noise. The conclusion is, therefore, that applying conjuncture to identify the most promising features improves results as it increases the information-to-noise ratio (Piccialli et al., 2021; Yarkoni & Westfall, 2017).

At the same time, Chapter V provides evidence contrary to previous findings (Funk et al., 2020; Gogoulou et al., 2021) that complex features such as word embeddings yield competitive results when paired with adequate sequential models. It is also the first work contributing evidence for the potential of pre-trained transformer models in DMHIs. Moreover, in Chapter II, the extended feature groups showed a strong growth trajectory, and the deteriorating effect of combining baseline and intervention features on small data was reversed on larger data sets. Combined with the clear superiority of larger feature numbers and Deep Learning models in other areas of medicine (Piccialli et al., 2021), these findings foster the expectations that increasing data set sizes may tip the scale towards the expected superiority of more sophisticated model-feature combinations.

In terms of more specific findings, Chapters II, IV, and V corroborate the previously reported difficulty of predicting intervention dropout with baseline data only. While mouse trajectories could not solve the problem of pre-intervention dropout predictions in the investigated set-up, Chapter IV contributes to the field of research by introducing a new and innovative feature type. While its key limitation is the small data set size caused by the high number of mobile users, its strength lies in the rigorous methodology aimed at producing generalisable results (Hullman et al., 2022; Yarkoni & Westfall, 2017). Further, publishing the tracker and code to process mouse trajectories advances the dissemination of this novel data type and adheres to open science principles.

However, at 0.66 AUC, the baseline prediction in Chapter II is higher than the other two studies (0.60 and 0.56), and related work (0.57) (Gonzalez Salas Duhne et al., 2022; Günther et al., 2023; Linardon et al., 2020). Investigating the feature importance reveals that whether a user filled out the optional baseline questionnaire parts was by far the most predictive feature group. Additionally, in Chapter V the number of useless (e.g., "*x*" or "-") answers was among the most predictive meta-text features for dropout (AUC=0.60). As such, recording questionnaire non-response as additional variables before imputing missing values is a promising approach that emerged from this thesis.

The three ML case studies using intervention behaviour data further emphasise its importance and predictive power for dropout predictions. The studies at hand do not sufficiently compare to be able to identify reasons for the differences in predictive power. How users evaluated the intervention after the first two sessions regarding time adequacy and usefulness combined with the lengths of their answers yielded the best dropout predictions in Chapter V. So far, none of the related work has discussed the user's answers about their experience with the intervention as a data type. Such questionnaires are standard in many non-clinical settings and can be easily implemented. Finally, adding baseline data to the intervention data improved results, though only slightly (Chapters II and IV). In summary, all models using data from the intervention

time achieved good or very good prediction power (Kraemer et al., 2003) while leaving enough time to act on the prediction information. Further, evaluation data was proposed as a promising new and low-cost feature type for dropout predictions.

In conclusion, the four ML case studies contribute detailed insights into a variety of feature and model pairs and their respective performance. Beyond increasing the available evidence on the predictive power of baseline and intervention behaviour data, it proposes two new (mouse trajectory, evaluation data) and provides evidence for a so-far unsuccessful (word embeddings) complex feature type. However, the advantages of sequential neural networks are non-existent or negligible in the current data sets and are arguably outweighed by the increase in requirements for Deep Learning models. Thus, simple models and feature types are concluded to be preferable for the data sets at hand.

4.3 RQ 3: User Adaption Hurdles

Chapters VI and VII explore two hurdles identified in the user adaption of ML-based DSTs in DMHIs. Chapter VI shows that explainability, in the form of local SHAP values, significantly increased therapists' trust ($p=0.01$, $d=0.43$, 95%-CI [0.07-0.76]) and the perceived transparency ($p<0.001$, RBC=0.69), and usefulness ($p=0.003$, rank biserial correlation=0.66) of individual predictions with noteworthy effect sizes. At the same time, the more extensive and validated Human-Computer Interaction Trust Score (Gulati et al., 2019) was not significantly higher, and the study failed to support the underlying assumption of change in behavioural intention, at least for intended support levels. Further, it has relatively low power (N=35), but as data collection is still ongoing, this limitation will be mitigated in the final manuscript. That final manuscript will also investigate to what extent therapists interpret SHAP causally, as this is a common and potentially risky pitfall (Lundberg & Lee, 2017). It, however, does not measure the final impact on patient outcomes but instead serves as an exploration of how to lower the widely reported adaption hurdle of DSTs (Devaraj et al., 2014; Gille et al., 2020; Jacobs et al., 2021). Ultimately, how ML algorithms arrive at their predictions and to what extent this should and does matter to clinicians using the predictions remains an epistemic question. At the same time, Chapter VI contributes to the knowledge necessary to translate ML predictions from theory into clinical practice.

Beyond local SHAP values, Chapter IV uses global SHAP values to provide additional insights into the inner workings of the models' decision processes. It further applies the checklist of Scott et al. (2021) to extensively discuss the clinical usefulness and implementation-relevant aspects of the proposed model. Among the discussed aspects are data quality, use cases, the effect of changing thresholds, and ways to display the predictions to clinicians. This discussion gives interested clinicians a much more detailed picture of the algorithm and its possible use cases. At the same time, it added three pages of text to the article, making it an arguably unfeasible standard for most Data Science publications. In terms of model implementation, however, it is a structured and comprehensive way of ensuring that the models' strengths and weaknesses are discussed and communicated.

Lastly, Chapter VII contributes to the question of how personalization can be (and is) implemented in DMHIs through the example of depressive symptoms. The key contributions are the proposed framework and the systematic analysis of 138 articles to quantify the current use of personalization, including studies evaluating its value-add.

A common logic to classify the activity of interest is a necessity when aiming to synthesise and evaluate research. Defining personalization as purposefully designed individual variation and differentiating it from other forms of adapting interventions offers such a categorisation. Researchers or clinicians designing a DST can use the categories and mechanisms to recommend what to do with a prediction. Chapter VII, therefore, sets a milestone for future (meta) research on personalization in DMHIs. Further, proving that 66% of interventions already apply some form of personalization, most of them as rule-based mechanisms, shows the theoretical potential for ML-based applications. Lastly, only three studies currently implement ML-based personalization, and only two investigated the value-add of personalization with null results on very small ($N<60$) samples, clearly showing the need for further research. Providing transparency to the state of research raises the issue that the amount of empirical evidence mismatches the commonly assumed value of personalization.

In conclusion, the contributed articles shine a light on the current state of knowledge regarding the role of trust and personalization options in DMHIs. Further, they propose and discuss different aspects necessary to translate proof-of-concept predictions into complex, high-stake and user-driven clinical processes.

4.4 Open Questions and Outlook

Despite this thesis' extensive and systematic engagement with a single problem, several decisive questions remain open regarding the value-enhancing implementation of dropout predictions in DMHIs. This section, therefore, attempts to elucidate the different levels at which ML applications in DMHIs could be furthered in the future.

First and foremost, any claim of the revolutionary paradigm change induced by ML-based DSTs bases on the conjuncture that personalization can improve outcomes and dropout at scale. However, much more evidence is necessary to investigate the extent and nature of this claim, as good predictions do not necessarily translate to useful clinical actions. A decisive aspect of future work is, thus, the prospective evaluation of the effects of Machine Learning-based DSTs on patients.

Secondly, several of the key problems discussed in this thesis originate from the mismatch between the ontological view of Data Science and the reality of DSTs in clinical mental health. In the end, most studies pair less favourable prerequisites with much higher stakes while often not adhering to the methodological standards of Data Science methodology (Cabitza & Campagner, 2021; DeMasi et al., 2017; Hullman et al., 2022; Yarkoni & Westfall, 2017). The consequences are possible contradictory or pluralistic results that often do not suffice to derive generalisable rules. Relevant aspects concern data set sizes, validation set-up and target leak (Hullman et al., 2022; Yarkoni & Westfall, 2017). Considering the availability of various domain-specific checklists (Cabitza & Campagner, 2021; Olczak et al., 2021; Scott et al., 2021), the field of

research would benefit from their enforcement as it is commonly done in classical statistical analysis (Hullman et al., 2022).

Thirdly, many more feature types and model options are possible and promising in predicting dropout. Examples are data from mobile phones, for example, GPS data (Terhorst et al., 2023). As an example, a recent study leveraged Large Language Models with multi-sensor mobile data to identify possible areas of interest, such as trends in sleep or movement (Englhardt et al., 2024). Adding aspects from outside the intervention could possibly enhance predictions and their actionability. Additionally, the research at hand only considers a limited number of supervised ML algorithms. While these are the most-used choice for the problem at hand, other modelling alternatives can be considered. A key example of such an alternative is recommender systems to directly personalize messages (Sadasivam et al., 2016) or content. Generally, more data-driven industries, such as e-commerce or entertainment, may serve as a source of inspiration.

However, as commonly said, more data is better than a better algorithm or feature type (Hullman et al., 2022). Therefore, fourthly, the availability of larger data set sizes determines much of the future success of more sophisticated approaches (Hullman et al., 2022; Yarkoni & Westfall, 2017).

As a fifth point, the ML approach assumes that the individual level of insights (predictions) renders population-based insights unnecessary. However, if high accuracy is not what drives adaption, this field of research risks ending up with what Hullman et al. (2022) call "*the worst of both worlds*". For instance, if therapists need causal information to trust and know how to act on predictions, local SHAP values are a bad proxy as they do not provide that information (Lundberg & Lee, 2017). Instead, simpler and easier-to-explain models may be preferable in their impact, even if they are not as accurate. This development is not unique to mental health, as data-driven researchers are moving from purely empirical towards critical realism (Hullman et al., 2022).

Sixth, and more specifically concerning dropout, as discussed by Beintner, Vollert, et al. (2019) and mentioned in Chapter I:1.1, many different dropout definitions are possible and may influence results. Standardisation is necessary to further this area of research by allowing comparisons and meta-analytic evaluations of findings.

In summary, moving away from context-specific and static proof-of-concepts on RCT data is imperative. Instead, prospective comparative studies in continuously used routine-care data settings are paramount in the reasonable advancement of the field. While some such studies are currently ongoing (e.g., Bjurner et al., 2024, PERSONAE at Region of Southern Denmark), they are still few. Further, for this development to be successful at scale, the legal frameworks and technological infrastructure must be developed, and interdisciplinary groups are necessary to ensure the domain knowledge to determine usefulness and the methodological rigour to ensure generalisability.

5 Conclusion

While DMHIs are an effective way of increasing and improving mental health care, such interventions are often plagued by high dropout rates (Beintner, Vollert et al.,

2019; Gan et al., 2021; Lipschitz et al., 2023). Leveraging the power of Machine Learning promises to efficiently allocate resources to those most at risk. However, 1) small data set sizes, 2) a plethora of features and model types, and 3) the gap between merely accurate and factually used predictions may be hindering this progress.

This PhD contributes to the field of research in various ways. First, it quantifies the risk of overfitting on small datasets and proposes a minimal necessary data set size of N=500-1000 in dependence on the setting. It further offers insights into several measures to mitigate this problem and provides evidence that pooling interventions improves and stabilises results. The systematic investigation of learning curves shows that complex models and feature types converge later. Moreover, all experiments except the mouse data were conducted on data set sizes between two and 19 times as high as the median data set size of the twelve related works. Thus, this thesis significantly increased the number of reliable publication results for dropout predictions in DMHIs.

Secondly, the four ML studies investigated the predictive power of both established and, in this setting, innovative data types. A critical common finding is that handcrafted theory-driven features paired with simple models prevail over more complex ones. At the same time, more sophisticated features and model types show promise as data set sizes increase.

Thirdly, this thesis investigated how the user adaption of ML-based DSTs could be facilitated. To that end, it found that local SHAP values increase the perceived trust and assumed usefulness of ML predictions. However, this did not translate into a bigger change in intended behaviour when therapists were asked to decide who should get increased human guidance levels. Further, the systematic review revealed that personalization is common in DMHIs for depressive symptoms, but very few studies use ML or provide evidence for the value of personalization. As such, the paper offers an extensive list of personalization options researchers and practitioners can draw from and improve on and calls for more evidence for the value of personalization.

At the same time, the synthesis of the findings reveals several unanswered questions. Many of the problems are enate in the attempt to translate the revolutionary success of AI from areas such as e-commerce into the very different and complex healthcare world. In the long run, it is unlikely that the often-proclaimed paradigm shift will be realised without larger data sets and the systematic measurement of the value-add of ML predictions in ongoing care. Despite these challenges, ML-based precision care remains a crucial lever in advancing mental health care, ultimately aiming to improve patients' lives in an overwhelmed system.

In summary, this PhD not only increases the body of research available but also furthers the academic and practical use of dropout predictions in DMHIs by addressing the three identified key problems. Beyond the exploration of research gaps, the findings challenge existing standards and provide critical insights, resulting in the provision of substantial empirical references to guide future research design and clinical use.

Chapter II: Learning Curves

Zantvoort, K., Nacke, B., Görlich, D., Hornstein, S., Jacobi, C., & Funk, B. (2024). Predictive Power, Variance and Generalizability – A Machine Learning Case Study on Minimal Necessary Data Sets Sizes in Digital Mental Health Intervention Predictions.

(Under revision at NPJ Digital Health since 28th, August 2024, submitted 21st, June 2024)

Abstract: Artificial Intelligence promises to revolutionize mental health care, but small dataset sizes and lack of robust methods raise concerns about result generalisability. As insights into minimal necessary data set sizes are scarce, this study explores domain-specific learning curves for intervention dropout predictions. Prediction performance is analysed based on dataset size (N=100-3,654), feature groups (F=2-129), and algorithm choice (from Naive Bayes to neural networks). The results substantiate the concern that small datasets (N≤300) overestimate predictive power. For uninformative feature groups, prediction performance was negatively correlated with dataset size. Sophisticated models overfitted in small datasets but were crucial for maximizing test results in larger datasets. While N=500 mitigated overfitting, performance did not converge until N=750-1500. Consequently, we propose a minimum dataset size of N=500-1,000, depending on feature complexity and information value. As such, this study offers an empirical reference for researchers designing or interpreting AI studies on Digital Mental Health Intervention data.

1 Introduction

The rapid advancement of artificial intelligence (AI) in various industries has spurred great anticipation for its transformative power in health care (Ben-Israel et al., 2020; Cruz Rivera et al., 2023). One area that particularly stands to benefit from AI-based improvements is mental health (Aafjes-van Doorn et al., 2021; Shatte et al., 2018). With 16% of global disability-adjusted life years attributed to them and staggering economic costs, mental disorders are immensely burdensome for individuals and societies alike (Arias et al., 2022). Further, mental disorders are heterogeneous in their treatment needs, and AI promises a resource-efficient way to personalize, scale and improve mental health care (Aafjes-van Doorn et al., 2021; DeMasi et al., 2017; Hornstein et al., 2023; Squires et al., 2023). However, among the central challenges in realizing AI's envisioned potential within mental health interventions (MHIs) is the limitation of data set sizes (Aafjes-van Doorn et al., 2021; Bzdok & Meyer-Lindenberg, 2018; DeMasi et al., 2017; Sajjadian et al., 2021; Squires et al., 2023).

In contrast to diagnostics or public health data (Shatte et al., 2018), median data set sizes of Machine Learning (ML) application studies with MHI data barely exceed 100-150 patients (Aafjes-van Doorn et al., 2021; Sajjadian et al., 2021; Squires et al., 2023; Vieira et al., 2022). Digital mental health interventions (DMHIs) allow for an easier collection of datasets than face-to-face (f2f) therapy (Bremer et al., 2020; Hornstein et al., 2023), but median data set sizes are still only 155-350 (Hornstein et al., 2023; Karyotaki et al., 2021; Zantvoort, Hentati Isacsson, et al., 2024). This is problematic because prediction power is notoriously known to be overestimated in such small data set sizes (Bates et al., 2022; Hastie et al., 2017; Lateh et al., 2017).

Sajjadian et al. (2021) found that MHI studies with small data set sizes reported significantly higher performance metrics than methodologically sound studies ($p=0.005$). Similarly, Zantvoort et al. (2024) reported that DMHI dropout prediction models trained on small data sets produced the highest cross-validation (CV) results but performed worst on the larger test set. As a result, several authors caution the interpretation of the current state of results and warn about possible consequences. Deploying an ungeneralisable model risks suboptimal care, deteriorating patient outcomes, wasted resources, and, thus, ultimately leads to the opposite of the intended effects (Chekroud et al., 2024; DeMasi et al., 2017; Hilbert et al., 2024; Sajjadian et al., 2021; Zantvoort, Hentati Isacsson, et al., 2024).

Despite their undebated relevance, minimal necessary sample sizes, as they are standard in classical statistical settings, are uncommon in ML applications (Balki et al., 2019). While no all-encompassing solution is available, a key approach for better understanding them are learning curves (Balki et al., 2019; Giesemann et al., 2023; Smeden et al., 2019). A recent study by Giesemann et al. (2023) produced such learning curves for dropout predictions in f2f psychotherapy and suggested 300 data points as a minimal necessary sample size. However, they only used eight patient-reported features and did not investigate overfitting or result variance. Further, only minimal insights are available into the interaction effect of sample sizes, model types and the

number and type of features in DMHI data. Flexible models approximate realities' complexity well, however, they risk overfitting, especially on small data sets (Bzdok & Meyer-Lindenberg, 2018; Perlich et al., 2004; Zantvoort, Hentati Isacsson, et al., 2024). Simple models tend to produce more stable results but risk disregarding valuable information (Atla et al., 2011; Fernandez-Delgado et al., 2014; Kwon & Sim, 2013). Additionally, the effectiveness of any model significantly depends on the nature and number of predictors (Kwon & Sim, 2013; Smeden et al., 2019). Especially for DMHIs, feature numbers can quickly grow into hundreds of variables (Bremer et al., 2020; Zantvoort et al., 2023). At the same time, data protection and adherence concerns call for a data minimalism approach (Bremer et al., 2020; Cote-Allard et al., 2022; Forsell et al., 2020). Moreover, several papers have reported that fewer handcrafted variables improved their results (Bremer et al., 2020; Bricker et al., 2023; Hentati Isacsson et al., 2023).

In conclusion, the key questions repeatedly arising in ML studies in DMHIs are 1) how the dataset size influences the results (Bremer et al., 2020; Giesemann et al., 2023; Sajjadian et al., 2021), 2) which of the ample algorithms to implement (Bricker et al., 2023; Cote-Allard et al., 2022; Fernandez-Delgado et al., 2014; Giesemann et al., 2023; Linardon et al., 2022; Zantvoort et al., 2023), and 3) which of the abundant possible variables to use (Bremer et al., 2020; Bricker et al., 2023; Zantvoort et al., 2023). The current study aims to investigate the interdependence of these questions by analysing the learning curves for dropout predictions across six models with varying levels of flexibility and six feature groups differing in their predictive power and extent. To derive insights into minimal necessary data set sizes, the results will be investigated not only regarding their performance level but also their variance, generalisability and trajectory. To this end, we leverage 3,654 users' data from digital eating disorder prevention interventions provided to the general public in Germany (Nacke et al., 2019). Eating disorders are highly prevalent (Galmiche et al., 2019) and associated with immense levels of suffering (American Psychiatric Association, 2006). While DMHIs are effective in preventing and treating EDs, intervention dropout is a substantial issue among them (Linardon et al., 2020). Measures such as guidance can mitigate dropout but are costly (Hilvert-Bruce et al., 2012; Pedersen et al., 2019). Using AI to identify users at risk of dropping out allows for optimising resource allocation and improving outcomes regardless of the availability of final symptom scores (Bricker et al., 2023; Hilvert-Bruce et al., 2012; Pedersen et al., 2019). As such, within a case study's limits, this paper seeks to provide insights to improve the design and interpretation of ML studies on DMHI data.

2 Results

2.1 Final Values

The final data set comprised 3,654 users, of whom 63% were classified as dropouts. Feature groups ranged from 2 features (F) (simple questionnaire), over 7 (simple behaviour), 13 (selected behaviour), 51 (extended questionnaire), and 64 (mixed) to a

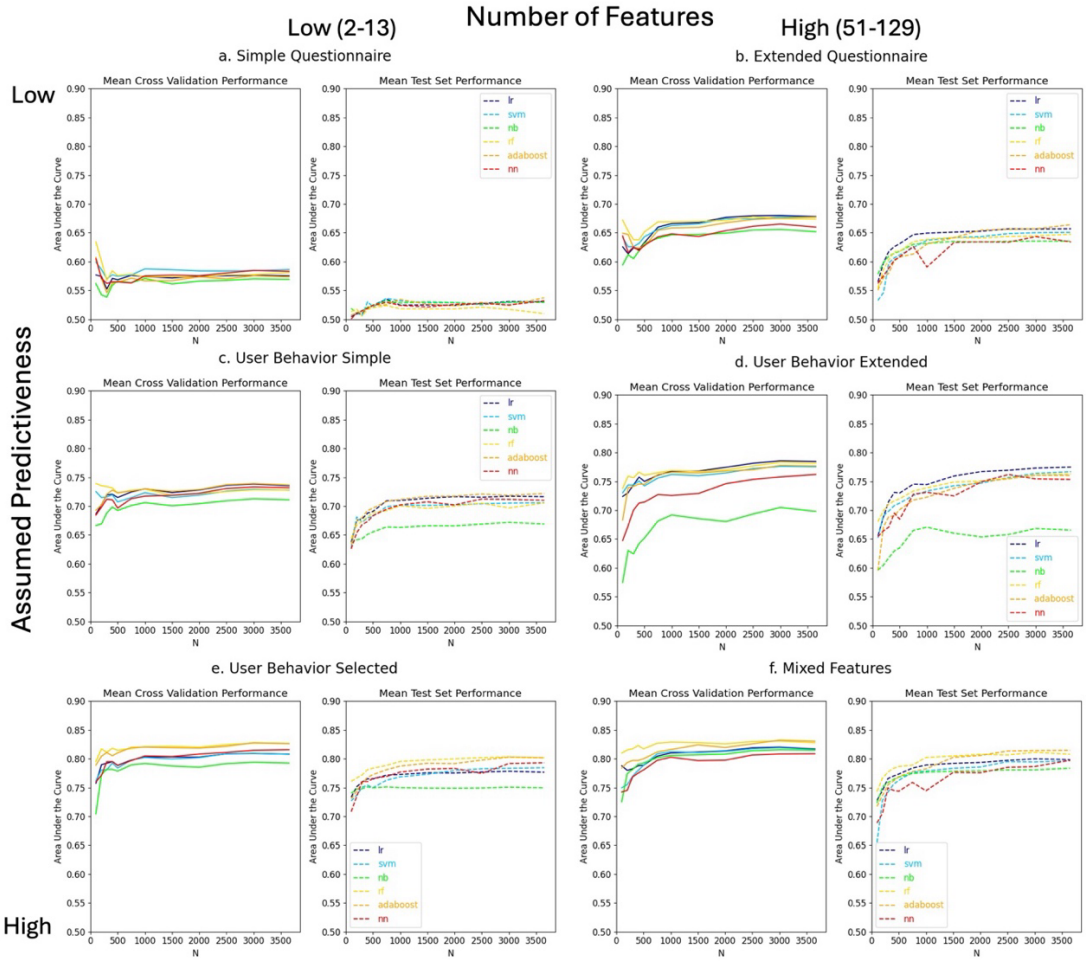
maximum of 129 features (extended behaviour) in addition to the intervention information. The descriptive statistics, including for the training and test set, can be found in Supplementary Table 1.

Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machines, (SVM) Random Forest (RF), adaBoost and Multilayer Perceptron Neural Network (NN) models were trained with 10-fold CV on 80% of the assumed data set sizes between 100 and 3,654 users and evaluated on the test set of 731 users. Hyperparameters differed across settings (e.g., regularisation for 7 vs 129 features), and are published in this study's GitHub repository. For clarity and brevity, the results are discussed on the area under the curve (AUC) score only, but the further metrics, including balanced accuracy, f1-score, and recall are published in Supplementary Table 2. Supplementary Table 3 holds the p-values for the DeLong tests.

2.2 Predictive Power of Feature Groups

Approximating the prediction performance via the best model on N=3,654, the assumed predictive power across feature types was confirmed. There was no information in the simple (0.53 test AUC) and only moderate (0.66 AUC) in the extended questionnaire data. The simple behaviour data already achieved an AUC of 0.72, which was increased to 0.77 for the extended and 0.80 for the selected behaviour data. From there, the mixed features only slightly increased results to 0.81 AUC. Since the simple questionnaire data had no predictive power, its results will only be discussed in the context of overfitting.

Figure 3: Training and test learning curves per feature type



Notes: Each panels shows the respective mean AUC score for the cross validation on the training data (solid line) on the respective left and mean test data performance (dotted line) on the right side

2.3 Overfitting on Small Data Set Sizes

Overfitting was a substantial problem for the small data set sizes (N≤300), as they were prone to produce much larger CV results than test results —however, the extent varied across the model and feature types.

Regarding models, at N=100, the share of CV results with at least +0.10 higher AUC than the test results was by far the lowest for NB (avg. 13%). The NNs, LR and SVMs followed with 25%, 33%, and 38%, respectively. On the highest end of the spectrum were the tree-based models, on average, overestimating predictive power by more than +0.10 AUC in 42% (adaBoost) and 45% (RF) of the cases. These numbers dropped substantially to an average of 7-8% for N=300. Increasing to 500 data points eventually lowered the share to 0% for all models and runs but the simple questionnaire data and the NB for selected behavioural data.

In terms of feature types, low-information feature groups (simple and extended questionnaire) were the most likely to overfit. For data set sizes of N≤300, the avg. difference between the training and test scores without NB was -0.07 (max. -0.12 AUC).

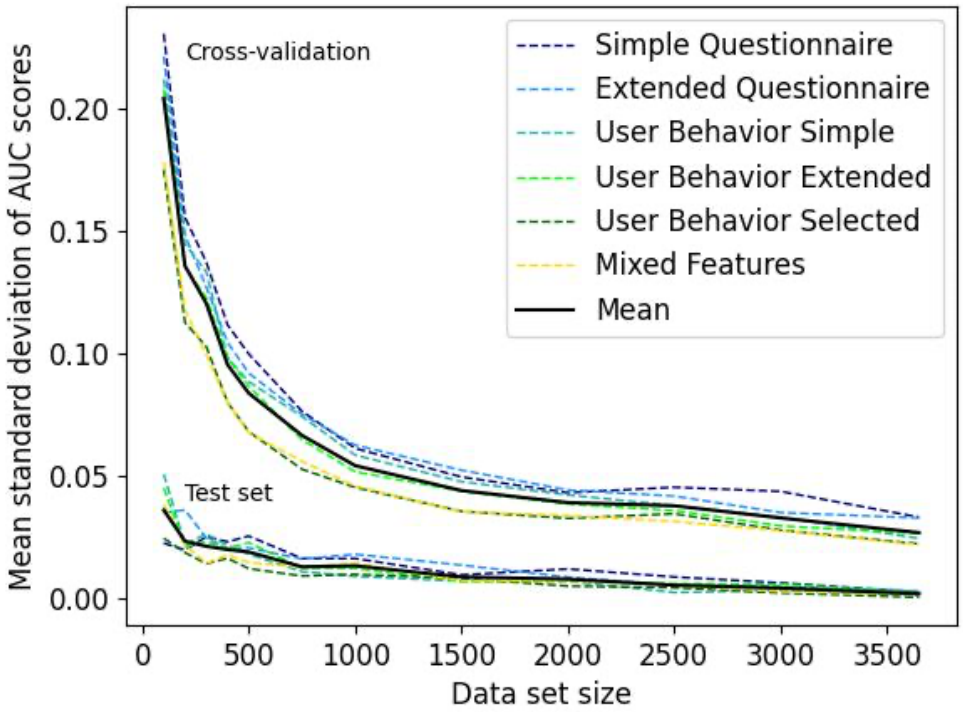
Choosing the winning model based on CV scores for the simple questionnaire data led to up to 70% of the results being >0.61 AUC despite a useless model. While the overfitting decreased for N>300 (avg. -0.02 AUC) in the extended questionnaire data, it continued to be -0.05 AUC for all models but the NB for the simple questionnaire data. Further, for these two feature types, up to N=300 training results got worse with increasing data set size (avg. -0.03, max. -0.06 AUC) as seen in Figure 3 a. and b. The same was visible in the simple behavioural data (Figure 3 c.) but less severe and only for RF and SVM (avg. and max. -0.02 AUC for N≤500).

For the extended behaviour, selected behaviour and mixed data, gaps between training and test set performance for N≤300 were also prevalent but less severe (avg. -0.05, max. -0.09 AUC). For these three most informative feature groups, both training and test results increased with data set sizes, and the models winning in the training scores consistently also produced the highest test scores.

2.4 Variance of Results

The results of the individual validation folds were highly unstable for small data set sizes. The AUCs' standard deviation (S.D.) steeply declined as the data set size increased. As such, the variability of AUC results was by far highest for N=100 (avg. S.D. 0.20 AUC) but quickly fell to half that value by N=400, as seen in Figure 4. After that, it continued to drop, with the lowest average value in our results being S.D. 0.03 AUC at N=3,654. The test set of 731 users also had a downward trajectory, however, with much smaller values (0.036-0.002 AUC for N=100-3654). Parallel to the observations in overfitting, the result variance was highest for the unpredictive feature groups. The single validation folds of N=100 in the simple questionnaire data covered the entire AUC score range from very bad to very good (AUC mean 0.60, ±S.D. 0.37-0.83, min. 0.00, max. 1.00). Variance was lowest but still very high for the selected behaviour data (AUC mean 0.70, ±S.D. 0.52-0.94, min. 0.10, max. 1.00).

Figure 4: Result variance per feature type



Notes: Mean standard deviation of the single folds' area under the curve score as dotted lines, mean across all features in black solid line.

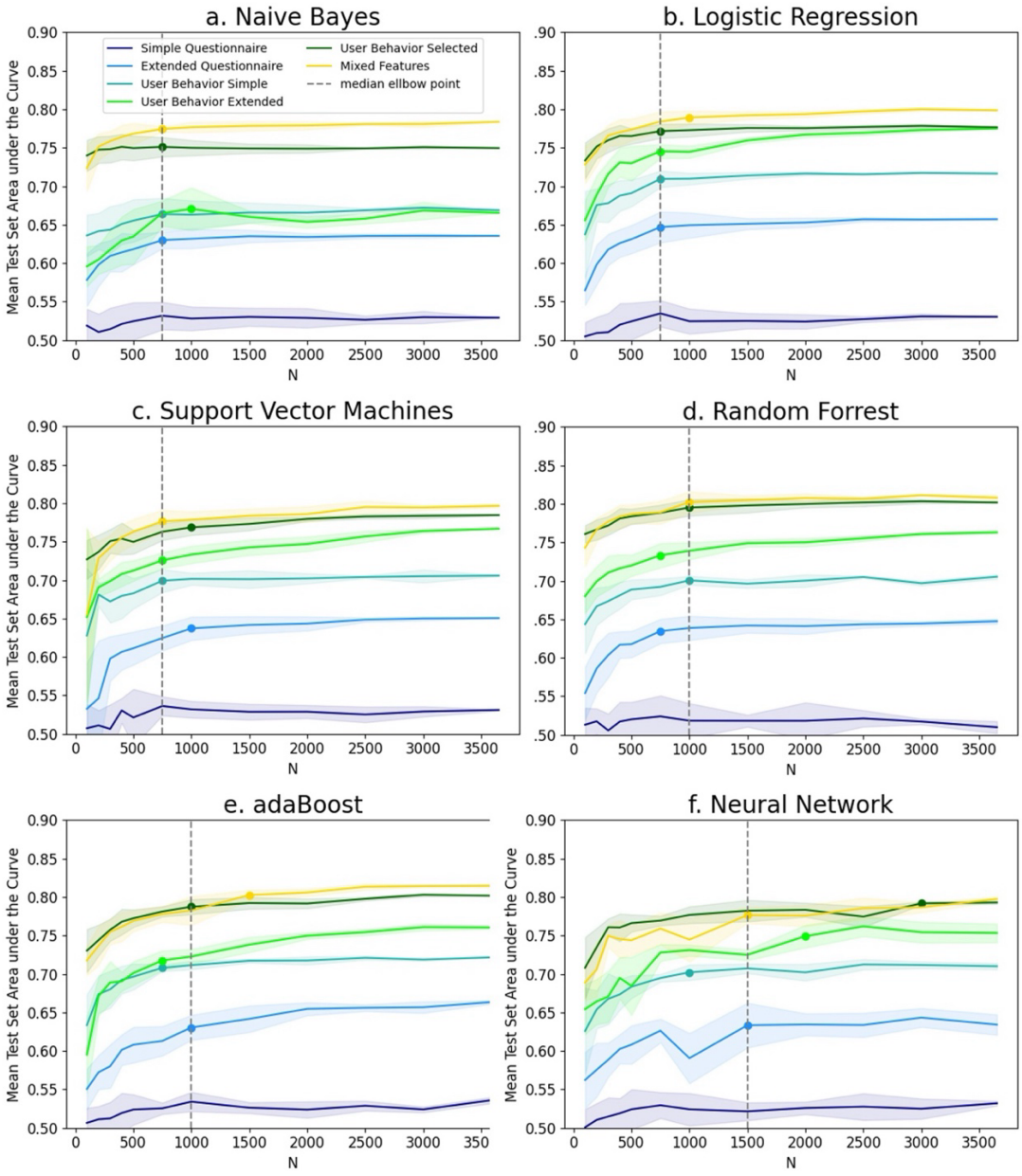
2.5 Performance Convergence per Model

The convergence points of the test set performance differed per model type and are shown in Figure 5. The simple questionnaire results are shown in the graphs but ignored in the calculations as there was no predictive power to converge towards.

The simpler models NB, LR and SVMs all had a median convergence point of N=750. The more sophisticated tree-based models converged later at N=1,000, followed by the NN at N=1,500. The NB had no performance improvement (+0% AUC) when provided with large data set sizes (N=3,654 instead of 750), whereas LR and RF, on average, grew +2%. SVMs and NNs could slightly better leverage the largest data set (+3%) but were surpassed by adaBoost on average increasing the AUC between N=750 and 3,654 by +5%.

It is noticeable that the NN showed oscillation and larger variability in the results for much longer than the other models, where this only occurred for very small data sets. Training it on the small data sets partly gave convergence warnings.

Figure 5: Test learning curve and convergence points per model type

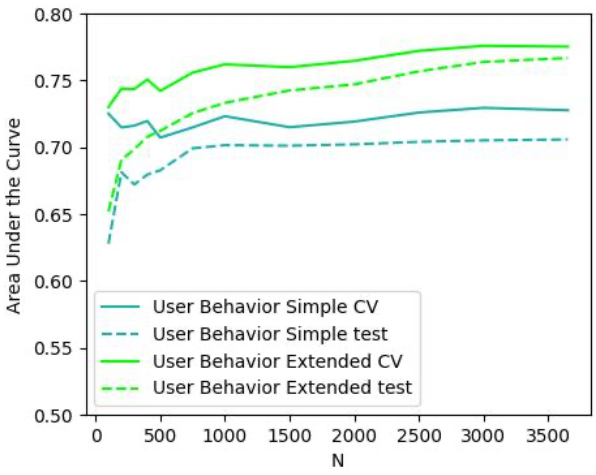


Learning curves on the test data per model: a) Naive Bayes, b) Logistic Regression, c) Support Vector Machines, d) Random Forest, e) adaBoost, and f) Multilayer Perceptron Neural Network. The colours indicate the different feature types, i.e., simple questionnaire (dark blue), extended questionnaire (light blue), simple behaviour (turquoise), extended behaviour (light green), selected behaviour (dark green), mixed features (yellow). The respective mean area under the curve score is shown as sold horizontal line and their S.D. as shaded area around it. Knee points indicate point of performance convergence as coloured scatter points for the individual and grey dotted line as median across feature types. Knee points are not shown for simple questionnaire due to lack of predictive power.

2.6 Marginal Value and Convergence of Additional Features

The marginal benefit of complex features was highest for large data set sizes, and more predictive feature groups tended to converge among the latest.

Figure 6: Random Forest training (CV) and test learning curve



Adding the extended questionnaire features to the simple ones (F=51) continuously improved results as the data set size grew (avg. 0.51-0.53 versus 0.55-0.66 test AUC for N=100-3,654). The same was the case for increasing the simple to the extended behaviour data (avg. 0.63-0.70 versus 0.64-0.75 test AUC). Due to overfitting on N=100, the best CV results for the simple behaviour were equal to those of the extended behaviour data (AUC=0.74), despite being lower on the test set (0.64 vs 0.68 AUC), as shown in Figure 6. This effect faded with increasing data set size, and at N=500, even the test set performance of the extended group surpassed the simple one's CV scores.

Similarly, using selected instead of extended behavioural data was most beneficial on the small data sets (+0.08-0.03 test AUC difference at N=100-3,654). Generally, for all models but the NB, the extended behaviour data curve was the steepest after N=1,000, such that it was closing the gap to the selected behaviour features. For LR, it even had already matched the selected behavioural data's performance at N=3,654 (Figure 5. b.).

Adding more than 50 questionnaire features to the selected behaviour data for the mixed data set first led to slightly less (N≤200, avg. difference in test AUC -0.02), then equal (N=300-500, 0.00), and ultimately slightly better performance (N>500, +0.01). As the only exception, using selected (F=13) instead of simple (F=7) behavioural data was always beneficial, but most so on the small data sets (avg. +0.12-0.08 test AUC difference for N=100-3,654).

2.7 Model and Feature Combinations

Naive Bayes (NB) obtained competitive test results (top3 models) for smaller data set sizes, specifically for the extended questionnaire (N≤750), mixed features (N≤400), selected behaviour (N≤200), and simple behaviour (N=100). However, NB never outperformed the respective other top3 models ($p>0.05$). Furthermore, as shown in Figure 3 c.-e., NB significantly underperformed compared to the other models for behaviour data, particularly for extended features and larger data set sizes ($p<0.05$).

Logistic Regression (LR), on the other hand, performed very well in almost all settings. It consistently outperformed most models for the extended questionnaire data for

N=200-500 ($p<0.05$). For N>500, LR continued performing well but was first matched by RF and later (N>2,500) by adaBoost. In the extended behaviour data, LR was below or equal to RF for N≤200 but significantly outperformed all models ($p<0.05$) with few exceptions after that.

Support Vector Machines (SVMs) mainly performed in the mid-field but were most competitive with a linear kernel in the two extended feature types. As such, they performed similarly to the top model LR on extended behaviour data for N>2500 ($p=0.06$-0.08) and regularly outperformed ($p<0.05$) NB, NN and adaBoost.

Similarly to LR, Random Forest models (RFs) performed very well, especially for the highest information feature types. They consistently outperformed all models for selected behaviour and mixed features, with the only regular exception being adaBoost for N>750 in selected behaviour and N>1,000 in mixed features.

adaBoost tended to perform better with larger data set sizes. For the highest information features, it progressively caught up to RF as of N>400. Additionally, adaBoost performed very well in the simple behaviour data (N>100) and the extended questionnaire data (N>1,500).

Multilayer Perceptron Neural Networks (NN) were among the top3 models for simple behaviour (N>750) and selected behaviour (N>200) data and occasionally performed well for extended behaviour data. NN's most competitive results were for data set sizes of 1,500 or more, where it was most likely to outperform NB, LR, or SVMs.

3 Discussion

Sophisticated ML models promise to disrupt mental healthcare through resource optimisation and personalization (Hornstein et al., 2023; Sajjadian et al., 2021), for example by lowering dropout (Pedersen et al., 2019) and improving health outcomes (Forsell et al., 2019). However, in DMHI settings, median data set sizes barely reach 155-350 (Hornstein et al., 2023; Sajjadian et al., 2021; Zantvoort et al., 2023). While such data set sizes are known to overfit and not suffice for many sophisticated models (Bates et al., 2022; Hastie et al., 2017; Lateh et al., 2017), very limited insights are available at what data set size these problems are mitigated. Therefore, the current study leveraged a dataset 10-24-times as big as the reported medians to evaluate performance levels, generalisability and variance across different feature groups (i.e., low to high predictive power with F=2-129) and six model types (Naïve Bayes, Logistic Regression, Support Vector Machines, Random Forest, adaBoost, and Multilayer Perceptron Neural Network models).

Our *first key finding* confirms that CV results on small, thus most common, data set sizes overestimate the prediction performance. Especially worrisome is that the effect was exacerbated for uninformative features, such that a useless model had up to a 70% likelihood of returning seemingly good CV scores. Further, we reproduced the negative correlation between data set sizes and CV results (Sajjadian et al., 2021; Zantvoort, Hentati Isacsson, et al., 2024) for N≤300 and partly N≤500 for the least predictive features. In these settings, such high training results were associated with the worst

test results (Chekroud et al., 2024; Zantvoort, Hentati Isacsson, et al., 2024). While overfitting was also prevalent in N≤300 for the more predictive features, it was lower, and the best training translated to the best test results. Further, among all features, the individual validation scores were highly unstable for N≤300 (S.D. 0.13-0.20 AUC). Evaluating on a single fold is common (Chekroud et al., 2024; Sajjadian et al., 2021), and publication bias risks an overrepresentation of the higher end of that variance in published studies (Andaur Navarro et al., 2021; Squires et al., 2023). Thus, we conclude that results from data set sizes of N≤300 imply a substantial risk of being inflationary and ungeneralisable, especially for features with low predictive power.

A *second, closely related key result* is that CV scores on small data sets risk underestimating the superiority of complex versus simple features. This is caused by, firstly, large data being necessary to leverage additional features and, secondly, simple features overfitting more. For the largest feature group (F=129), our data set size even may have been too small as it continued catching up to the already converged selected feature's performance. However, more research on larger data sets is necessary to investigate this hypothesis. Therefore, we tentatively confirm previous findings (Bremer et al., 2020; Hentati Isacsson et al., 2023) that hand-crafted and theoretically driven selected features are preferable, especially for small data sets.

The *third key result* confirms that simpler models are less likely to overfit but converge earlier and are less competitive for higher data set sizes. More flexible models, on the other hand, heavily overfit small data sets but produce the best results on the high information features, especially for large data set sizes. Consistent with theory and empirical evidence (Atla et al., 2011; Nettleton et al., 2010), particularly NB gave robust results but was not very competitive overall. On the other end of the spectrum, especially RF and SVMs seemed very competitive on noisy and small data sets but actually overfitted (Atla et al., 2011; Rodriguez-Galiano & Chica-Rivas, 2014; Saseendran et al., 2019). adaBoost performed badly on small but was most effective in leveraging large data sets. RF was one of the two most competitive algorithms across settings, already efficiently leveraging mid-sized data sets for predictive features (Atla et al., 2011). LR was the second competitive algorithm, confirming its balance of overfitting less on small data sets (Saseendran et al., 2019) but only partly being outperformed in large data sets. The fact that LR is easier to interpret and faster to train than the tree-based models emphasises its essential role as a staple baseline model to beat (DeMasi et al., 2017; Hentati Isacsson et al., 2023; Zantvoort et al., 2023).

The *fourth key finding* is that prediction performance in our study did not converge until N=750 for simpler and 1,000-1,500 for more sophisticated models. Both are substantially above Giesemann et al.'s findings that their results stopped improving at N=300. A possible explanation is that their study on f2f-therapy investigated only eight features, which all fall within our extended baseline definition. As a result, their maximum test AUC score (N=10.000) was 0.62, which our extended baseline data also achieved at N=300. Further, in our data, more predictive features partly converged later than those with less information value. One possible hypothesis could, therefore, be that their earlier convergence point may be due to the limited predictive information

in the features used. Thus, we conclude that more sophisticated models paired with larger data set sizes (N>750) are necessary to approximate the true potential for the common feature groups in DMHIs.

Beyond the potentially still too-small sample size of 3,654, this paper has several limitations. Firstly, it is only one case study, and while concurring with previous knowledge, this study per design does not suffice to reliably differentiate between setting-specific and generalisable tendencies. Regarding sample bias, the interventions considered are preventative and the sample only comprises self-referred female participants. Further, the five study arms were heterogeneous in their content, lengths, and user symptom strength (Nacke et al., 2019). As pooling interventions already mitigates overfitting (Zantvoort, Hentati Isacsson, et al., 2024), results may differ if repeated on a single intervention. However, this also implies that overfitting in this study may be underestimated, making the proposed increase of minimal data set sizes even more critical. Hence, the current study presents first insights, but more research is necessary to confirm the proposed minimal data set sizes. As a second limitation, while the operationalisation of the outcome and feature groups was empirically and theoretically founded, many other options (Bremer et al., 2020; Smink et al., 2021; Zantvoort et al., 2023) are possible and may influence results. We proposed six different feature groups representing low to high predictiveness for intervention dropout, but they would, for example, differ in health outcome predictions (Hornstein et al., 2021; Zantvoort et al., 2023). Further, although recent works substantiate the assumption that our findings still apply (Hilbert et al., 2024; Sajjadian et al., 2021; Vieira et al., 2022), features such as neuroimaging or biological data are not considered in the current study. The same limitation applies to pre-processing steps and model choice, including more sophisticated neural networks than the MLP used. Fourthly, while using the elbow method allows an analytical approach to determining convergence, it does not consider the trade-off of the cost that additional data points induce. Further, oscillations can influence elbow points, though mitigated by choosing the global instead of local elbow point.

In terms of recommendations, we, firstly, strongly discourage mistaking CV or, even less so, single test set results for suitable performance measures on small data set sizes (N=100-300). Doing so exacerbates publication bias and causes ungeneralisable result expectations (Andaur Navarro et al., 2021; Chekroud et al., 2024; Hilbert et al., 2024; Zantvoort, Hentati Isacsson, et al., 2024). A key step against overfitting is separating the validation set for the hyperparameter decision from the model choice, for example, through nested CV (Bates et al., 2022). Further, especially for complex features or ones with unknown or low information value, having a reasonably sized test set is indispensable (Beleites et al., 2013; Chekroud et al., 2024). Based on our results and previous suggestions (Beleites et al., 2013), we, therefore, propose a minimal data set size of N=500 for predictions in DMHIs to mitigate overfitting.

Secondly, even though N=500 started producing reliable results, it did not suffice to approximate many of our feature groups' maximum predictive power. Performance did not converge until N=750 for LR, SVM and NB, and for the more flexible models,

it even required N=1,000-1,500. Further, the predictive power of additional and mixed features increased in higher data set sizes. We, therefore, suggest N=1,000 as a minimal data set size when comparing simple to more complex feature groups.

Lastly, and closely related to the other points, we recommend being mindful of the interaction between the nature and number of features, data set sizes and models. While ML methods can theoretically handle many features, for small data set sizes, the noise of additional features and the models' ability to overfit it must be considered (Atla et al., 2011; Kwon & Sim, 2013; Nettleton et al., 2010; Saseendran et al., 2019). Further, the hyperparameters, especially those concerning regularisation, need to be chosen accordingly. To determine the adequacy of the set-up, we suggest implementing and reporting a learning curve approach leading up to the maximum available data set size. On the one hand, a downward CV trajectory suggests substantial overfitting. On the other hand, a continuously steep upward trajectory of both CV and test results suggests an underestimation of the predictive power due to a lack of data.

In summary, this paper contributes to the field of research by providing insights to aid the design and interpretation of predictions in DMHI settings. As such, it aims to combat unrealistic result expectations and the consequent disenchantment in a field where AI can be of great value but is only gradually gaining a foothold.

4 Methods

4.1 Case Study Background – everyBody Study

The everyBody dissemination study (ISRCTN13716228) provided evidence-based eating disorder (ED) prevention and health promotion programs (Beintner, Emmerich, et al., 2019; Jacobi et al., 2007, 2012, 2022) in Germany (Nacke et al., 2019). Participants (N=3,654) were adult women without full-syndrome EDs recruited from the general population between November 2016 and May 2019. All participants gave informed consent to participate in the study, and participation was anonymous. This primary study was a stratified, nonrandomised, parallel-group interventional design where intervention content matched risk and symptom levels. From the total sample, 452 users were allocated to the Basic intervention, 397 to Original, 1,386 to Plus, 80 to AN, and 1,339 to Fit. The interventions comprised 4 to 12 weekly online sessions (20 to 60 minutes) based on cognitive-behavioural principles, including psychoeducation, exercises to promote body image and balanced eating, and—if applicable—to reduce ED symptoms. Four out of five interventions were supplemented with daily or weekly online diaries. Four interventions had access to moderated peer group discussions, and two included weekly coach feedback messages.

Questionnaires were completed at screening, baseline, mid-intervention, post-intervention, 6-month, and 12-month follow-up. Analysis of pre-post changes of weight-related concerns within the completer subset revealed notable decreases in weight-related concerns across four of the five study arms (effect sizes d = -0.45 to d = -0.94) (Nacke et al., 2024).

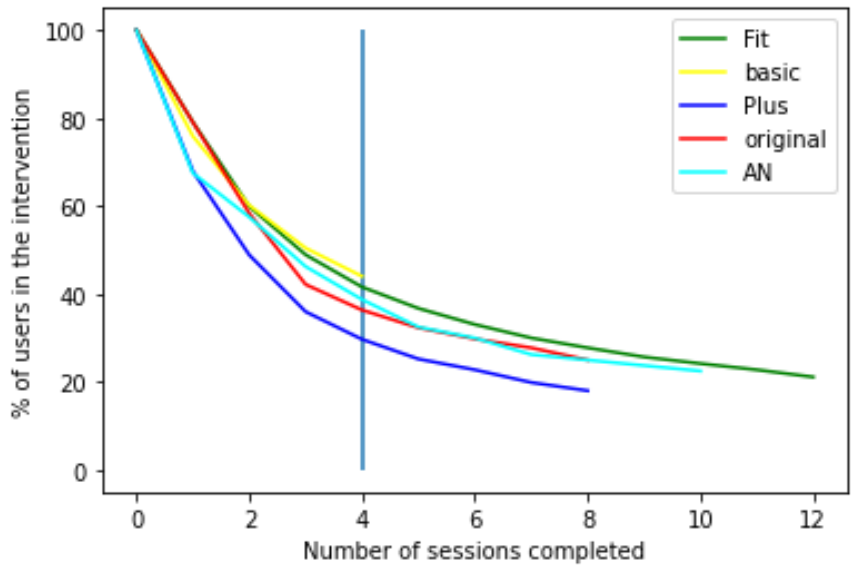
The screening and allocation process, individual intervention design and data generation is described in detail in Supplementary Note 1. Additional information can be found in the pre-registration protocol of the study (Nacke et al., 2019) and its primary publication (Nacke et al., 2024).

4.2 Definition of Outcome

Session completion was chosen to operationalise dropout, as it was found to be the most closely connected to intervention outcome (Donkin et al., 2013). While the different interventions had variable numbers of sessions (4-12), they presented similar dropout patterns, as seen in Figure 7.

Therefore, completing less than four sessions was defined as dropout to account for the minimum length of four weeks in the shortest intervention. This definition led to 56% dropout in the Basic intervention, 64% in Original, 70% in Plus, 61% in AN, and 58% in Fit. While many other dropout definitions are possible (Beintner, Vollert, et al., 2019), this operationalisation presents the possibility of identifying the users most at risk of leaving across interventions while ensuring sufficient time left to intervene (Forsell et al., 2019).

Figure 7: Dropout curves per intervention arm with cutoff



Notes: Dropout curves per intervention arm defined by the share of users that finished each number of sessions across intervention, i.e., Fit (green), basic (yellow), Plus (dark blue), original (red), AN (turquoise). Vertical blue line indicates the cutoff point of four models,

4.3 Feature Groups and Pre-Processing

The most common overarching categories for dropout predictors are questionnaire data and intervention user behaviour data (Bremer et al., 2020; Zantvoort et al., 2023). For the current study, feature groups were categorised based on the number of features and their empirically proposed predictive power regarding dropout. The categories

considered and their key details are shown in the overview in Table 3 and briefly described in the text below. Across all feature groups, the basic information of which intervention the user participated in, its lengths in weeks, and the starting year was also added.

The translated original questions and units can be found in Supplementary Table 1. An overview of the almost 200 features' description including their number of missing values is provided in Supplementary Note 2. All data processing was done in Python, primarily relying on the NumPy (Harris et al., 2020) and Pandas (McKinney, 2010) libraries. Missing values were imputed with a multivariate iterative imputer (Roderick & Rubin, 2002) using the training sets questionnaire and weekly aggregated user behaviour variables described below.

Table 3: Overview feature groups

Name	Description	Key Aspects	#
Simple Questionnaire	Primary symptom scores (WCS) (Killen et al., 1994) at screening and baseline	Assumed low predictive power (Bricker et al., 2023; Günther et al., 2023; Linardon et al., 2022; Zantvoort et al., 2023), available before intervention start	2
Extended Questionnaire	Variety of self-report questionnaires incl. WCS (Killen et al., 1994) and further eating disorder (Fairburn & Beglin, 2008; Tylka, 2006), depression (Kroenke et al., 2001), and anxiety (Spitzer et al., 2006) symptoms and behaviour patterns, personality (Rammstedt et al., 2012), self-regulation (Carey et al., 2004) and self-esteem scores (Rosenberg, 1979), psychiatric and weight loss history, alcohol use (Bush et al., 1998), socio-demographic information, and user expectations.	Assumed low predictive power (Bricker et al., 2023; Günther et al., 2023; Linardon et al., 2022; Zantvoort et al., 2023), theoretically available before intervention start but with high time-invest from users	51
Simple User Online Behaviour	Sum of logins per day of the first week	Assumed high predictive power, very simple to obtain	7
Selected User Online Behaviour	Single aggregation for the first week of time to complete sessions, seconds spent, number of logins, number and length of answers, diary entries and messages to coaches and group chat	Assumed high predictive power (Bremer et al., 2020; Hentati Isacsson et al., 2023; Zantvoort et al., 2023; Zantvoort, Hentati Isacsson, et al., 2024) with effort into researching and choosing most promising options and aggregation measures	13
Extended User Online Behaviour	Variables from log files aggregated per day of the first week, incl. sessions completed, seconds spent, log ins, time spent in beginning/ mid/ end of the week and morning/ day/ evening, session completion, count and number of characters of diaries, group, and coach messages, exercises, answers to the sixteen most common closed questions as mean, min and max	High predictive power but possible loss due to complexity (Bremer et al., 2020; Hentati Isacsson et al., 2023; Pedersen et al., 2019), automatically collected during first week of intervention with limited time invest	129
Mixed Features	Extended questionnaire + selected user online behaviour	Mixed, with reported increase of predictive value (Zantvoort et al., 2023)	64

Questionnaire Data: For the primary dissemination trial, various questionnaire data were collected before intervention start, ranging from the standard primary symptom data up to less common measures such as personality scores. As pre-intervention questionnaire data has limited predictive power regarding dropout by itself (Bricker et al., 2023; Günther et al., 2023; Linardon et al., 2022; Zantvoort et al., 2023), it was used to investigate a low predictive power setting.

For the *simple questionnaire data*, only the screening and baseline primary symptom questionnaires (Weight Concern Scale (Killen et al., 1994)) were used. For the

extended questionnaire data, another 49 measures on psychological symptoms and characteristics, socio-demography, and user expectations were chosen based on their availability from the primary study and assumed usefulness. As a result, missing data was minimal, with five variables with <1.5% and six variables with <15% missing entries. The six latter were voluntary, and most users either answered all or none. Therefore, an additional variable was added to indicate this choice.

User Intervention Behaviour Data: For the intervention user behaviour data, log files and user submissions were aggregated into a set of simple, selected, and extended features. Only data from the first week of the intervention was used to leave sufficient time to intervene against dropout. The *simple behaviour data* followed related work on generalisable features in DMHIs and counted the users' number of logins per day for the first week of the intervention (Bricker et al., 2023; Cote-Allard et al., 2022). For the *selected user behaviour,* features were selected based on the related work (Bremer et al., 2020; Hentati Isacsson et al., 2023; Hornstein et al., 2021; Zantvoort et al., 2023; Zantvoort, Hentati Isacsson, et al., 2024) and aggregated per week, mitigating sparsity, multicollinearity, and complexity. For the *extended user behaviour,* the same raw data instead was separately aggregated per day and included additional less known or theoretically less informative features as well as more aggregation forms (e.g., mean, minimum and maximum).

Mixed Features: To consider possible interaction effects between the two types of features (Zantvoort et al., 2023), the selected behaviour and extended questionnaire data were added together for the last group.

4.4 Algorithms

Six common ML algorithms (Fernandez-Delgado et al., 2014; Hastie et al., 2017) were included in a trade-off of investigating different models while maintaining a reasonable computational load and ability to present results. For the simple algorithms, Naïve Bayes (NB) (Zhang, 2004), Logistic Regression (LR), and Support Vector Machines (SVMs) (Cortes & Vapnik, 1995) with a linear and radial kernel option and classifier were trained. In terms of more sophisticated tree-based models, first, Random Forest (RF) models were used due to their high flexibility and good performance in similar settings (Fernandez-Delgado et al., 2014; Zantvoort et al., 2023; Zantvoort, Hentati Isacsson, et al., 2024). Second, to leverage the upsides of sequentially combining several tree learners, adaBoost decision trees were included. Lastly, a Multilayer Perceptron covered the family of deep neural networks. These models have been extensively discussed in various sources (Hastie et al., 2017; James et al., 2021) and will, therefore, not be further detailed here.

4.5 Learning Curves and Training Set-up

To estimate training performance, 10-fold cross validation (CV) with grid search was implemented. The best resulting estimator was re-trained on the entire training dataset and evaluated on the previously set aside test set of 20% of the data. A standard scaler was incorporated into the pipeline. Regarding the hyperparameter ranges, default

values were expanded upon if the outermost values appeared insufficient or excessive within the training data results.

Following authors such as Giesemann et al. (2023), Balki et al. (2019), and Perlich, Provost, & Simonof (2004), learning curves were used to provide insights into the effect of sample size on prediction performance. For the data set sizes, the space of 100, 200, 300, 400, 500, 750, 1,000, 1,500, 2,000, 2,500, 3,000, and 3,654 was explored to balance a comprehensive investigation with computational costs. Samples were stratified for dropout, and the models were trained on 80% of the respective N to represent the data set sizes. Further, training was repeated on samples drawn with different seeds ten times for small data set sizes (≤500), five times for the mid data set sizes (≤2,000), and three times for the remaining large dataset sizes (Giesemann et al., 2023). The model training was implemented with the scikit-learn (Pedregosa et al., 2011) library in Python, and the code is publicly available in this paper's GitHub repository.

4.6 Evaluation and Result Analysis

The area under the curve (AUC) score was used to compare results across all settings without depending on a threshold. In terms of evaluation, the scores were classified into no (0.50-0-0.56 AUCs), low (0.57-0.64), moderate (0.65-70), good (0.71-0.75) and very good (>0.75) predictive power (Kraemer et al., 2003). Predictive power per feature group was approximated through the test score for the model type with the highest training scores at N=3654. A two-tailed DeLong test (DeLong et al., 1988; Sun & Xu, 2014) with a significance threshold of $\alpha=0.05$ was used to compare the test AUCs between models. The DeLong test was chosen because it is non-parametric, aimed at comparing AUCs and sufficiently computationally efficient (Sun & Xu, 2014). The test returns the p-value for the null hypothesis of equal performance, hence the assumption that no model performs better than the other. Failing to reject the null hypothesis ($p>0.05$) leads to possible differences in the AUC being assumed to be due to random chance.

The variability of results was determined through the standard deviation of single validation results across repetitions. To determine overfitting, first, the difference between the mean training and test score was considered. Next, the per centage of CV scores at least +0.10 AUC higher than the mean test set was investigated. The threshold 0.10 was chosen as it is a step that definitively jumped one results categorisation introduced above, meaning, for example, a "*low*" score would become "*good*". Performance convergence was investigated by considering the diminishing marginal benefit of adding more data through the so-called elbow method. To this end, the kneed algorithm (Satopaa et al., 2011) Python implementation was used and set to find the global convergence point.

Chapter III: Intervention Data Pooling

Zantvoort, K., Hentati Isacsson, N., Funk, B., & Kaldo, V. (2024). Data set size vs homogeneity – A Machine Learning study on pooling intervention data in E-Mental Health dropout predictions. SAGE Digital Health, 10, 1–11.

Objective: This study proposes a way of increasing dataset sizes for Machine Learning tasks in Internet-based Cognitive Behavioural Therapy through pooling interventions. To this end, it (1) examines similarities in user behaviour and symptom data among online interventions for patients with depression, social anxiety, and panic disorder and (2) explores whether these similarities suffice to allow for pooling the data together, resulting in more training data when prediction intervention dropout.

Methods: A total of 6418 routine care patients from the Internet Psychiatry in Stockholm are analysed using (1) clustering and (2) dropout prediction models. For the latter, prediction models trained on each individual intervention's data are compared to those trained on all three interventions pooled into one dataset. To investigate if results vary with dataset size, the prediction is repeated using small and medium dataset sizes.

Results: The clustering analysis identified three distinct groups that are almost equally spread across interventions and are instead characterized by different activity levels. In eight out of nine settings investigated, pooling the data improves prediction results compared to models trained on a single intervention dataset. It is further confirmed that models trained on small datasets are more likely to overestimate prediction results.

Conclusion: The study reveals similar patterns of patients with depression, social anxiety, and panic disorder regarding online activity and intervention dropout. As such, this work offers pooling different interventions' data as a possible approach to counter the problem of small dataset sizes in psychological research.

1 Introduction

Modern societies struggle to provide adequate mental health care (Becker et al., 2018; Ebert et al., 2019; Lancet Global Health, 2020), as traditional therapy alone is not meeting the increasing need (Lamo et al., 2022). Internet-based Cognitive Behavioural Therapy (ICBT) promises to improve care levels by achieving similar goals as face-to-face therapies through efficient digital means (Becker et al., 2018; Cuijpers et al., 2014). With the rise of ICBTs, a large variety of user online activity data becomes recordable. Applying advanced analytics to this data holds great promise to individualise and improve care (Bremer et al., 2020; DeMasi et al., 2017; Hornstein et al., 2023). One task that presents itself to be solved with Machine Learning (ML) is that of intervention dropout predictions (Bremer et al., 2020; Lamo et al., 2022). A patient dropping out of an intervention is significantly less likely to have positive outcomes (Donkin et al., 2011). Yet, upon starting the intervention, costs occur, and scarce resources are occupied (Kaltenthaler, Sutcliffe, et al., 2008). Measures such as guidance from therapists lower dropout rates (Baumeister et al., 2014), but are often too costly to be provided to all patients (Forsell et al., 2022). Identifying patients at risk of dropout early on allows for the personalized allocation of measures. If more individuals' needs for support are met, resource allocation is optimised, and health benefits increase (Forsell et al., 2019). In contrast to the direct prediction of health outcomes, dropout predictions include patients who otherwise tend to be ignored due to missing symptom data (Barrett et al., 2008; Wu et al., 2022).

Initial studies using ML to predict dropout based on user behaviour data show promise (Bremer et al., 2020; Cote-Allard et al., 2022; Linardon et al., 2022; Moshe et al., 2022; Pedersen et al., 2019; Smink et al., 2021; Wallert et al., 2018). Nevertheless, there are still few ML applications in ICBTs, especially dropout predictions (Bzdok & Meyer-Lindenberg, 2018; Lee et al., 2018; Moshe et al., 2022; Pedersen et al., 2019; Symons et al., 2019). A recent systematic literature review found that, despite the widespread consensus about its value, only three out of 94 digital interventions use ML to personalize interventions for depression (Hornstein et al., 2023). This scarcity of work is attributed to the small size of available data sets for training (Bzdok & Meyer-Lindenberg, 2018; Hornstein et al., 2023), as it limits the accuracy and generalisability of predictions (Dietterich, 1998; Lateh et al., 2017; van Smeden et al., 2019). Collecting large health data sets is costly (Pasini, 2015), and the median data set size across 59 studies using ML for outcome predictions in depression treatments was found to be only 115 patients (Sajjadian et al., 2021). Similarly, the median for related work predicting dropout in online interventions was 342 patients (Bremer et al., 2020; Cote-Allard et al., 2022; Linardon et al., 2022; Moshe et al., 2022; Pedersen et al., 2019; Smink et al., 2021; Wallert et al., 2018). Thus, with ML approaches in ICBTs, the question arises of how to produce accurate predictions despite small data sets (Aafjes-van Doorn et al., 2021; DeMasi et al., 2017; Symons et al., 2019).

Albeit the lack of large data sets, the number of smaller data sets available was already reported to be in the hundreds five years ago (Carlbring et al., 2018). Data sharing between providers creates larger training data sets but poses significant challenges

regarding data privacy, data interoperability, and conflicting interests (Loftus et al., 2022). However, many providers themselves offer similar interventions for different but related primary symptoms, such as depression and anxiety disorders (Smink et al., 2021). As symptoms and behaviour are interconnected (Beard et al., 2016), common behavioural patterns between patients could be leveraged by pooling several interventions into one data set. If successful, this not only improves prediction results but also lowers model maintenance efforts. However, pooling the data may be detrimental if contexts and user behaviours differ significantly. Most papers predicting dropout focus on a single intervention (Bremer et al., 2020; Moshe et al., 2022; Pedersen et al., 2019; Smink et al., 2021; Wallert et al., 2018), while one gathers two different interventions for the same symptoms (Linardon et al., 2022), and one gathers interventions for three different symptoms (Cote-Allard et al., 2022). While this shows that different options are possible, no study comparatively evaluates the approaches. It, generally, remains unclear how patients with different but related target symptoms vary in their online intervention behaviour. Clustering analyses have successfully identified user archetypes in mental health apps in general (Aziz et al., 2023) and within specific interventions (Chien et al., 2020). Applying such a clustering analysis to users of different but related ICBTs could offer valuable insights into the similarities in user behaviour.

In this study, we use demographic, symptom, and online user behaviour data of 6,418 routine care patients from the Internet Psychiatry in Stockholm, Sweden. The main research question is if the value of pooling intervention data for social anxiety, depression, and panic disorder outweighs the downside of losing data homogeneity when predicting intervention dropout. The goal is to identify patients who will end up leaving the intervention early without benefiting, already in intervention week four of 12. For this, four different supervised ML methods (i.e., Logistic Regression, Support Vector Machines, Random Forest, and AdaBoost classifiers) are compared. We further investigate the relationship between prediction performance and data set size. To this end, the training on individual versus pooled data sets is repeated on samples of the median data set sizes of related work. A clustering analysis is the intermediate step to understanding the differences and similarities between the interventions' data. Through these proposed steps, this study aims at 1) exploring the heterogeneity of online intervention data across three highly prevalent mental disorders and 2) providing insights into what pooling these different intervention data sets into one data set yields for dropout prediction. Finally, this work adds to the limited body of research investigating dropout predictions across three large routine care interventions.

2 Methods

2.1 Interventions and Participants

This study uses routine care outpatient data from the Internet psychiatric clinic in Stockholm, Sweden from 2008 to 2020 (Titov et al., 2018). The data comprises all available patients undergoing treatment while the platform in question was used. The data sets consist of 1633 panic disorder (PD), 1907 social anxiety disorder (SAD), and 2902 major depressive disorder (MDD) patients. The treatments have previously been

evaluated with positive results (El Alaoui et al., 2015; Hedman et al., 2013, 2014). After a psychiatric assessment, each patient received 12 weeks of disorder-specific intervention. The assessment and treatment evaluation are based on established patient self-rating measures; For PD, the Panic Disorder Severity Scale-Self Report (Houck et al., 2002); for SAD, the Liebowitz Social Anxiety Scale-Self Report (Baker et al., 2002), and for MDD, the Montgomery–Åsberg Depression Rating Scale-Self Report (Montgomery & Asberg, 1979; Svanborg & Asberg, 1994) were used. Each of the three interventions was administered in the same clinical context where patients self-refer and then went through a web-based screening and the established semi-structured M.I.N.I diagnostic assessment interview (Sheehan et al., 1998) with a psychiatrist. During the interview, the clinician provides information about (I)CBTs in general, and the expected effort required from the patient to complete treatment. As such, all patients are ensured to have a relevant diagnosis as well as being informed about the treatment, and sufficiently motivated to engage in it.

Included patients receive the same therapist support routine across all three interventions. The interventions reside on the same technical platform and are very similar in structure, but clearly differ in therapeutic content. All interventions start with psychoeducation and consist of cognitive behavioural therapy techniques specific for each condition divided into 10 modules. For example, the SAD and PD interventions include symptom-specific exposure exercises, whereas the MDD has a focus on behavioural activation. Each intervention's modules consist mainly of exercises, including homework, messages from and to a therapist and weekly symptom assessments. More detailed information about the interventions is summarised in Appendix 1 and has previously been published for social anxiety (El Alaoui et al., 2015), depression (Hedman et al., 2014), and panic disorder (Hedman et al., 2013).

2.2 Features

The prediction is based on the first four weeks of data as a trade-off between gathering sufficient data versus maintaining sufficient time to intervene to prevent dropout (Bremer et al., 2020). A previous study has shown personalizing the support level in week 4 to have a positive effect on patients at risk (Forsell et al., 2019). For the features, thus, all data gathered after week 4 is disregarded to prevent target leak. First, we include the common socio-demographic variables, age and gender (Moshe et al., 2022). Second, given their importance (8,44), the symptom measures at screening and at the beginning of weeks 1, 2, 3 and 4, respectively are included. In addition to the actual scores, the time needed to fill out the questionnaires is included. Third, we use the basic information of the intervention set-up (i.e., year of start, start in winter or summer, and target disorder) (Moshe et al., 2022). Fourth, the character length of the homework assignments and messages are each summed per week, which has previously been found to be predictive of both dropout and health outcomes (Zantvoort et al., 2023). To account for the therapist messages, the per centage of characters sent in the conversation produced by the therapist is included. Following the insights from Bremer et al. (2020), several features are generated from patients' log data (Bremer et al., 2020). This includes the sum of time spent on the intervention, the number of pages clicked, the number of sessions, and number of days that patient logged in. In addition, the time patterns of the login behaviour are gathered across all weeks by looking into

the per centage of sessions per weekdays and daytimes. Further, we record how many days a patient needed to finish each module. Further information on the pre-processing and feature engineering steps can be found in Appendix 2.

The operationalisation of the dropout variable is aimed at identifying the patients who are most likely to leave the intervention too early to sufficiently benefit (Beintner, Vollert, et al., 2019). The intended symptom improvement is determined by either the final symptom score below the absolute cutoff for remittance (8 for PDSS-SR (Furukawa et al., 2009), 35 for LSAS-SR (von Glischinski et al., 2018), and 11 for MADRS-S (Fantino & Moore, 2009)) or a 50% improvement since the start of treatment (Karin et al., 2018). If no symptom score after week 8 is available, their symptom improvement is classified as "*unknown*". Module completion has been found to be the adherence measure with most consistently positive power towards explaining therapy outcome (Donkin et al., 2013). For this study, we use module eight of ten as it contains all unknown leavers, includes all of the content introductions, as the last two modules are repetition and maintenance (52), and produces considerably balanced classes. A more detailed explanation of the dropout variable can also be found in Appendix 2. The averages, SD and units of the resulting data set can be found in Appendix 3. All steps, including the subsequent modelling, are implemented in Python, using the pandas (McKinney, 2010), Numpy (Harris et al., 2020), Kneed algorithm (Satopaa et al., 2011) and Scikit-learn (Pedregosa et al., 2011) libraries.

2.3 Exploration of Heterogeneity Between Interventions

This analysis addresses the first goal; The exploration of heterogeneity in patients across the three interventions using the demographic, intervention, symptom, online activity, and character counts variables. The general purposes of clustering are to gain insight into the data, identify natural groups, and be able to summarise them based on segment prototypes (Jain et al., 1999). First proposed in 1967, k-means algorithms are among the most used clustering approaches due to their computational efficiency and easy implementation (Jain et al., 1999; Sinaga & Yang, 2020). The k-means algorithm is an optimisation algorithm that iteratively finds a set of k centroids, such that the total sum of distance between each point and its nearest centroid is minimized (Hastie et al., 2017). As such, it optimises for groups that are as similar as possible within themselves but as different as possible from each other. (Hastie et al., 2017). The number of clusters needs to be decided apriori and is inferred using the Elbow (or Knee) Method (Bholowalia & Kumar, 2014). In essence, this method uses the explained data variance to reveal where the marginal gain of a new cluster is outbalanced by the increasing number of clusters. As commonly done, a Principal Component Analysis (PCA) is conducted prior to the clustering to lower the number of dimensions, reduce multicollinearity and facilitate the visualisation of clusters (James et al., 2021). The number of principal components is also automatically identified by using the Kneed algorithm (Satopaa et al., 2011). The critical question is whether the algorithm will rely on the target disorder variable identifying each intervention as primary splitting criteria. If the different target symptoms result in different online behaviour, the clustering algorithm would be expected to reproduce the three intervention groups. Only if patients behave sufficiently similar across interventions, mixed clusters can be expected. To

ensure comparability, this process is done only on the first four weeks of data also available to the prediction models. This excludes the dropout and health outcome variable for all patients, which will, however, be added after completing the clustering for the evaluation of clusters.

2.4 Dropout Prediction

The second goal is investigating the effects of pooling patients from interventions for depression, social anxiety, and panic disorder to one data set when predicting dropout. To this end, models trained on each of the interventions' data individually are compared to models trained on all three interventions pooled. As the data has almost twice as many MDD as PD patient, we add a pooled run where we under sample the large interventions to have balanced ratios of one-third each. The training data set sizes at hand are in the four digits, which is already unusually big (Sajjadian et al., 2021). To increase the usefulness of results for future studies, the training process is repeated on smaller samples - The median data set sizes of related work for outcome prediction (115) (Sajjadian et al., 2021) and dropout prediction (342) (Bremer et al., 2020; Cote-Allard et al., 2022; Linardon et al., 2022; Moshe et al., 2022; Pedersen et al., 2019; Smink et al., 2021; Wallert et al., 2018). Taking away an assumed 15% data for a holdout test set results in training data of 98 and 291 patients per single intervention.

While the training data differs in data set size, all models are evaluated on the same 20% stratified test set to maintain comparability across runs. The final evaluation is done 1) relatively and 2) absolutely, both focusing on balanced accuracy (BACC). For the relative evaluation, the single vs pooled data results are compared. For the absolute comparison, the benchmark of 1) better than chance and 2) 67% balanced accuracy as minimally necessary to be valuable in an ICBT to adapt treatment as proposed by Forsell et al. (2022) are used (Forsell et al., 2022). Further, to adhere to the standards of medical ML studies, accuracy, balanced accuracy, specificity, recall, and Area Under the Curve (AUC) are provided for the test set performances (Cabitza & Campagner, 2021).

In terms of algorithms, Logistic Regression (LR), Support Vector Machines (SVMs) (Cortes & Vapnik, 1995), Random Forest (RF) and AdaBoost (Schapire, 2013) classifiers are chosen as they cover a range from simple and robust to more sophisticated and flexible options (Aafjes-van Doorn et al., 2021; James et al., 2021). As extensively argued (Bates et al., 2022; Cawley & Talbot, 2017), choosing the algorithm to use in the same step as optimising the hyperparameters comes with a significant risk of overfitting. Therefore, the model selection will be done through 5x10 nested-cross validation (CV). The inner cross validation optimises hyperparameters via grid search, and the outer CV score determines the one algorithm to use. That algorithm is then retrained on the whole training data with a 10-fold CV, returning a single model to be evaluated on the test set. An intervention-based scaler is added to the pipeline, such that a standard scaler is fitted to the training data per intervention and then applied on the respective holdout fold (Cabitza & Campagner, 2021).

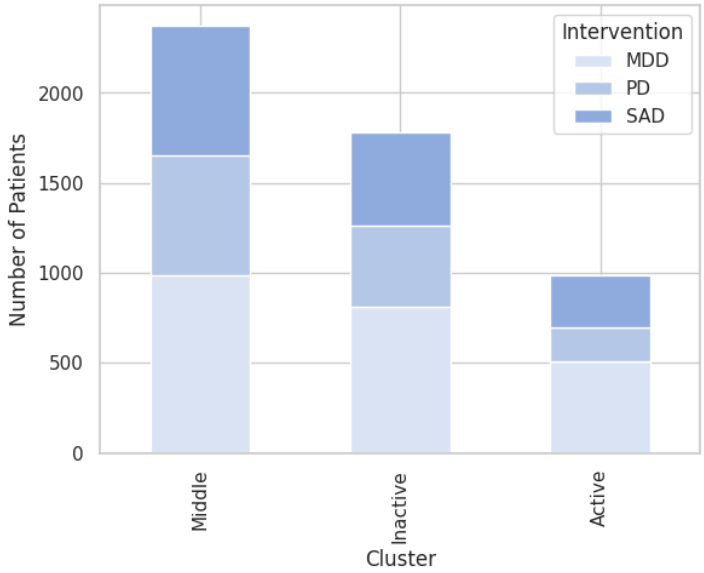
To choose the range of hyperparameters, initial values are run and added to if the outer points seem too low or high. For the algorithms that allow for balancing class weights, the class weights are balanced. For the LR, the choice of L1 and L2 feature selection is optimised as a hyperparameter for the liblinear solver (Fan et al., 2008). The C value is searched across the range [0.001, 0.01, 0.05, 0.1, 0.20, 1]. The SVMs optimise over an RBF and a linear kernel with respective C values [0.001, 0.01, 0.1, 0.25, 0.5, 1]. The RF model searches across the number of estimators [5, 10, 25, 50, 500, 1200], the minimum samples [10, 25, 50, 100, 200], the maximum depths [5, 25, 50, 100, 500, 750], and a binary indicator for bootstrapping. Lastly, the AdaBoost Classifier trades off the number of estimators [1, 2, 5, 10, 25, 100, 1500] with their respective learning rate [0.001, 0.01, 0.1, 1, 2, 2.5].

3 Results

3.1 Data Heterogeneity

To better understand the differences in online behaviours, user characteristics and symptom patterns, the 1631 PD, 1906 SAD, and 2881 MDD patients and their 57 input variables that result from pre-processing as described in Appendix 2 are clustered. The kneed method suggests four principal components to represent the input data. Feeding these components into the k-means algorithm with k ranging from one to 11 suggests three most prevalent clusters. However, this optimal value only coincides with the number of interventions, as each intervention spreads comparatively evenly across clusters (Figure 8). The biggest intervention group, MMD, makes up 42-51% of each cluster, SAD accounts for 29-30%, and the smallest intervention, PD, spreads at 20-28% per cluster. To facilitate understanding, the clusters will be referred to as active, middle, and inactive clusters from now on for the reasons explained below. The middle cluster is by far the biggest as it contains 46% of all patients, with the inactive cluster following at 35% and the very active cluster tailing at 19%. All cluster means reported in this section can also be found in the overview table in Appendix 3.

Figure 8: Distribution of patients per intervention and cluster



Notes: PD = Panic Disorder, MDD = Major Depression Disorder, SAD = Social Anxiety Disorder

Inactive patients are more than six times as likely to have missing symptom scores (0.84/5) in the first four weeks as active patients (0.13), who are similar to the middle cluster (0.16). Further, the average lengths of messages and homework in the first weeks are six times as high for the active cluster (459/ 1331 characters) as for the inactive cluster (81/ 224 characters), with the middle clusters averaging at 157/ 1008 characters. Similarly, login data such as sessions, pages and duration per week are almost all 2-4 times as high for active as inactive patients, with the middle cluster somewhere in-between. The most extreme differences are in the durations where inactive patients, on average, spent one-third or -fourth of the time (12,071, 6698, 6198, and 6878 seconds per week 1-4) that active patients (32,393, 25,850, 24,738, and 24,171) spent. This is also reflected in the average number of modules completed in the first four weeks, with 2.7 for inactive, 4.2 for the middle and 4.7 for the active patients.

While the starting symptom scores barely differ, inactive patients have a higher average symptom improvement between screening and treatment start (-12%) than middle (-9%) or active (-8%) patients. However, this changes in the following weeks.; While inactive patients still see 12% improvement in week 2, they have next to no change (0%, -1%) in week 3 or 4. While the strength of change also lessens for middle and active patients, they still continuously improve (middle: -16%, -3%, -4% and active: -15%, -4%, -5%). The least differentiating variables are the start year, time variables (e.g., time and weekday of intervention use), symptom questionnaire duration, if they started the intervention in the winter, and the patient's age and sex. Retrospectively joining the dropout variable to the clustered data shows that patients from the active cluster are less than half as likely to drop out (25%) as patients from the inactive cluster (65%), with the middle cluster being closer to the active cluster (36%). Despite the dropout rates heavily differing across interventions (PD 28%, MDD 45%, and SAD

57%), the differences in dropout probability per cluster remain. Dropout ratios for the inactive, middle, and active groups are 68%, 36% and 24% for MDD, 41%, 23%, and 16% for PD, and 80%, 49%, and 33% for SAD. Doing the same for the health outcomes shows that 53% of active and middle patients are treatment successes while only 34% of inactive patients are. For 15% of inactive patients, their health outcome is unknown, whereas the middle cluster has 5% and the active cluster only 2%. As a result, the per centage of not successful treatments is close together between active (45%), middle (43%), and inactive (50%) patients.

3.2 Prediction

The train-test split leads to a maximum of 5132 training data points and 1289 test data points, for which the averages of all variables can be found in Appendix 3. Of these data points, 45% are MDD, 30% are SAD, and 25% are PD patients, resulting in the unbalanced pooled training data sets in Figure 9. The small and medium balanced pooled training data have the same total as the unbalanced run; however, they have balanced ratios of one-third per intervention. For the large data, balancing is dictated by the smallest intervention (PD), resulting in a sample size of 1304 each. The nine single intervention runs (three per disorder) with 98, 291, and 1304/ 1524/ 2303 are not separately shown in Figure 9.

Figure 9: Training and test data sample per pooled run

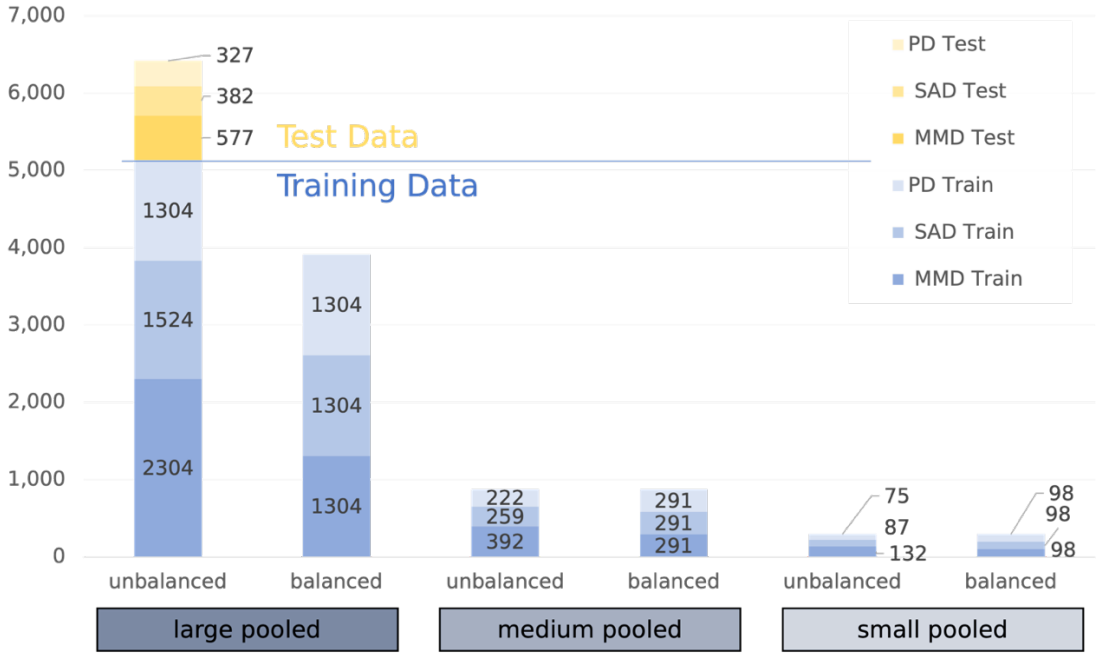
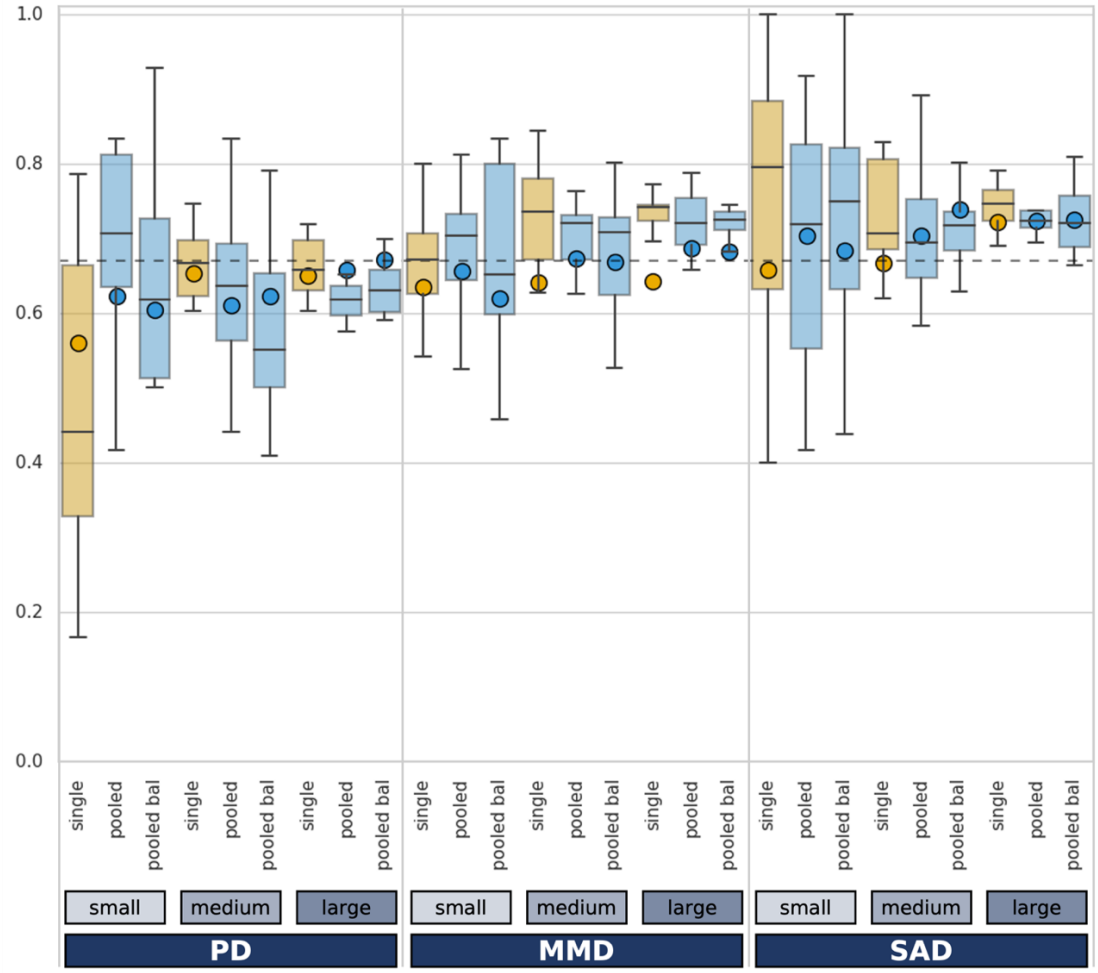


Figure 10 presents the BACC results for each intervention, data set size, and type of training data. The box plots show the 10 outer CV scores of the training data, while the single bullet point shows the performance on the test set. An ideal result graph has a high y-axis value (balanced accuracy) with a narrow boxplot (low variance in training results). The boxplot's horizontal stripe (median) should be close to the test set point (neither overfitting nor unexpectedly high results). For pooled data results in the following, the unbalanced (UB) results are mentioned first, followed by the balanced

(B) results. To answer the main research question, the single data set runs are compared to their respective pooled counterparts (e.g., 98 datapoints single intervention vs 3*98 = 294 datapoints pooled run). All results, including the numbers discussed here, accuracy, recall, specificity, AUC score, and the model type chosen in the CV can be found in the result table in Appendix 4.

Figure 10: Balanced accuracy for training and test results



Notes: Box plot of outer CV balanced accuracy of the training data with test balanced accuracy as single bullet point and clinical threshold as dotted grey line.

Training data results. Only looking at training results, single runs outperform the pooled data sets in predicting dropout. The single intervention runs have a higher median outer CV score than both pooled runs in six out of nine cases with an average advantage of 0.037 BACC. Further, single runs are higher than at least one pooled run in two more cases. Hence, the only exception is the small (98 data points) PD data set, where training results for the single run are lower than both pooled data BACCs. For MDD, the median CV score increases as the data set size increases from small to medium to large, with a total difference of +0.04 BACC. The opposite pattern is visible for PD, such that the training scores decrease as more data is added with -0.05 BACC. Similarly, SAD has the highest score for the small data but then has the smallest score for the medium and an average score for the large data with a total range of -0.08. Considering the range between the first (Q1) and third (Q3) quartile of the CV scores,

pooled runs have a lower average range than single interventions. The spread of the training results (Q3-Q1) is lowest for the large data sets at an average of 0.045 in BACC. For the small data sets, it is more than four times as much (avg. at 0.202), with the medium data set size in-between (avg. 0.101).

Test data results. Which setting (single vs pooled) performs best is reversed when looking at the test instead of training results. Here, in seven out of nine cases, both pooled runs outperform the single intervention. Additionally, the unbalanced pooled data outperforms the single intervention for the smallest MDD run (0.64 vs UB 0.66/ B 0.62). Hence, the single intervention is only superior in one case, the medium PD run. As such, PD patients have both one of the biggest gains and biggest loss (+0.063/-0.044 BACC) when using the pooled model instead of one trained on the PD patients alone. The runs with all available PD data barely differ (+0.007 for pooled). For SAD patients, the small data gains somewhat from pooling (UB +0.046/ B +0.026), the medium data set size has an average and the highest gain (+0.037/ +0.072), and there is barely any difference in the large data (+0.000/ 0.002). On the contrary, MDD patients almost always benefit from pooling the data, and the gains grow with the data set size (small data UB +0.021/ B -0.016, medium data +0.032/ +0.028, large data +0.045/ +0.040). Keeping the natural ratio is superior for all three small data runs and the medium and large data MDD runs. Balancing the data is favourable for all medium and large data sets of SAD and PD. The largest impacts of balancing the data are -0.037 BACC for the small MDD and +0.035 for the medium SAD data sets. For MDD and SAD, the pooled data's higher BACC also means a higher recall than the single intervention data sets. However, in the case of PD, the pooled data sets have high specificity (avg. 0.80) but lower recall (avg. 0.46) whereas the single intervention runs are more balanced (avg. specificity: 0.57, recall: 0.67).

Difference between training and test results. In terms of potential overfitting, the pooled runs have a smaller absolute difference between the test and training results in seven out of nine cases. The only exceptions are the medium and large PD data sets. This results in an average absolute train-test difference in BACC of 0.063 for single and much lower 0.034/ 0.037 (UB/B) for pooled data. The smallest data set of SAD has the largest gap (-0.14) with the highest training results out of all runs but the lowest test results of all SAD models. Similarly, the training and test results of the MDD models are twice as much apart for the single data as for the pooled data, and the pooled model achieves better test results in all cases. The only exception to this rule is the medium PD data set, where the single model achieves better and closer test and training results. However, if pooled runs outperform the single interventions in the training data, they consistently also outperform it in the test set.

Absolut evaluation. Regarding the absolute evaluation, all models achieve a balanced accuracy of more than 0.5 BACC, and are, therefore, better than chance. Further, 13 of the 27 models achieve a BACC on the test set of 0.67 or higher, with results differing across interventions and settings. For SAD, all but the single intervention small patient models achieve the threshold, with a maximum BACC of 0.74. For PD and MDD, not one model trained on the single intervention achieves clinically relevant prediction results. However, for MDD, both pooled model for the medium (0.67/ 0.67) and large

(0.69/ 0.68) data achieve the threshold. For PD, only the largest balanced pooled model achieves clinically relevant prediction results on the test set.

4 Discussion

Researching the heterogeneity in patient data for ICBTs for MDD, SAD, and PD, we found intervention-overarching patient groups in the first four weeks of the interventions. Despite differing dropout rates per intervention (28-57%), the algorithm identified the respective most likely and least likely clusters to dropout. The active, middle and inactive clusters' correspondence to low, middle and high dropout is in line with previous findings (Beintner, Vollert, et al., 2019). Our first finding, that SAD, PD and MDD patients have similar clusters of activity patterns may help the design and delivery of both individual and transdiagnostic interventions.

The answer to the first research question already hints towards the answer to the second; Pooling the data was almost always favourable and doubled the likelihood of achieving clinically relevant test results. Most noticeably, having 873 mixed intervention training data points outperformed having 2304 individual intervention MDD or 1524 SAD patients. A possible hypothesis for this is that pooling different interventions forces the model to focus on general patterns rather than intervention-specific noise. Beyond better results, pooling data comes with the upside of less resources necessary for deploying and maintaining one versus three models. PD patients' overall low results might partly be explained by their high class imbalance regarding dropouts and completers.

Two further interrelated key findings are the importance of independent test sets and risk of overfitting on small data sets. If the decision about whether to pool the data was made on the training CV scores, single intervention runs would have been preferred. Further, even with pooled data, in two out of three interventions the small data sets seemingly outperform the much larger datasets in the training score. This aligns with Sajjadian et al.'s (2021) findings that data set size is significantly negatively correlated to the reported prediction accuracy (Sajjadian et al., 2021). Our study's large test sets of 327-577 patients provides evidence that these good training results are biased as they fail to generalise. Sajjadian et al. (2021) further find that many studies do not even use an adequate training set-up, instead relying on a single train-test split (Sajjadian et al., 2021). As can be seen in the box plots, this can result in extremely high or low results, neither of which represent the expectable prediction performance. Making a deployment decision on such ungeneralisable training results comes with a myriad of problems: Risk of suboptimal care, wasted resources and ultimately the corrosion of trust in the use of ML in clinical care (Cabitza & Campagner, 2021; DeMasi et al., 2017; Sajjadian et al., 2021). As this paper shows, pooling different interventions enables providers to mitigate at least some of the risks when presented with the challenge of limited data availability.

The paper, thus, contributes to e-mental health care by exploring the trade-off between data heterogeneity and data set size and discussing the risk of overfitting (Sajjadian et al., 2021).

5 Limitations

At the same time, several limitations apply. For one, the routine care data in this study only includes self-referred patients, which leaves it unclear if the insights generalise to different patient selection methods. Further, it is yet to be investigated if the similarities between patients translate to the same clinical actions against dropout being effective. Third, using k-means for the clustering analysis is an industry standard (Jain et al., 1999), but generative (Hastie et al., 2017), or density-based methods (Kotu & Deshpande, 2014) may allow different insights. For the prediction task, the arguably biggest challenge in scaling the proposed approach is the availability of comparable interventions. While differing in content, the interventions at hand have a lot in common, the technical platform, the structure of treatments, the clinical routines for referral, assessment, therapist support, and the clinical staff. Therefore, our results do not warrant any prognosis about how the absence of these similarities would affect results. Lastly, this paper neither compares the gains of pooled data to other options such as federate learning (Loftus et al., 2022), nor offers definitive insights on what minimal data set size is necessary to produce generalisable results. In the end, pooling data in the proposed way is only one possible tool in the attempt to produce more generalisable and useful prediction models in psychological research.

6 Conclusion

Using ML to improve mental health care is a promising and growing research field. However, the lack of large data sets available hamper generalisability and cause biased results. This paper addresses this issue by investigating the effects of pooling data from different interventions together to increase the training dataset size available.

A total of 6,418 routine care patients' data from ICBTs for depression, social anxiety, and panic disorder is used to 1) investigate heterogeneity in patient online behaviour between interventions and 2) analyse the benefits of data pooling when predicting intervention dropout. Regarding the first question, the cluster analysis suggests three intervention-overarching groups that are defined more by their online behaviour and other clinical characteristics than by which ICBT-program they are in. The finding that patients across the three interventions have similar behavioural patterns is further supported in the prediction results. Ultimately, data pooling doubles the number of results that reach the threshold of clinical usefulness on the test set results. We, therefore, answer the second research question by concluding that data pooling is the superior approach based on our data set.

Chapter IV: Computer Mouse Trajectory

Zantvoort, K., Matthiesen, J., Bjurner, P., Bendix, M., Brefeld, U., Funk, B. & Kaldo, V. (2024). The Promise and Challenges of Computer Mouse Trajectories in DMHIs – A Feasibility Study on Pre-Treatment Dropout Predictions.

(Under consideration at Internet Interventions, submitted 3rd, September 2024)

Abstract: With the impetus of Digital Mental Health Interventions (DMHIs), complex data can be leveraged to improve and personalize mental health care. However, the majority of work relies on a limited number of often costly data types. Computer mouse trajectories are an unintrusive, cost-efficient and scalable data type that can be seamlessly integrated into current baseline processes. Empirical evidence suggests that how one moves their mouse holds information on motivation and attention, both valuable aspects otherwise difficult to measure at scale. Further, mouse trajectories can already be collected on pre-treatment questionnaires, making them a promising candidate for early predictions informing treatment allocation. Therefore, this paper discusses how to gather and process mouse trajectory data on questionnaires in DMHIs. Covering different levels of requirements, hand-crafted features analysed with a non-sequential model and the temporal-spatial raw mouse data analysed with a specific sequential neural network are proposed. The pipeline for the latter includes task-specific pre-processing to convert the variable length trajectories into a single prediction per user. As a feasibility study, we collected mouse trajectory data from 183 patients filling out a pre-intervention depression questionnaire. While the hand-crafted features slightly improved baseline predictions, the temporal-spatial model did not succeed. However, considering the evidence in contrast to our small data set size, we propose more research to investigate the potential value of this novel and promising data type.

1 Introduction

Digital Mental Health Interventions (DMHIs) are pivotal in expanding much-needed psychological treatment (Andersson et al., 2019). However, high dropout and moderate success rates make it unclear for whom this treatment form is best suited (Haller et al., 2023; Lipschitz et al., 2023). Machine Learning (ML) models show promise in individualising care, with most papers focusing on questionnaire or treatment data (Bricker et al., 2023; Cote-Allard et al., 2022; Prasad et al., 2023; Sajjadian et al., 2021).

Self-reported measures are a central information source in psychological research. However, they suffer from human bias in their design, selection and answer patterns (Breakwell et al., 2020; van Berkel et al., 2020). Further, adding measures increases patients' time requirements with an often limited marginal informational gain due to high multicollinearity (Patel et al., 2008; Sander et al., 2021; Tomitaka & Furukawa, 2021). Lastly, despite a plethora of attempts to predict or explain how treatment will go on pre-treatment questionnaire data only, no successful approach is known (Forsell et al., 2020; Gonzalez Salas Duhne et al., 2022; Günther et al., 2023; Haller et al., 2023; Linardon et al., 2022; Zantvoort et al., 2023). Other data types, such as medical imaging, genetic or heart rate variability data, have not shown much better results, and many of them are resource-intensive to obtain (Hilbert et al., 2024; Hornstein et al., 2022; Sajjadian et al., 2021). Identifying patients at risk before treatment starts would allow the allocation of more promising treatment, thus saving substantial resources for patients and providers (Günther et al., 2023; Kaltenthaler, Parry, et al., 2008). Further, even after treatment start, most ways of measuring relevant concepts such as adherence (e.g., time spent) are known to be noisy and poorly related to health outcomes (Donkin et al., 2011; Gan et al., 2021; Lipschitz et al., 2023).

Despite the patients' possibly lengthy decision processes and time on the page, so far, only the final answer to each questionnaire item or click off the page is recorded (Bremer et al., 2020). While methods such as eye-tracking have provided valuable insights into decision processes (Aimone et al., 2016; Gidlöf et al., 2013; Vachon & Tremblay, 2019), they are costly and difficult to scale to real-world settings. However, computer mouse cursor movements are highly correlated with gaze and can be unobtrusively and automatically gathered, making them a low-cost, scalable alternative (Chen et al., 2001; Milisavljevic et al., 2021).

Ample studies investigate the insights computer mouse dynamics offer into user's emotional state (Cepeda et al., 2021; Fu et al., 2017; Lepora & Pezzulo, 2015; Matthiesen & Holte, 2019; Tzafilkou & Protogeros, 2018; Yamauchi & Xiao., 2018; Zimmermann et al., 2003). Further, mouse trajectory data has been leveraged as a proxy for attention and motivation (dos Santos & Santana, 2022; Gledson et al., 2021; Leiva & Arapakis, 2020; Yusupova, 2021). Given this information value, mouse trajectories are a promising candidate for predictions in DMHIs. However, considering the complexity of temporal-spatial data, the question is whether and in what form information can be efficiently extracted from it. This article, therefore, discusses the collection of

mouse trajectory data in DMHIs and different ways of processing it (i.e., temporal, spatial and hand-crafted features).

To account for the requirements of DMHI settings where sequences are long and only one prediction per patient is needed, we propose a task-specific time-series pre-processing method. To illustrate the process along an exemplary use case, we leveraged mouse movement data from 183 patients filling out a pre-treatment depression symptom questionnaire to predict treatment dropout. This paper aims to introduce a novel data type in DMHIs and disseminate the mouse tracker and modelling code necessary to leverage it.

2 Mouse Trajectory Data

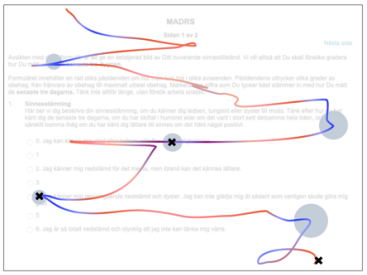
Whenever using a computer mouse, users move the cursor across the screen to reach the displayed option of choice. This dynamic decision-making process involves sensorimotor control subsystems to progress the hand towards the desired location (Yamauchi & Xiao, 2018). Previous research emphasised the cerebral connections between motor control and emotions, motivation and decision-making (Mendoza & Foundas, 2008; Mink, 2008; Wichmann & DeLong, 2014; Yamauchi & Xiao, 2018). As a result, computer mouse dynamics are influenced by users' mood (Zimmermann et al., 2003), state of confusion (Hucko et al., 2019), level of satisfaction or frustration (Cepeda et al., 2021; Matthiesen & Holte, 2019; Tzafilkou & Protogeros, 2018), stress and anxiety (Pepa et al., 2021; Yamauchi, 2013) and other emotional experiences (Yamauchi & Xiao, 2018).

2.1 Gathering Mouse Trajectories

Collecting the above-described information via the computer mouse does not require additional hardware and does not affect the user experience. A tracker to gather mouse trajectories on web applications is publicly available on GitHub[1]. It is based on JavaScript and PHP and can be implemented by inserting the initialisation script into any website. The tracker records a mouse cursor's position on the page (x- and y-coordinates) every 40 milliseconds until the page is closed or a pre-set limit of seconds has passed. The tracker further records scroll speed and event names, i.e., (1) whether the mouse is moving, 2) if the user is scrolling, and 3) clicking recorded as *mouse-up* and *mouse-down* events (see Figure 11). Next to the page coordinates, the client's screen coordinates are tracked to reflect the mouse's movement to the interface and the user's screen, which allows for normalisation.

[1] https://github.com/jjmatthiesen/evtrack/tree/setup_karolinskaInstitutet

Figure 11: Example of mouse data representation of a single user on interface



timestamp	x-coord (client)	y-coord (client)	mouseaction	scroll speed	
1612448647902	361	266	mousemove	0	movement
1612448647942	343	290	mousemove	0	
1612448647982	307	315	mousemove	0	
1612448648022	282	315	pause	0	pause
1612448648062	282	315	pause	0	
1612448648102	282	315	pause	0	
1612448648142	282	315	mousedown	0	click
1612448648182	282	315	pause	0	
1612448648222	282	315	mouseup	0	
1612448647783	282	315	scroll	16	scroll up & down
1612448647823	282	315	scroll	22	
1612448647863	282	315	scroll	-18	

Notes: User screen coordinates are not displayed

2.2 Processing of Mouse Data

Mouse trajectories are a time-ordered series of x- and y-coordinates such that the raw data holds the temporal-spatial information of a user's decision process. Three overarching ways of leveraging them are focusing 1) on the temporal sequence, 2) the spatial shape, or 3) aggregation to hand-crafted features.

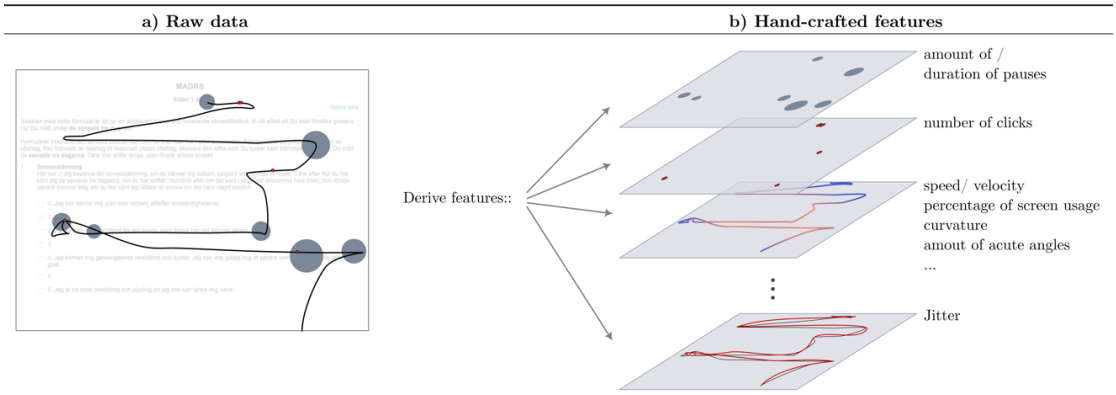
Firstly, mouse trajectory data are commonly modelled as a time series, hence, a sequence of coordinates representing the change of x and y over time. Accordingly, the models used are the likes of Long-Term-Short-Term memory models (Arapakis & Leiva, 2020) or one-dimensional convolutional neural networks (1D-CNN) (Antal et al., 2021; Chong et al., 2018, 2019). As neural networks require fixed input sizes, a key challenge with the time-series approach is how to use differently sized inputs. Typical approaches pad short and truncate longer sequences (Arapakis & Leiva, 2020) and produce a prediction for each chunk (Antal et al., 2021; Chong et al., 2019). However, this approach is wasteful on highly variable lengths of trajectories. Further, producing several predictions per user has been suspected to be subject to target leak (Leiva & Arapakis, 2020). Consequently, we propose to first split the trajectories across lengths but rejoin them before updating the network (see further details in Appendix A.2). Additionally, using a sliding window approach retains the context of each data point when fed into the model (Chong et al., 2019).

Secondly, the focus can be put on the spatial aspect (e.g., shape and patterns such as loops) of mouse trajectories by plotting them on a two-dimensional positioning plane, thus representing them as images. Accordingly, image processing models such as two-dimensional convolutional neural networks (2D-CNN) are paired with this form of data representation (Arapakis & Leiva, 2020; Chong et al., 2018, 2019). A vital drawback of this method is the lack of consideration of the underlying user interface, such that differently-sized screens yield varying images of the same trajectories. Further, creating images from the mouse trajectories is a resource-intensive step.

Thirdly, raw mouse trajectories can be aggregated into different hand-crafted features. Subsequently, the variables are processable for simpler, non-sequential models such as Logistic Regression or tree-based models. As shown in Figure 12, ample hand-crafted feature options are possible. Many have proven beneficial in producing insights

and predictions (Arapakis & Leiva, 2016; Feher et al., 2012; Matthiesen & Brefeld, 2020; Yamauchi, 2013).

Figure 12: Raw mouse trajectories to feature examples



A first common group is the temporal information of a given event, for example, pauses occurring due to hesitation (Arapakis & Leiva, 2016; Feher et al., 2012; Matthiesen & Brefeld, 2020). Second, events, such as the number of clicks, may reflect aspects such as a patient changing their answer many times. Third, speed measures how much distance the mouse crosses in a given time (Gamboa & Fred, 2004; Pepa et al., 2021). Fast movement could, for example, be an indication of frustration or lack of patience (Yamauchi, 2013). Further, the movement patterns can be formalised through characteristics such as angles, curvature or jitter (Hassan Hosny et al., 2022), which can, for example, reflect a user's stress level (Martín-Albo et al., 2016; Naegelin et al., 2023).

Parallel to intervention login data, hand-crafted mouse features can be aggregated in different ways, such as sums, means or minimum and maximum values (Bremer et al., 2020). Which of these options makes the most sense depends on the nature of the feature at hand. For example, average and maximum speed could be interesting, while the minimum will likely be 0 for all users. In the end, the number of such multicollinear features must be appropriate for the data set size and model used (Zantvoort, Nacke, et al., 2024). A key drawback of hand-crafted features is the time-intensive manual pre-processing and consequent bias in their selection. However, a critical upside is the transformation into meaningful, humanly interpretable features that require less computational resources. A detailed overview of exemplary calculation of the features can be found in Appendix A.1, and the code to produce them on data sets from the above-mentioned tracker is available in this study's GitHub repository.

2.3 Data Science Methodology

As common in the medical domain, DMHI data sets tend to be small, with median data set sizes of 155-350 (Hornstein et al., 2023; Karyotaki et al., 2021; Zantvoort, Hentati Isacsson, et al., 2024). Small data sets risk overfitting, especially when paired with inadequate validation methods and flexible models (Bates et al., 2022; Cawley & Talbot, 2017; Hastie et al., 2017; Lateh et al., 2017; Piccialli et al., 2021). Further single test sets and even cross-validation (CV) results vary widely, with higher, improbable

results likely being overrepresented in publications (Ahmed & Lofstead, 2021; Hullman et al., 2022; Piccialli et al., 2021). Consequently, concerns about the generalisability of prediction results in mental health are increasing (Hilbert et al., 2024; Sajjadian et al., 2021; Vieira et al., 2022; Zantvoort, Nacke, et al., 2024).

We, therefore, propose using nested CV, such that the inner CV optimises hyperparameters, and the mean outer CV score estimates the model performance (Bates et al., 2022). Further, repeating the experiment and reporting the mean score, including the confidence interval, accounts for the variance of results due to randomness in the model (Ahmed & Lofstead, 2021; Cabitza & Campagner, 2021; Scott et al., 2021).

3 Case Study

In the following, an exemplary case study on gathering and processing mouse trajectories at baseline to predict intervention dropout in DMHIs in Sweden is discussed.

3.1 Interventions and Participants

The data was gathered from a subgroup of two ongoing studies at the Karolinska Institutet in Stockholm, Sweden, between the beginning of March 2023 and mid-March 2024. The data gathering ended when the treatment platform the tracker was implemented to was taken out of service. All participants gave informed consent to participate in the studies. As pooling different interventions has been shown to decrease overfitting and improve results (Zantvoort, Hentati Isacsson, et al., 2024), all patients will be considered as one data set. Both studies concern Internet-based cognitive behavioural therapy (ICBT), a form of DMHI.

The first trial (SOPHIA) evaluates the clinical benefits of an ML-based Decision Support Tool for ICBT for depression, social anxiety, and panic disorder. Patients receive twelve weeks of the respective treatment program, all three of which have previously been described in detail and were evaluated with positive results (El Alaoui et al., 2015; Hedman et al., 2013, 2014). The screening and randomisation process is further described in the original studies pre-registration (Bjurner, Isacsson, et al., 2024). The second trial (DANA) includes women in pregnancy weeks 8-29 with mild or moderate major depression. Patients receive 10-week therapist-guided pregnancy-adapted ICBT for depression, partly including supportive counselling sessions with perinatal health staff. The screening, randomisation, and treatment processes are further detailed in Appendix A.3.

3.2 Dropout Definition

The number of modules completed is used to operationalise dropout in this study as it is most associated with symptom outcomes (Donkin et al., 2013; Gan et al., 2021). The goal is to identify patients at risk of prematurely leaving the intervention before assumingly sufficiently benefiting (Beintner, Vollert, et al., 2019). For the treatments in the Sophia trial, seven modules have previously been identified as suitable dropout measures (Zantvoort, Hentati Isacsson, et al., 2024). As the Dana trial comprised different content, target groups, and treatment design, the cutoff is set to six modules.

3.3 Baseline Data

To ensure the evaluation of the mouse data against a minimal baseline approach (DeMasi et al., 2017), depression scores at screening and pre-intervention, age and gender, and target disorder are used as baseline variables.

3.4 Mouse Data

As they are routinely gathered across all interventions, all mouse dynamics were gathered on the Montgomery–Åsberg Depression Rating Scale-Self Report (Montgomery & Asberg, 1979; Svanborg & Asberg, 1994). The MADRS-S questionnaire comprises nine questions regarding depressive symptoms (0-6 scale) and, in the setting at hand, is spread over three different pages. The inactivity limit to stop tracking is based on the 95%-per centile (1000 seconds) of 2500 patients who had previously filled out the questionnaire (Zantvoort, Hentati Isacsson, et al., 2024).

Time-series data: Considering previous success and its lower computational requirements (Antal et al., 2021), time-series representation is chosen for the case study. Thus, each trajectory comprises a one-dimensional time series with two-dimensional elements (x, y). As tracking pauses do not serve a purpose in the time series, they are removed. To 1) encompass the time instance into the spatial information and 2) to ensure translation invariant sequences, the derivatives of the coordinates ($\delta x/ \delta t$, $\delta y/ \delta t$) are calculated.

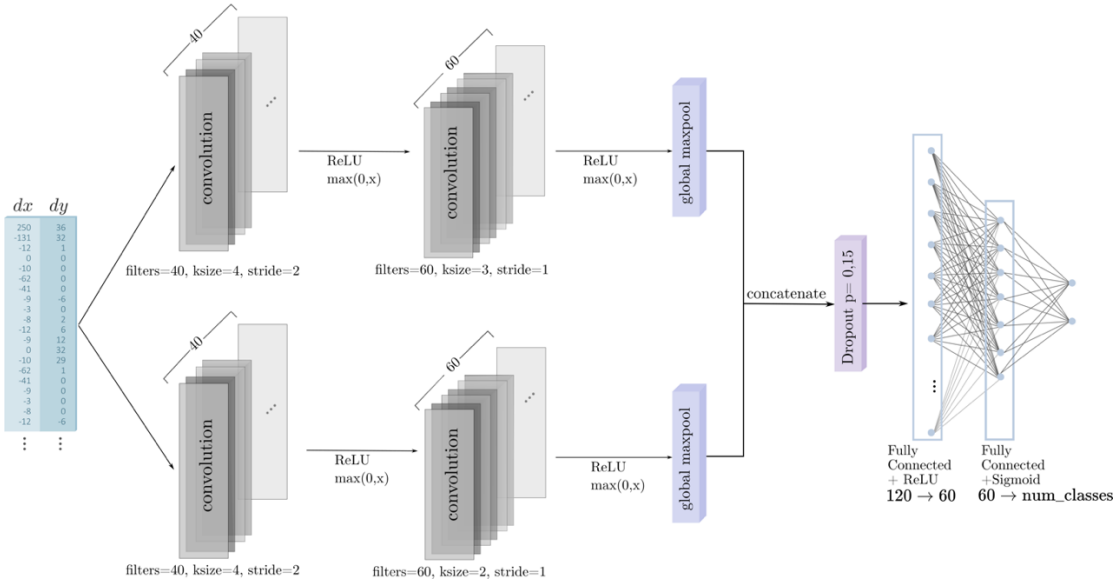
Hand-crafted features: While ML models can theoretically handle large and complex feature groups, in DMHI settings, fewer hand-crafted features have been shown to improve results, especially for small data sets (Bremer et al., 2020; Hentati Isacsson et al., 2023; Zantvoort, Nacke, et al., 2024). Therefore, based on related literature, we select a subset of a small (three) and mid-sized (ten) group of the most promising ones. The small group only focuses on the basics, temporal aspects: Average speed, total time of pauses and scrolling speed. The ten further comprise those three temporal features but also the moved distance, the number of data points, and more abstract features (i.e., the average change in angle, the amount of acute and obtuse angles, and jitter).

3.5 Prediction Models and Experiment Set-up

To mitigate the risk of overfitting on the small data set (Bates et al., 2022; Cawley & Talbot, 2017), the models are trained and evaluated through 5x5 nested-cross validation (CV) with five different seeds. Further, both pipelines include a standard scaler. The evaluation of the outer CV score is done in three steps: Firstly, following Occam's razor principle, to warrant the additional effort needed, more sophisticated features need to outperform simpler ones. Secondly, we externally evaluate based on the above-cited paper that reported an AUC of 0.57 for baseline predictions (Günther et al., 2023). Lastly, we use Kraemer et al.'s (2003) evaluation categories for clinical significance, classifying no (0.50-0-0.56 AUCs), low (0.57-0.64), moderate (0.65-0.70), good (0.71-0.75) and very good (>0.75) predictive power.

Sequential Neural Network: Following related work (Antal et al., 2021; Chong et al., 2018, 2019), a one-dimensional convolutional neural network (1D-CNN) is implemented. The architecture comprises the concatenation of two towers of each two convolutional filters, one dropout layer and two fully connected layers, and a sigmoid activation function (see Figure 13). A fixed window size is used with an overlap of 50% with the previous sequence to contain the context of every mouse sequence. Different block sizes [100, 128, 265] are tested within the inner fold. The network outputs for each block are aggregated to one prediction per user before providing feedback to the network. Further, the model is trained for 35 epochs with a learning rate of 0.001. We use the Adam optimiser and the binary cross entropy as a loss function.

Figure 13: Architecture of the 1-dimensional convolutional neural network



Non-sequential models:

To analyse the hand-crafted features, a Random Forest model is trained in a trade-off between simple but flexible enough to detect non-linear and interaction effects. Missing depression scores are imputed, assuming the mean change between screening and pre-treatment from the training data with the patient's respective symptom level. The RF model searches across the number of estimators [3, 5, 10, 25], minimum samples for a split [5, 15, 25, 50], the maximum depths [3, 5, 10, 25, 50], and a binary indicator for bootstrapping.

4 Results

4.1 Final Data Set

The final baseline data set comprises 408 patients, but 55% of them used a mobile device and, hence, did not produce mouse trajectories. Therefore, the mouse trajectory analysis includes only 184 desktop computer or laptop users. One user had insufficient mouse data points (<10) and was therefore excluded. The final descriptive numbers across studies can be seen in Table 4. For the screening and pre-intervention MADRS-

S scores, there are 13 (2.5%) and 17 (4.2%) missing values, but no patient misses both time points.

Table 4: Descriptive statistics for a) baseline data and b) mouse data set

a) Mouse data	**Target Disorder**	**N**	**Dropout in %**	**Mean Age (SD)**	**Females (%)**
Dana	Depression	75	48%	34.67 (3.83)	100%
Sophia	Depression	52	44%	42.25 (10.51)	54%
	Social Anxiety	36	50%	35.97 (10.30)	78%
	Panic Disorder	20	31%	42.40 (12.58)	50%
b) Baseline data					
Dana	Depression	144	56%	34.28 (3.95)	100%
Sophia	Depression	115	54%	41.54(10.59)	66%
	Social Anxiety	85	63%	36.21 (9.72)	74%
	Panic Disorder	64	38%	41.09 (10.98)	58%

4.2 Prediction Results

Providing the model with only baseline data resulted in an average inner CV (training) score of 0.59 and outer (test) score of 0.56 (95%-CI 0.54-0.57) AUC. Adding the ten hand-crafted mouse features increased the inner CV score to 0.65 but did not improve the outer CV score 0.56 AUC (95%-CI 0.52-0.61). Only adding the average speed, pause time and scroll speed, on the other hand, increased the inner CV to 0.66 and the outer CV score to 0.58 (95%-CI 0.55-0.62). The NN produced an inner CV score of 0.51 and an outer CV score of 0.50 (95%-CI 0.44-0.54).

5 Discussion

Considering their information power regarding users' mindset and emotional state (Fu et al., 2017; Lepora & Pezzulo, 2015; Maldonado et al., 2019; Yamauchi, 2013), mouse trajectories are a promising candidate for predictions in mental health care. However, they are complex in nature, and no insights into their use in DMHIs exist. Therefore, the study at hand discusses how to gather and process mouse trajectory data and provides an exemplary use case for intervention dropout predictions.

Mouse trajectories can easily and unobtrusively be gathered through the provided tracker. They can then be transformed into time series, images of hand-crafted features and paired with the respective model type.

In the case study, while adding the three simplest selected hand-crafted mouse features negligibly increased prediction results, the time-series approach did not succeed on the data set at hand. As such, this paper is in line with several works finding more sophisticated models to fail on small-sized e-mental health data (Bremer et al., 2017; Funk et al., 2020; Gogoulou et al., 2021; Hentati Isacsson et al., 2023; Smink et al., 2021; Zantvoort et al., 2023; Zantvoort, Nacke, et al., 2024). Neural networks are known to require larger data sets (Alzubaidi et al., 2023; Piccialli et al., 2021), especially as the task of predicting behaviour several weeks in the future is a challenging one (Gogoulou

et al., 2021). Future research should, therefore, be conducted on larger mouse trajectory datasets.

Despite the small data set comprising four heterogeneous interventions, the baseline prediction of 0.56 AUC is comparable to related work (Gonzalez Salas Duhne et al., 2022; Günther et al., 2023; Linardon et al., 2022). However, overfitting is clearly a risk for the mouse features as the inner CV score increased by up to 0.09 AUC but did not generalise to the outer CV score. Overfitting is presumably also the issue of the neural networks, as the training accuracy of 0.90 did not generalise to the CV scores (0.50 AUC). Relying on an inadequate validation set-up with the data set size at hand would have led to an overestimation of the performance (Hilbert et al., 2024; Sajjadian et al., 2021).

Beyond the data set size, this study has several limitations. Firstly, the recorded data neither accounted for the devices used (e.g., touchpad, vertical or normal mouse) nor non-accurate movement caused by the hardware. Secondly, albeit proposed to increase prediction accuracy and stability (Zantvoort, Hentati Isacsson, et al., 2024), the heterogeneity of the different interventions could be too pronounced considering the small data set size at hand. Thirdly, we only explore the prediction power of mouse trajectories on the pre-treatment depression questionnaires, which is a very limited part of the assessment, especially as it is not the primary symptom for all patients. Future research could explore the second proposed use case of leveraging mouse trajectories to improve adherence measures, for example, through pauses, the explored content or predicted attention (Leiva & Arapakis, 2020). Lastly, more than half of the patients accessed the questionnaire via mobile devices, which emphasises the importance of also investigating touch data (Yang et al., 2021). Only by combining both input channels can predictions for all patients be made.

In conclusion, the paper at hand introduces mouse trajectory data, a tracker to gather, and the process and code necessary to leverage them. While the case study results show that hand-crafted features slightly increase prediction results, 183 patients are arguably too few to leverage complex feature and model combinations. Considering their successful implementation in other domains, more research on larger data sets is necessary to determine the predictive value of mouse trajectories as unobtrusive and rich data type in DMHI predictions.

Chapter V: Intervention Text Features

Zantvoort, K., Scharfenberger, J., Boß, L., Lehr, D. & Funk, B. (2023). Finding the Best Match — a Case Study on the (Text-)Feature and Model Choice in Digital Mental Health Interventions. J. Healthcare Informatics Research 7, 447–479.

Abstract: With the need for psychological help long exceeding the supply, finding ways of scaling, and better allocating mental health support is a necessity. This paper contributes by investigating how to best predict intervention dropout and failure to allow for a need-based adaptation of treatment. We systematically compare the predictive power of different text representation methods (metadata, TF-IDF, sentiment and topic analysis, and word embeddings) in combination with supplementary numerical inputs (socio-demographic, evaluation, and closed-question data). Additionally, we address the research gap of which ML model types – ranging from linear to sophisticated deep learning models – are best suited for different features and outcome variables. To this end, we analyse nearly 16.000 open-text answers from 849 German-speaking users in a Digital Mental Health Intervention (DMHI) for stress. Our research proves that – contrary to previous findings - there is great promise in using neural network approaches on DMHI text data. We propose a task-specific LSTM-based model architecture to tackle the challenge of long input sequences and thereby demonstrate the potential of word embeddings (AUC scores of up to 0.7) for predictions in DMHIs. Despite the relatively small data set, sequential deep learning models, on average, outperform simpler features such as metadata and bag-of-words approaches when predicting dropout. The conclusion is that user-generated text of the first two sessions carries predictive power regarding patients' dropout and intervention failure risk. Further, the match between the sophistication of features and models needs to be closely considered to optimise results, and additional non-text features increase prediction results.

1 Introduction

Estimates suggest that even before 2020, only a third of people affected by mental health problems received the help they needed (Rommel et al., 2017; Wang et al., 2005). This unmet need is accelerated by the psychological aftermath of the COVID-19 crisis, with estimated growth rates in the prevalence of major depression and anxiety disorders of more than 25% (Santomauro et al., 2021). Consequently, offering effective help on a larger scale is of paramount importance for individuals and, considering costs and devastating impacts, for societies as a whole (Ebert et al., 2019).

Digital Mental Health Interventions (DMHIs) help to provide psychological treatment as they are easily accessible, economical, and scalable (Karyotaki et al., 2015). DMHIs pursue related goals to face-to-face therapy but are conducted through the means of online education formats. They mainly consist of self-help texts, video or audio manuals, and exercises and can be accessed independently of time and location. DMHIs can be unguided self-help interventions or can include guidance by e-coaches, for example, via calls or messages (Andersson et al., 2019). Meta-analyses demonstrate their efficacy in treating various mental health problems like stress (Heber et al., 2017) and stress-related disorders such as depression (Karyotaki et al., 2015; Reins et al., 2020) and anxiety (Domhardt et al., 2020). At the same time, to be effective, a participant must finish at least a certain amount of the intervention to show health benefits (Donkin et al., 2011; Gan et al., 2021). However, it is estimated that in unguided DMHIs, three out of four participants drop out too early. At one in three participants, the odds are better, yet still problematic, in guided DMHIs (Richards & Richardson, 2012). Such dropout is a key factor identified for participants' variance in response rates, causing Gan et al. (2021) to call for strategies to help those who struggle. Measures such as e-coaches' guidance, reminders, and personalization positively influence overall completion rates and health outcomes (Baumeister et al., 2014; Gan et al., 2021; Hilvert-Bruce et al., 2012). However, the extent of guidance necessary differs among individuals and many complete and benefit from interventions with little or none of the usually costly support. Hence, in order to optimally allocate the limited resources and effectively help as many as possible, participants in need of attention must be identified (Forsell et al., 2019).

Machine learning (ML) models can make individual predictions and have previously been used to estimate intervention dropout and failure probabilities (Shatte et al., 2018). Most of these attempts focus on user journey data, including log-in data and other indicators of online behaviour (Bremer et al., 2020; Chekroud et al., 2021; Pedersen et al., 2019). At the same time, human language is the primary tool in psychiatry and psychology (Abbe et al., 2016; Funk et al., 2020). Accordingly, DMHIs often include asynchronous text-driven communication with participants, generally involving (1) open-text intervention exercises and (2) direct communication with e-coaches (Calvo et al., 2017). Such texts are known to hold valuable information regarding a user's mental state and intentions that e-coaches can use to best support their participants (Bone et al., 2017). Extracting information from these texts is a promising but time-consuming task and thus poses a major challenge with respect to scalability.

Natural Language Processing (NLP) is a field of computer science specifically designed to handle text data. Using NLP methods to automate or augment parts of the e-coaches' work is a largely unexplored field of research (Shatte et al., 2018). First advances train ML models on the users' text to predict binge eating behaviour (Funk et al., 2020) as well as intervention outcomes for social anxiety (Hoogendoorn et al., 2017), and depression interventions (Gogoulou et al., 2021). NLP methods are ample and differ in both their complexity and their requirements. Obtaining descriptive numbers (e.g., length of the text) and simple counts of words (i.e., bag-of-words approaches) is straightforward from a technical point of view. However, the amount of human decision-making and manual pre-processing is high, and the contextual meaning captured is essentially non-existent. Word embeddings based on neural networks can account for the context of words (Devlin et al., 2019) and have set various NLP prediction performance benchmarks outside of DMHI text data (Nobles et al., 2018). However, first applications to intervention text data are disenchanting when paired with simple classifiers (Funk et al., 2020; Gogoulou et al., 2021). With some results worse than random chance, Gogoulou et al. (2021) conclude that "*the task of predicting treatment outcome based on patient text is very difficult*" [28, p. 578]. These results notwithstanding, word embedding features are successfully combined with more complex ML models in the related field of mental health diagnostics (Cohan et al., 2018; Nobles et al., 2018). As these conflicting findings show, deciding on a suitable combination of text representation techniques and ML models remains a largely unexplored problem in DMHIs. In addition, the predictive power of newer deep learning models, such as bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019) is yet to be explored in the context of intervention text data. Beyond the issues discussed thus far, Funk et al. (2020) point out that the isolated investigation of text data overlooks the likely interaction with non-text features such as the age of participants – a hypothesis supported by several other authors' findings (Calvo et al., 2017; Hoogendoorn et al., 2017; Howes et al., 2014). Hence, the main motives driving this research are (1) the open question of how to best combine automated text analysis with non-text features to optimise resource allocation in DMHIs, (2) the hypothesis that previous performances of word embeddings in DMHIs are limited by the subsequent classification models used, not the word embeddings themselves, and (3) the proposition that a BERT model pre-trained on a general corpus will have predictive power in the intervention setting as well.

Joining the rising efforts of ML applications and automatization in the health sector (Oesterreich et al., 2020; Wołk et al., 2021), we tackle the problem of machine-learning-aided decision-making in E-Mental health research. Within this research area, our clinical application is optimising resource allocation to relieve an overstrained system by identifying those that most need additional support. The findings of our case study on 849 participants allow for the derivation of more concrete hypotheses for the further investigation of empirical generalisation (Tsang, 2014). More precisely, our contribution is threefold: First, we systematically compare the predictive power of different text representation methods (i.e., metadata, TF-IDF, topic analysis, sentiment analysis, and word embeddings) in combination with supplementary numerical inputs (socio-

demographic, evaluation, and closed-question data) for intervention dropout and failure. We complement related work by investigating which ML model types – ranging from linear to sophisticated deep learning models – are best suited for different features and outcome variables. Second, we account for the relatively long and sequential input texts by designing a task-specific neural network architecture which (in many settings) outperforms existing word embedding approaches on intervention text. Third, we demonstrate the potential of BERT models (Devlin et al., 2019) pre-trained on generic text corpora in dropout prediction. To this end, this paper is structured as follows: we summarise related work (Section 2), describe our research approach (Section 3), present the text representation techniques (Section 4) and ML models (Section 5), and thoroughly evaluate different combinations of text representation methods and ML models (Section 6). Finally, we discuss the limitations of our study and highlight future research directions (Section 7).

2 Background

In medical research, the number of ML applications has greatly increased in recent years as they promise improved care, scalability, and cost efficiency (Eloranta & Boman, 2022). Such improvements are particularly needed in mental health care, where patients often go undiagnosed (Cepoiu et al., 2008), and require long-time monitoring and care (DeMasi et al., 2017). While many data types (e.g., log-in or questionnaire data) are available (Shatte et al., 2018), text data presents itself as a propitious option in a field that has always primarily relied on language for diagnosis and treatment (Becker et al., 2018; Funk et al., 2020). Several research branches emerged to leverage text data's vast occurrence in the context of mental health (Calvo et al., 2017). As their nature, accessibility, and use significantly differ, Becker et al. (2018) call for differentiation between research on *pre*-intervention and intervention data. This chapter briefly explains both and outlines related work to derive the research gaps addressed in this study.

2.1 Pre-Intervention Text Data

Pre-intervention data is gathered before and, thus, outside of a clinical intervention. Use cases focus on diagnosing mental health disorders and generating insights. For this purpose, much attention has been placed on social media data (Becker et al., 2018; Masino et al., 2018; Paul et al., 2021; Yeruva et al., 2019). These datasets gather users' natural communication with each other on platforms like Twitter or Reddit. One example are Cohan et al. (2018), who tackle a multi-class diagnosis problem on a dataset of 20.406 self-reportedly diagnosed and 335.952 control users' social media posts. They find that sequential neural network approaches outperform their non-sequential models trained on Term-frequency Inverse Document Frequency (TF-IDF) (Spärck Jones, 1972) features in eight out of nine conditions. Yeruva et al. (2019) compare insights on obesity and healthy eating - topics related to eating disorders (Marcus & Wildes, 2012) - from 103.609 Tweets versus 6.602 academic abstracts from PubMed. They propose a pipeline to construct social and contextual word embeddings, which

produce valuable insights. Wongkoblap et al., (2021) predict depression diagnoses for 4.169 Twitter users. On the one hand, they compare the dictionary-based Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al., 2015), a language model, topic analysis, and Usr2Vec (Amir et al., 2017) features paired with Logistic Regression (LR) or Support Vector Machines (SVMs). On the other hand, they pair word embeddings with a one-dimensional convolutional neural network (CNN), as well as two task-specific (attention-based) neural network architectures. At AUCs of 0.91-0.93, their sequential models outperform their non-sequential models with AUCs of 0.79-0.88. They explain this gap with the information loss non-sequential models suffer when features are aggregated across words. More recent studies in Mental Health diagnostics go one step further by using a more novel pre-trained BERT model (Devlin et al., 2019), which yields good results (Bucur et al., 2021; Wołk et al., 2021) and thus shows promise for other areas of text data in E-Mental Health research. As much more work exists than can be discussed here, reviews such as Chekroud et al., (2021), Calvo et al., (2017), or Le Gaz et al., (2021) can be referred to for a more detailed picture.

While pre-intervention text data is usually publicly and easily accessible on large scales (e.g., through crawlers), it lacks health labels such as a reliable clinical diagnosis and must depend on self-published information. Further, anonymity and limited information verification options can cause issues with data quality (Calvo et al., 2017; Eloranta & Boman, 2022). In summary, pre-intervention text data was produced in a non-clinical setting and primarily generates diagnoses and epidemiological insights.

2.2 Intervention Text Data

In contrast, intervention data comes from a clinical setting designed to help an already diagnosed user. Here, text is produced by health staff (e.g., Electronic Health Records (Ewbank et al., 2020; Le Glaz et al., 2021)) or by the users themselves (Becker et al., 2018). In DMHIs, users primarily produce answers to open-text questions or conversation data with health staff. Because of the controlled setting, high-quality socio-demographic, longitudinal symptom, and user behaviour data is usually available. However, gathering intervention data requires resource-intensive steps such as screening, diagnosis, and the assurance of weeks-long (guided) interventions. Consequently, such data points tend to be costly, and data sets stay small (Pasini, 2015). Additionally, access to existing datasets is extremely limited due to privacy concerns (Andersson et al., 2019; Gan et al., 2021). As a result, Shatte, Hutchinson and Teague (2018) find that only 1% of studies investigating ML in a Mental Health setting investigate intervention data, and barely any consider NLP methods. In agreement with these findings, several authors conclude that NLP on intervention data is vastly understudied despite its substantial potential (Calvo et al., 2017; Funk et al., 2020; Le Glaz et al., 2021; Shatte et al., 2018).

In mental health interventions, lack of adherence and responsiveness to treatment are major concerns (Andersson et al., 2019; Donkin et al., 2011; Eysenbach, 2005). As shown by Forsell et al. (2019), Pedersen et al. (2019), and Pihlaja et al. (2020), targeted measures such as human guidance can improve upon these problems but, in an already

overstrained system, cannot be offered to all participants. Here, supervised ML models provide great value by identifying those users that require additional care and allowing for individually targeted measures (Forsell et al., 2019). To present a comprehensive picture of previous work of NLP for dropout and intervention failure prediction, we search PubMed with the query ("Natural Language Processing" OR "NLP") AND ("Psychology" OR "Psychiatry" OR "DMHI*") AND ("Predict*" OR "Machine Learning") AND ("Outcome" OR "Dropout" OR "Adherence"). We include papers that used ML models to make individual dropout or outcome predictions based on user-generated open-text data in DMHIs. We then follow the citations in the related work section for more relevant papers. Further, a PubMed search including a similar query with the term "BERT" did not lead to any studies including user-generated intervention data.

Howes et al. (2014) predict intervention outcomes based on chat data between therapists and 167 English-speaking users of a depression and anxiety intervention. Simple LR, linear SVMs, and Decision Tree (DT) models are trained for classification. The authors conclude that a combination of demographic and metadata yields better results than the slightly more sophisticated sentiment and topic analysis. The best-reported f1 measure improves the baseline from 0.57 to 0.7. However, as they point out, they split several messages of one patient between test and training set in their 10-fold cross validation. With limited patients available, the combination of age, gender, and therapist can already allow a model to identify an individual participant and infer the result from the training example.

Hoogendorn et al. (2016) retrieve information about sentiment, topics, writing style, and word usage from German emails written by 69 social anxiety patients, together with meta and demographic data. They investigate 1) averages and 2) trends per person. They choose the 20 features most correlated with their outcome variable – symptom levels at week twelve – mainly covering single words (17), topics (2), and writing style (1). For classification, they train LR, DT, and random forest (RF) models, arguing that these model types give reasonably good and understandable results. While socio-demographic data alone has no predictive value, complementing it with text data up to week six significantly enhances the prediction performance of their RF model (AUC 0.83).

Smink et al. (2021) use 770 participants' first four out of an average of 20 emails written in a DMHI for alcohol abuse to predict dropout. They retrieve word count and LIWC (Pennebaker et al., 2015) features and combine them with socio-demographic data. The classifiers used are LR, a neural network, XGBoost, and a Mixed-effect RF model. First, they aggregate the features as means across all four emails for the non-sequential models. Second, they input the features per email into their sequential neural network and RF. Hence, while sequential models are included, they only consider the order of emails, not the sequentially of language itself. The winning XGBoost model performs worse than their baseline, leading to the conclusion that they could not associate their simple email text features with intervention dropout.

Funk et al. (2020) use 372 participants' English messages and intervention text snippets to predict binge eating episodes in the next 24 hours. A total of 100 of these participants also have the 6-month follow-up health outcome. The authors compare an array of different methods of text representation: metadata, bag-of-words models including topic and sentiment scores, word embeddings, and Part-of-Speech tagging. To predict short-term symptom severity, they train an LR and an RF model, resulting in a maximum AUC of 0.57 for new users. Additionally, they use LASSO regression to determine the best out of their 220 variables for the long-term outcome prediction. None of the 50 element-wise averaged embedding dimensions are among the most informative features.

Gogoulou et al. (2021) compare TF-IDF, Word2Vec, FastText (Bojanowski et al., 2017), and Doc2Vec (Le & Mikolov, 2014) text representation on Swedish homework reports of 1.986 users of a depression intervention. The three word embeddings are trained in advance on an additional 4.835 users' texts from other interventions. In their approach, TF-IDF outperforms the word embeddings in almost all settings, and in some cases, the latter even perform worse than the naïve baseline. With a maximum f1 score of 0.69 (baseline of 0.58), they conclude there is a signal in intervention text data regarding outcome prediction, but word embeddings do not serve to extract it. While their paper has the by far largest sample for intervention text data and considers three different methods of word embeddings, it only uses a simple linear classifier. As such, they do not put their focus on leveraging the sequential nature of word embeddings (Mikolov et al., 2017).

In conclusion, only one paper investigates the prediction of dropout based on intervention text data, with little success. However, as the authors propose, features other than the two simple ones included should be investigated (Smink et al., 2021). For outcome prediction, several studies find that combining text features with non-text features such as socio-demographic data leads to the best results (Calvo et al., 2017; Hoogendoorn et al., 2017; Howes et al., 2014). This results in the first research focus of this paper presented in the introduction; the question of how to best combine text analysis with non-text features to optimise resource allocation in DMHIs. The works so far suggest that simpler text representation features are superior in their predictive performance. However, datasets were almost always smaller than 250 users, and the focus has been on linear and simpler tree-based classifiers. Those papers including more sophisticated models, only used simple features. Thus, the performance of more sophisticated models, such as ensemble methods and deep neural network classifiers in combination with complex features, remains to be investigated in typical DMHI prediction tasks. This leads us to our second research proposition: Previous performance of word embeddings in DMHIs is limited by the subsequent classification models used, not the word embeddings themselves. Further, successful examples from research on pre-intervention data (Bucur et al., 2021; Wołk et al., 2021) let us arrive at our third proposition for this paper: That a BERT model pre-trained on a general corpus will have predictive power in the intervention setting as well.

3 Study Set-Up

This study addresses the gap in existing research by systematically exploring the predictive power of text (i.e., different metadata, TF-IDF, sentiment and topic analysis, Word2Vec and FastText word embeddings) and non-text data types (i.e., socio-demographic and symptom data, evaluation data, and closed-question data) and their interplay with different model types (i.e., LR, SVMs, XGBoost, AdaBoost, LSTMs and BERT). We investigate these results for intervention dropout and outcome to provide insights into the use of ML methods to optimise resource allocation. The final goal is better outcomes with equal or lower costs (Forsell et al., 2019; Pedersen et al., 2019). A key focus of this paper is the investigation of the gap between the word embeddings' theoretical power and the lack of its manifestation when used on intervention text data. To this end, two different word embeddings are trained and then (1) averaged for non-sequential models and (2) used as they are with a sequential model. Furthermore, we employ BERT to make predictions based on the intervention text data, which – to the best of our knowledge – has not yet been investigated. At the same time, Occam's razor principle suggests that – ceteris paribus – the simplest model is preferable (Blumer et al., 1986). Because of this, feature extraction methods and models of different sophistication levels are pit against each other in this exploratory study of how to best predict intervention failure and dropout. With 849 participants, the dataset at hand is larger than all but one of the previous works on intervention text.

3.1 Data Description

For our case study, we consider the data of 927 participants from six randomised controlled trials (Table 5) of an internet-based stress management intervention called GET.ON Stress (Heber et al., 2017). The training program comprises seven sessions, planned to be held on a weekly schedule. Each session consists of general information, quizzes, audio and video files, downloadable worksheets, and interactive exercises. The interactive exercises are the most important element in each session. Users work through the exercises by reading or listening to short instructions and then writing their answers into text boxes. In subsequent sessions, many of the text inputs are picked up and displayed again to the user by the system. The core stress coping strategies included in the training program are problem-solving (D'Zurilla & Nezu, 2001) and emotion regulation (Berking & Whitley, 2014). At the beginning of the program, participants write about their stressors, goals, and motivations. In each subsequent session, the participants are asked to choose pleasant activities, plan to implement them into their lives, and to reflect on how it went in the subsequent session. In the second and third sessions, participants learn a systematic six-step problem-solving method that can be applied to their problems, again reflecting on it in the subsequent sessions. In sessions four to six, participants learn and practice different emotion regulation techniques, such as muscle and breathing relaxation (Berking & Whitley, 2014). In the seventh session, participants reflect on their goals for the training and plan how to continue practising stress coping in the future. Four weeks after completing session seven, an optional booster session eight is provided. Depending on the trial,

participants went through the program as a self-help intervention, were able to ask for feedback, or automatically received written feedback by e-coaches after every session. For more detailed information on the set-up of the intervention and each of the studies, please refer to the primary publications cited in Table 5.

Table 5: Overview of the intervention studies included in this analysis.

Study	German Clinical Trial No.	Publication	Level of human support
1	DRKS00004749	Heber et al., (2016)	Intensive guidance[a]
2	DRKS00005112	Ebert, Lehr, et al., (2016)	Guidance on demand[b]
3	DRKS00005384	Ebert, Heber, et al., (2016)	No guidance[c]
4	DRKS00005687	Nixon et al., (Preprint)	Guidance on demand
5	DRKS00005990	Ebert et al., (2021)	Guidance on demand
6	DRKS00005699	Nixon et al., (2022)	No guidance

Notes: [a] participants receive written feedback (avg. 30 minutes) after each session; [b] participants receive feedback on demand; [c] participants receive technical support only

In this study, intervention dropout is defined as having finished less than the six core sessions out of eight total sessions. Sessions 7 and 8 are not considered core sessions as they do not convey new material but instead serve as a reflection and repetition session, respectively. As such, the dropout definition follows the consensus of operationalising dropout reported by Donkin et al. (2011) and is recommended to use by Gan et al. (2021). The second session is chosen as the point of prediction due to the trade-off between text gathered and time left to intervene (Bremer et al., 2020). Choosing this prediction point results in 849 German-speaking participants who completed exercises in the first two sessions – 25% of whom are considered dropouts. Intervention failure is defined as an improvement of fewer than 5.16 points on the Perceived Stress Scale (PSS) (Jacobson & Truax, 1991; Schneider et al., 2020), the primary health outcome metric. This threshold value of 5.16 is based on the reliable change index indicating a clinically meaningful change in symptomatology introduced by Jacobson and Truax (1991). The average baseline PSS score is 25 and, after finishing an average of 6.6 sessions, ends at 17. In total, 37% of users considered are intervention failures. A total of 40 participants did not fill out the PSS questionnaire after finishing the intervention and, therefore, cannot be considered for intervention failure prediction. Losing many data rows because participants did not fill out the final symptom questionnaire is a common problem when predicting intervention outcome. For example, Gogoulou et al. (2021) disregard 38% of their participants because their low adherence prevents the calculation of the target features. Attempting to predict the 6-month follow-up, Funk et al. (2020) even lose 73% of their data. In this dataset, from those with unknown outcomes, 85% dropped out. Excluding these participants runs the risk of ignoring those most in need of additional support. Therefore, we provide insights into both, dropout (keeping more participants) and intervention failure (the more exact outcome measure) predictions.

3.2 Non-Text Data

Related work suggests that a combination of text and non-text features is most promising when retrieving information about a user's mental state (Calvo et al., 2017). Thus, unsurprisingly, a myriad of the above-mentioned studies includes non-text variables in their analysis. We train benchmark models on each of the non-text and text feature types by themselves and then combine them to be able to differentiate between individual, and interaction effects.

Baseline variables such as socio-demographics or symptom data have been thoroughly investigated in terms of their predictive power for dropout and intervention failure, howbeit with limited consensus in results (e.g., (Christensen et al., 2009; Hedman et al., 2014)). We include these variables in our analysis based on the assumption that ways of expressing oneself are dependent on users' characteristics such as age and gender (Calvo et al., 2017). Asking users to fill out a baseline questionnaire before starting the intervention is common, as seen in the related work section. Our eleven socio-demographic variables cover different information about the participants' age, gender, educational background (2 features), occupation (5), and family status (2). 77 participants did not indicate their income level, they are accounted for in an additional feature. The descriptive statistics and data types of all included socio-demographic features can be found in the Supplementary Material 1. The majority of participants identify as female (78%), hold a college degree (60%) and are on average 42 years old, where the age distribution is bimodal with two peaks around 30 and 50 years. In addition to the socio-demographic variables, five symptom-related variables provide the baseline PSS subscores of Helplessness and Self-Efficacy (Christensen et al., 2009), and carry information about previous experiences with training and therapy (3 features). The mean values of the PSS subscores before the intervention are 16 and 9, respectively. The aforementioned variables are supplemented by the intervention support level and an indicator of whether the user found out about the intervention via their health insurance company.

Evaluation data providing information on the user's attitude towards the intervention can easily be argued to be an evident factor for their intention to continue it. Therefore, this data proposes a promising alternative to the resource-intensive process of text-analysis. At the same time, it requires an additional questionnaire after each session, hence straining the limited user attention available. To investigate this trade-off, it will be included in the analysis. The users evaluate the (1) easiness and (2) usefulness of each session on a scale of 1 (very useful/very easy) to 5 (not useful at all/very difficult). Furthermore, the users were asked to estimate the time they needed to complete the respective session on a rating scale from 1 (less than 30 minutes) to 4 (more than 90 minutes). On average, users rate the easiness with 2.3, the usefulness at 1.8 and the time required between 30 and 90 minutes. Furthermore, users have the chance to articulate well-liked and improvable aspects of each session in an open-text format. For the text representation, we append this text to the rest of the user's generated text of the corresponding session. In total, 735 participants answered the evaluation questions

for at least one of the first two sessions, and missing values are accounted for in an additional feature.

Closed-question data is structured data in the form of questionnaire items that have a limited set of pre-defined answer options, which Cook et al., 2016 found to have better performance than open-text questions when predicting suicidal intentions. Such closed questions are often inherent in the intervention design and are easier to handle than unstructured text data from a technical standpoint. Exemplary impressions of how the users saw such questions can be found in the Supplementary Material 2. In our dataset, three closed-form intervention exercises sum up to an additional 13.298 user entries. These questions address the perceived stress levels, the per centage of successfully implemented goals from the previous session, and the intended day of finishing the upcoming session. We extract the relevant numbers and – depending on the nature of the question – include them as they are or aggregate them (i.e., sums, averages, or counts). We fill missing values with 0s and create additional features indicating missing values.

4 Text Representation

In total, the 849 users produced 61.290 open-text answers to intervention exercises and another 3.647 answers to the open-text evaluation questions. Given the point in time of the prediction, only the text from sessions one and two are used. This leaves 15.773 entries, 1.597 of which are open-text evaluation answers. As a first step, 1.064 entries that do not contain any relevant information (e.g., "*xxx*", "...", "-") are deleted, which are found via the investigation of the answers with less than five characters. A feature counting the number of such entries is included in the simple metadata. Since text representation techniques typically cannot handle numbers well [26], digits are replaced by '#'. Second, we scrape a list of commonly used German abbreviations and manually adjust and supplement them to better fit the context of this intervention. The abbreviations are replaced with their long-form, and special characters, as well as smileys, are deleted. A spell check based on the Hunspell package is tried but does not increase cross-validation scores and, therefore, is not used in the final results. Third, we lemmatize the participants' text using the Python library SpaCy. As upper-case letters carry significant meaning in German (Fehle et al., 2021), the texts are only lower-cased after lemmatization. Since bag-of-words methods usually benefit from lemmatized texts (Fehle et al., 2021), while neural network approaches are not expected to(Camacho-Collados & Pilehvar, 2018), we keep both. Lastly, we aggregate the text per user and session, resulting in a concatenated string of all user text inputs that can be used as-is or be further aggregated across sessions 1 and 2.

4.1 Metadata

Especially when thinking about dropout as the binary manifestation of engagement, the effort invested in the exercises is a promising candidate for its prediction (Karyotaki et al., 2015). Assuming that a longer answer to a given task requires more effort than a short one, the arguably most straightforward measure is the length of the

answer. Hence, we create a *simple metadata* representation of the participants' texts by measuring the word and character count. An average intervention text in sessions 1 and 2 together contain 617 words in 4.105 characters and an additional 48 words in 313 characters for the evaluation questions. To account for the participants' willingness to answer the intervention questions, a feature counting the number of *useless* (defined as above) entries is added. Additionally, the usage of upper cases, exclamation, question marks, and positively or negatively connoted smileys are counted before they are deleted in the text cleaning. The *advanced metadata* is based on Ewbank et al.'s (2020) finding that different therapeutic intentions and topics have different impacts on outcomes in face-to-face intervention. In sessions 1 and 2, tasks aim to gather information about the user's motivation and build skills in problem-solving, stress analysis, behaviour reflection, and behavioural planning. Thus, all text snippets are categorised, and text lengths per category are retrieved to investigate whether this additional information can improve predictions.

4.2 Bag-of-words

Bag-of-words approaches count the occurrences of each word in a document (i.e., intervention answers) in an attempt to extract similarities or differences in texts. A popular bag-of-words method is *Term Frequency-Inverse Document Frequency* (TF-IDF) (Spärck Jones, 1972). The word occurrence count is rescaled based on the relative occurrence of all documents. The scikit-learn TF-IDF vectorizer is used on the word level, considering uni- and bigrams to produce the vector per participant. This approach results in a very large and highly sparse matrix; both attributes that many ML models cannot handle well. To reduce the size of the matrix, features used by more than 70% of documents are discarded, as they are assumed to be stop words. In order to keep fewer features than data points (Funk et al., 2020), the number of TF-IDF features kept is determined by the number of users minus the number of additional non-text features. In two additional steps, sentiment and topic analysis are used to reduce the matrix dimensionality by grouping similar words. *Sentiment properties* of polarity and subjectivity are retrieved per text snippet to extract variation in sentiments depending on the exercise (e.g., "*What stressed you today?*" vs "*What makes you feel good?*"). The German version of the text blob package - a rule-based approach – is used on the lemmatized text, as per the recommendation of Fehle, Schmidt and Wolff (2021). Sentiment polarity is recorded on a scale from [-1,1], with the minimum indicating a negative and the maximum indicating a positive connotation. In addition, the subjectivity variable indicates the level of opinion, emotions, or judgments between 0 (objective) and 1 (subjective). As both the average sentiment and the range of sentiment are considered valuable information (Funk et al., 2020), the mean, max, and minimum scores across sessions are included as features. Another way of reducing the dimensions is *Latent Dirichlet Allocation (LDA),* which tries to identify latent topics in the documents. LDA assumes that a document touches upon different topics operationalised by a list of common words associated with each topic (Blei, 2003). Considering the number of relatively small entries and the likely tendency that similar exercises produce similar answers, this step is done on the already aggregated text, and the

number of topics considered is set to 10 (Funk et al., 2020). The topic model is calculated on the training data corpus only and then applied to the test data text.

4.3 Embeddings

Based on the assumption that similar words appear in similar contexts, word embeddings attempt to analyse word co-occurrences and represent each word by *n*-dimensional vectors of real numbers. Thus, words used in akin contexts tend to be mapped to vectors with small distances. Word2Vec (Mikolov et al., 2017) and FastText (Bojanowski et al., 2017) are frequently used word embedding techniques based on neural networks. Word2Vec offers two different network architectures to learn word representations by (1) predicting a current word based on its surrounding words (CBOW) or (2) predicting the surrounding words based on a current word (Skip-gram). While the learned representations of the words in the training corpus are mostly meaningful, unseen words cause difficulties. In order to find a vector representation of these words, a fraction of rare words is typically mapped to an out-of-vocabulary (OOV) token during training allowing unseen words to be mapped to this generic OOV vector. FastText (Bojanowski et al., 2017) is an extension of Word2Vec, which tries to tackle the problem of unseen words by building embeddings for each word in the corpus as well as the n-grams each word consists of. Hence, word vectors for unseen words can be generated based on the n-grams in a more meaningful way. Both word embeddings can be trained from scratch on custom datasets or word embeddings pre-trained on large text corpora in languages (e.g., Wikipedia or News articles) can be used. Since the text produced by the study participants is different in its structure from generic corpora, we follow related work (Funk et al., 2020; Gogoulou et al., 2021) and train the word embeddings on an extended dataset using the Gensim library. To enhance our small training dataset, we also use the texts generated by control group users and train the word embeddings at the sentence level. We treat the vector dimension *n* and the model architecture (i.e., CBOW or Skip-gram) as hyperparameters which are optimised during the training of our recurrent neural network (Section 5.2). To compare the sequential approach to results from related work (Funk et al., 2020; Gogoulou et al., 2021), we process the generated word embeddings by calculating the element-wise averages of every participant's text and use these averaged word embeddings as inputs for non-sequential models (Section 5.1).

5 Machine Learning Models

In the following, we present the different ML models that are trained to predict dropout and intervention failure. To match the complexity of the text representation methods, we consider three different model categories: (1) traditional ML models for non-sequential data, (2) deep learning models for sequential data, and (3) advanced pre-trained transformer-based models. While non-text features, metadata, bag-of-words, and averaged word embeddings are combined with traditional ML models, we extend related work in this field by additionally maintaining the sequential nature of text by training recurrent neural networks as well as a BERT classification model. We set apart

a hold-out test set (20% of the participants) beforehand to evaluate the models' out-of-sample performance (Chapter 6).

5.1 ML Models for Non-Sequential Data

We use four different classification models: LR, SVMs, AdaBoost, and XGBoost. The corresponding model hyperparameters are optimised in a 5-fold cross validation (CV), where each hyperparameter space is defined by initially choosing small intervals around the default values and incrementally considering adjustments if the boundaries perform best in the CV. For each data input (i.e., combinations of text representations and supplementary numerical inputs), one final model, chosen based on the CV scores, is trained on the entire training data, and evaluated on the hold-out test data. To account for the class imbalance in the dropout data, we create synthetic data of the minority classes by using SMOTE oversampling (Chawla et al., 2002). The sampling ratio is treated as a hyperparameter for all four models and is optimised during the CV.

Logistic Regression as a linear model for binary classification is very popular due to its fast training times, good explainability, and reasonably good results. In light of the dataset size, the liblinear solver is chosen. Given the partially high number of predictors, L_1 or L_2 regularisation are optimised as a hyperparameter in the CV, together with the respective penalization strength (0.01-10). *Support Vector Machines* classify by drawing decision boundaries between classes. SVMs can either use the feature space as is or use a non-linear kernel to map it into a higher dimensional space to make classes linearly separatable (Cortes & Vapnik, 1995). The use of a linear or a Radial Basis Function kernel is optimised as a hyperparameter, each with their own set of regularisation parameters (C: 0.1-1000, gamma: 0.001-1) to balance over- and underfitting. For both, LR and SVMs, a scaler is added to the ML pipeline. *XGBoost* is a fast and efficient implementation of a Gradient Boosting Tree that also allows for the regularisation of features and thus avoids overfitting on smaller datasets (Chen & Guestrin, 2016). As the XGBoost classifier has many non-trivial hyperparameters, Bayesian Search CV is used to allow a less computationally expensive grid search (Guyon et al., 2011). To constrain the architecture of the trees, the max. depth (3-5), and the minimum weight of a child (0.5-1) are optimised. Further measures against overfitting are the per centage of rows (0.5-1), and columns (0.5-1) used to build each tree, as well as the regularisation parameters gamma (0/1) and lambda (1/2). The number of estimators (50-1000) is also investigated with the learning rate for each step (0.01-0.5). *AdaBoost* classifiers leverage the advantages of ensemble learning by combining a variety of weak learners to achieve better predictions (Schapire, 2013). The number of estimators (3-2000) used stands in a trade-off to the learning rate (0.001-2) – the weight given to each estimator – because of which these are optimised together. We implement our models in Python using the Scikit-learn and xgboost libraries. The non-sequential models can be trained on a standard laptop, and training times partially depend on the number of features. Including grid search, LR and SVMs usually need mere seconds while the AdaBoost model, on average, takes several minutes. Training times are the longest for the XGBoost models, where iterating through the entire

hyperparameter space often takes longer than for the AdaBoost model, despite the use of Bayesian Search CV. Including the large number of TF-IDF features implies the longest training times at one or two hours each for the Ensemble models.

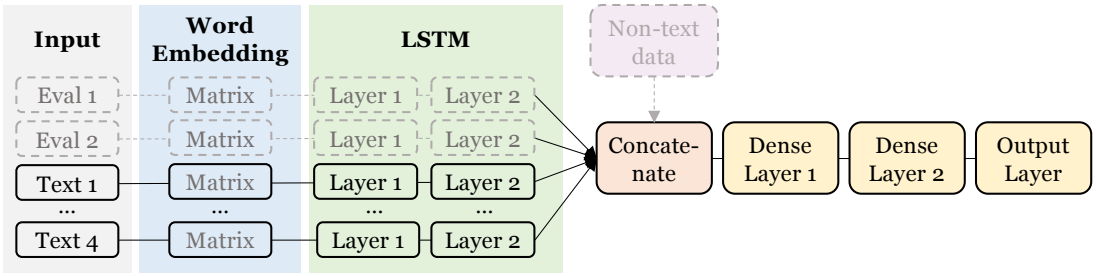
5.2 Recurrent Neural Network

Related work in this field demonstrates the inferior performance of word embeddings when element-wise averaged and used as inputs for models from the previous section (Funk et al., 2020). Due to the relatively long input sequences in the second session (on average 370 words respectively 392 with evaluation texts), we assume that a carefully designed recurrent neural network can better leverage the potential of word embeddings than averaged word vectors and thus possibly achieve better results on our two classification tasks. To avoid enlarging the input sequence length further, we do not include text generated in the first session.

A naïve bidirectional LSTM-based (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997) model architecture, which consists of one input containing all text inputs of a given participant, barely achieves baseline performance on our validation set. This may be grounded in challenges arising from these long input sequences. Therefore, we decide to design a more sophisticated, task-specific model (Figure 14) for our problem. The core model has four different blocks which aim to encode the participants' texts with respect to one of the four categories used in the second session – problem solving, reflection, stress analysis, and behavioural planning – and thus naturally reduces the input sequences' lengths. Each block consists of an input layer, an embedding layer (i.e., our pre-trained word embedding matrix), and two bidirectional LSTM layers. All outputs from the last bidirectional LSTM layers are concatenated and passed to a fully-connected neural network with dropout. Adding a further bidirectional LSTM layer after concatenation does not improve performance on our validation set. We consider the embedding dimension (FastText: 10, 25, 50, 100; Word2Vec: 25, 50, 100, 300), input sequence length (30, 50, 100, 200 words), number of units per LSTM layer (First layer: 0, 16, 32; Second layer: 16, 32), number of neurons per dense layer (16, 32), and the dropout rates (0.1, 0.2) as hyperparameters which are optimised during training. If further text inputs are considered (i.e., evaluation texts), we extend our core model by two blocks processing the two different evaluation categories (i.e., feedback about liked contents and suggested improvements). If numerical inputs are considered (i.e., demographical data, numerical evaluation data, or extracted numbers from text), we extend our core model by another input layer which is normalised and directly passed to the concatenation layer. We try to account for the imbalanced class distribution by using a weighted binary cross-entropy loss function. The class weights are considered hyperparameters which are optimised during training. The Adam optimiser is used to train this network architecture where the learning rate (0.01, 0.001, 0.0005) yields the final hyperparameter. To tune all hyperparameters, we use 20% of the training data as a validation set and re-train our tuned models on the entire training set for 25 epochs with early stopping. Since the performance does not increase when fine-tuning the embedding layers, we freeze the embedding weights

and only train the remaining weights of the network. The network is implemented in TensorFlow, and the hyperparameter tuning is executed on an Nvidia Tesla P100, which takes approximately six hours for each of the four different data inputs.

Figure 14: Task-specific LSTM-based model architecture



5.3 BERT

To represent the more complex recent transformer model architectures, we investigate the prominent bidirectional encoder representations from transformers (BERT) model (Devlin et al., 2019) to predict dropout and intervention failure. In contrast to the previous approaches of separating the steps of text representation and training classification models, the BERT model combines these tasks (Devlin et al., 2019). While training BERT from scratch requires a substantial amount of data, BERT models pre-trained on large datasets can be leveraged and are easily adaptable to new NLP tasks. On NLP benchmark tasks, pre-trained BERT models that are fine-tuned on custom datasets achieve better results than carefully crafted task-specific model architectures. Therefore, we also follow this approach to both maintain the sequential structure of the texts and to reduce the manual effort in designing an appropriate architecture.

We build our classification model based on the BERT model pre-trained on three large German datasets (bert-base-german-cased from Huggingface's model repository) and fine-tune it on our dataset. To adapt this model to our two classification tasks, we slightly modify the model architecture: we use the 768-dimensional representation vector produced by the BERT model and feed it into a new classification head consisting of two hidden layers and a sigmoid output layer. When considering additional numerical inputs (i.e., baseline, evaluation, or closed-question variables), we concatenate the 768-dimensional vector with the supplementary variables. The design of the classification head is optimised during training where the number of neurons per hidden layer (16, 32, 64) and the dropout rate (0.1, 0.2) are considered hyperparameters. Despite BERT's ability to handle input sequences up to 512 words, we only consider shorter lengths (64 and 128 words) due to the required computational resources. To compensate for the class imbalances, we make use of a weighted binary cross-entropy loss function and treat the class weights as additional hyperparameters. The aforementioned hyperparameters and the learning rate ($5 \cdot 10^{-4}$, 10^{-5}, $5 \cdot 10^{-5}$, 10^{-6}) of the Adam optimiser are optimised during training using a validation set of 20% of the training data. The final model is trained on the entire training data for 20 epochs with early stopping.

6 Results

We evaluate our final models on the test set of 170 (dropout) and 163 (intervention failure) participants. Although our test set is large when compared to most related work, this size still implies the risk of unrepresentative results. Since the area under the receiver operating characteristics (AUC) accounts for class imbalance and thus eases the comparison of results of the two classification tasks, we choose this evaluation metric (Bradley, 1997). The two result tables for intervention failure (Table 6) and intervention dropout (Table 7) summarise the AUC scores on our test set, where columns represent different text representation methods and rows define supplementary non-text features. The benchmark model (BM) column provides a reference score trained exclusively on the corresponding numerical features. To identify the most predictive features, we calculate SHAP values (Lundberg & Lee, 2017) or use included feature importance measures for the non-sequential models.

6.1 Intervention Failure

Exclusively considering *text variables*, sentiment analysis (AUC of 0.65) outperforms the other text representation techniques on our test set. Other methods, such as Word2Vec combined with our LSTM architecture and advanced metadata with an LR model, achieve solid results (0.61 resp. 0.59) as well. While averaged word vectors combined with boosting classifiers perform very poorly (0.50-0.52), leveraging the sequential nature of the texts by using deep learning architectures yields benefits (AUC 0.55-0.61). Thus, also performing equally or better than TF-IDF features with an AdaBoost model (0.55).

Our benchmark model (BM) trained on the numerical *baseline data* achieves an AUC score of 0.69 and, hence, is not outperformed by the vast majority of text representation techniques. This is most likely due to the initial PSS subscores included in the baseline data, which are expected to be important variables in predicting intervention failure (Forsell et al., 2019). While the additional baseline variables increase the performance of all text representation methods (compared to text-only models), only advanced metadata and sentiment analysis (both combined with LR) achieve better AUC scores (0.71 and 0.72) than our baseline benchmark. In both cases, age is the most important feature, followed by the baseline PSS subscores and income category. While PSS subscores are among the five most important variables of our benchmark model as well, age and income are not, which possibly indicates a moderating function for text features. The deep learning approaches are the only approaches that do not benefit from adding baseline data and perform worse than the averaged word vectors combined with LR. Similar to baseline data, additional **evaluation data** (both textual and numerical) enhances the performance of nearly all representations. Besides the winning task-specific Word2Vec LSTM architecture (0.68), advanced metadata, sentiment analysis, and LDA (all using LR) achieve better results than the evaluation data benchmark by itself (0.62). TF-IDF, averaged word vector, and the remaining deep learning approaches cannot attain the benchmark scores, suggesting that more variables can have a harmful effect on the information-to-noise ratio. *Closed-question data* adds little

value and, in some cases, even decreases the model performance when compared to text-only results. Only the task-specific FastText LSTM model leverages this additional information and achieves an AUC result of 0.62. While this clearly outperforms the benchmark (0.53) as well as the averaged word vectors on this task, various other approaches on different data inputs achieve better results.

To predict intervention failure, baseline data containing initial PSS subscores clearly benefits the models' performances. On our test set, sentiment analysis and advanced metadata approaches yield solid results which perform better than benchmark models and other approaches considered. On average, advanced metadata (0.63) performs slightly better than simple metadata (0.60), offering evidence that the nature of the exercise done matters for the intervention outcome. Analysing the model coefficients of our advanced metadata reveals that the largest coefficients are assigned to the text length of tasks concerning problem reflection, behavioural planning, and motivation. Since this model aims to predict rather than to explain, further research is necessary to investigate the causality. Among the two best-performing non-sequential approaches, LR and SVM are most frequently chosen (6 out of 8). BERT, TF-IDF (primarily used with boosting classifiers), and averaged word embeddings often perform below our benchmark. Although we demonstrate, on our test data, that word embeddings combined with our task-specific architecture often outperform the averaged word vector approaches, the deep learning models fail to achieve benchmark scores in many cases.

Table 6: Result table intervention failure prediction

AUC	BM	Sim. MD	Adv. MD	TF-IDF	Sent-iment	LDA	W2V Avg.	FT Avg.	W2V NN	FT NN	BERT
Pure Text	0.500	0.546	0.588	0.545	0.649	0.550	0.500	0.522	0.605	0.553	0.550
Baseline	0.688	0.687	0.710	0.687	0.719	0.687	0.687	0.687	0.635	0.617	0.581
Eval.	0.624	0.623	0.631	0.520	0.649	0.629	0.615	0.596	0.676	0.592	0.591
Closed-Q.	0.529	0.524	0.578	0.554	0.577	0.534	0.508	0.530	0.572	0.623	0.564
Average	**0.614**	**0.595**	**0.627**	**0.577**	**0.649**	**0.600**	**0.578**	**0.584**	**0.622**	**0.596**	**0.572**

Notes: BM = Baseline; Sim./ Adv. MD = Simple or advanced Metadata; W2V = Word2Vec; FT = FastText; NN = Task-Specific neural network; Eval. = Evaluation Data; Closed-Q = Closed Questions

6.2 Dropout

Despite the theoretically assumed interrelation between dropout and intervention failure (Barrett et al., 2008; Donkin et al., 2011; Gan et al., 2021), well-performing text representation approaches and ML models differ significantly on our dataset. While TF-IDF and the deep learning approaches perform poorly in many settings when predicting intervention failure, these approaches, as well as the simple metadata approach, dominate the results for dropout prediction. On pure *text data*, simple metadata combined with a non-linear kernel SVM classifier yield the best AUC score (0.65), closely followed by TF-IDF combined with XGBoost (0.63) as well as the Word2Vec (0.64) and FastText (0.63) task-specific LSTM models. Word embeddings combined with our LSTM architecture increase performance in comparison to averaged word embeddings and an SVM classifier (0.53 and 0.61). Previously well-performing approaches

such as advanced metadata and sentiment analysis score mediocre results (0.54 resp. 0.56) on the task of dropout prediction. Akin to the intervention failure prediction, model performances generally benefit from additional *baseline and evaluation variables*. Yet, for dropout prediction, evaluation data has a stronger impact, supporting the hypothesis that a participant's opinion on the intervention is a good predictor for discontinuation. The task-specific LSTM-based approach on the Word2Vec embeddings scores the best results (0.70) when using additional baseline variables, and other deep learning approaches also perform well (0.65) in this setting. TF-IDF features used with LR likewise achieve an AUC score (0.64) well above the benchmark (0.60) on this task. Simple metadata combined with LR (0.69), and our fine-tuned BERT model (0.67) yield the best results when harnessing supplementary evaluation data. SHAP values, calculated for the evaluation benchmark and simple metadata model, suggest that the number of useless entries, the session's perceived usefulness, and time adequacy are the most important features in this setting. Our FastText approach slightly surpasses the benchmark of 0.65 on the evaluation data. Using additional *closed-question data* mostly enhances the performance. BERT (0.67), FastText (0.66), Word2Vec (0.64), and SVM trained on TF-IDF features (0.65) clearly outperform the benchmark model (0.58). Averaged FastText (0.63) features combined with SVM achieve solid results, however, they cannot reach the results of our LSTM architecture.

In most cases, adding non-text data increases the model performance, most evidently in the case of evaluation data. The most basic approach considered (simple metadata) outperforms all other approaches when working on pure text data as well as in combination with evaluation data. Thus, a participants' attitude in combination with how much they write is an easily attainable and well-performing prediction set-up. Among the non-neural approaches, at nine times, SVMs with the non-linear kernel are the most commonly chosen classifiers, with an additional two wins for linear SVMs. At six or seven each, LR, AdaBoost and XGBoost do not differ much in how often they were chosen. The more sophisticated approaches (TF-IDF, word embeddings in combination with task-specific LSTM architectures, and BERT) constantly achieve good results, and on average yield the best AUC scores on our test set. We notice a pattern in the embedding dimension and input sequence length hyperparameters: the most prominent embedding dimension among Word2Vec models is 25 with a maximum input sequence length of 100, whereas FastText models prefer shorter sequences of 50 words and embedding dimensions of 10 or 25. These findings also hold when predicting intervention failure, thus indicating the need to treat these numbers as hyperparameters instead of choosing default values. Furthermore, most models do not benefit from the second (optional) LSTM layer, which points towards an overwhelming model complexity considering our dataset size.

Table 7: Result table dropout prediction

AUC	BM	Sim. MD	Adv. MD	TF-IDF	Sent-iment	LDA	W2V Avg.	FT Avg.	W2V NN	FT NN	BERT
Pure Text	0.500	0.651	0.541	0.626	0.559	0.547	0.531	0.610	0.644	0.630	0.618
Baseline	0.596	0.614	0.633	0.642	0.588	0.617	0.662	0.551	0.696	0.645	0.646
Eval.	0.649	0.686	0.592	0.651	0.636	0.643	0.639	0.601	0.636	0.656	0.668
Closed-Q.	0.584	0.554	0.563	0.652	0.599	0.606	0.575	0.632	0.639	0.663	0.668
Average	**0.610**	**0.626**	**0.582**	**0.643**	**0.596**	**0.603**	**0.610**	**0.607**	**0.654**	**0.649**	**0.650**

Notes: BM = Baseline; Sim./ Adv. MD = Simple or advanced Metadata; W2V = Word2Vec; FT = FastText; NN = Task-Specific neural network; Eval. = Evaluation Data; Closed-Q = Closed Questions.

6.3 Discussion of Clinical Usefulness

As discussed by several authors such as Olczak et al. (2021), Cabitza and Campagner (2021), and Scott, Cater and Coiera (2021), prediction performance metrics are only one subdimension when evaluating ML models in healthcare settings. Therefore, we use the ten questions proposed by Scott, Cater and Coiera (2021) to summarise and evaluate the prospective clinical value of the proposed winning models.

(1) What is the purpose and context of the algorithm? The pain points the respective algorithms address are (1) high dropout rates and (2) low response rates in DMHIs in light of limited resources. The proposed models provide insights into who will likely drop out or not benefit after two out of eight sessions. As such, these predictions serve to adapt individual treatment plans (e.g., through additional guidance, sessions, or reminders) only if and where necessary.

(2) How good were the data used to train the algorithm? We use the five categories (i.e., completeness, correctness, concordance, plausibility, and currency) to assess data quality for clinical research proposed by the review of Weiskopf & Weng (2013). Regarding completeness, the data consists of all information else provided to the interventions' e-coach for decision-making. Further, it spans a large variety of what previous work found relevant for intervention dropout and outcome. While additional outside information, such as previous health records or expert assessments, could possibly improve the predictions, the effort necessary to collect them requires extensive steps, deteriorating the cost-value ratio. Since the data stems from RCTs, research staff monitored the completeness of entries and missing data was very low, as seen in Supplementary Material 1. As for correctness, all non-text dimensions were manually investigated by the two first authors to find mistakes, and data quality was found to be high. The fact that spelling-mistake correction did not increase cross-validation scores indicates a good quality of the text data. Concordance of the data was, for example, internally validated by cross-checking modules completed with the submitted answers, running pivot tables for related variables (e.g., current employment status and leadership responsibility), and ensuring the correct time sequence of the entries. To check for plausibility, every feature's range and distribution were manually checked by two authors. Questions and findings, including averages and ranges, were discussed with the

third author, who was involved in the data collection to check for plausibility, and no issues remained open. The currency of the data is high as the nature of the online set-up allows the instant use of the data as soon as the patient submits their answers. As such, a deployed model could inform clinical decisions immediately.

(3) Were there sufficient data to train the algorithm? The data set at hand is comparatively small for Data Science applications in general, thus presenting one of the major limitations of this study. Especially deep neural networks usually require large amounts of data to perform well. At the same time, this is a prerequisite that is rarely met in E-Mental Health research (Pasini, 2015), and with almost 850 participants, the data set is large for DMHI standards. As seen in the related work section, only one other paper considering intervention data exceeds the dataset size presented in this work. A literature review found a dataset size of 100 to be minimally adequate for outcome predictions in DMHIs (Sajjadian et al., 2021), but only 44% of the 56 studies investigated complied with this criterion. Further, they found that only 29% used a hold-out test set or adequate cross-validation method. At a test set size of 163/170 that was not used for training at any point, the results at hand can be considered among the more generalisable of the works currently available (Sajjadian et al., 2021). To address the small dataset size, we extend the pre-training corpus with texts generated by control group users and train the word embeddings at the sentence level. Further, our use of a pre-trained BERT model comes with the significant advantage that - as researchers from a field struggling with data collection - we can leverage large unrelated but available data sets (Pedersen et al., 2019). The results for the deep learning models are stable and good within and across different settings. This suggests an at least minimally adequate data set size for them to compete with classical Machine Learning models.

(4) How well does the algorithm perform? With almost all average AUC scores well above 0.5, it can be concluded that the considered features have predictive power regarding intervention outcome and dropout. With the best scores reaching an AUC of 0.70 (dropout) and 0.72 (intervention outcome) after just two weeks, results are competitive with related work. For example, Bremer et al. (2020) achieved an AUC of 0.6 when using the user journey data (e.g., time spent) of their first two out of seven sessions to predict dropout. The best prediction models proposed by us achieve a balanced accuracy of 0.66 and 0.67. Forsell et al. (2019) did not reach similar balanced accuracy scores predicting outcome with only symptom data until week 3 or 4. The comparison to other related works is limited due to differences in baselines and time horizons. The performance in the sense of clinical usefulness will be discussed in question 8.

(5) Is the algorithm transferable to new clinical settings? The specific models with their respective (hyper)parameters and, in the case of the NNs, task-specific architecture, can likely not be deployed on a different intervention. However, the proposed process to train the two best-performing models can be replicated on any dataset including intervention text and socio-demographic data. As can be seen in the related work section, these are very common data types to be collected in a standard DMHI setting. The text pre-processing steps are generalisable for any German text and would only have to be slightly adapted for English text (i.e., different handling of capital

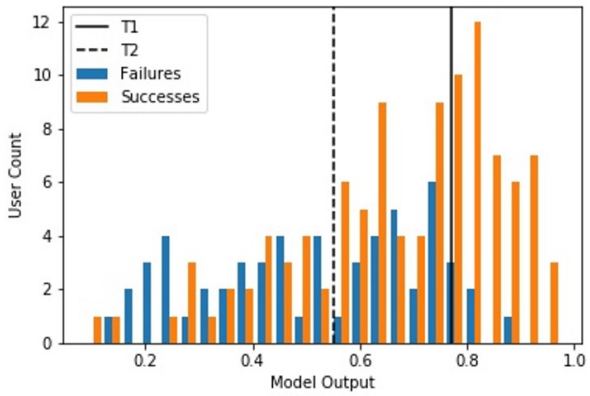
cases). Transferring models from one language to another in the clinical context has been shown to be possible in other tasks, especially for languages from the same family (Névéol et al., 2018). The fact that pre-trained neural networks for English text are more in number and more specific in problem-fit (Ji et al., 2021) indicates that the prediction results of the neural networks could even improve for the English Language. Once text features are produced, they can easily serve a variety of outcome measures. The related work section shows several options, ranging from 24-hour symptom prediction to 6-month follow-ups. Other options could be to use it to personalize content or adapt the time of intervention.

(6) Are the outputs of the algorithm clinically intelligible? Considering the transparency of the decision process, neural networks' black-box nature is one of their major drawbacks. For the non-sequential models, SHAP values and built-in feature importance measures give first insights into the decision-making process. These efforts can easily be extended per the suggestions made by Yang (2022) but are left for future research as interpretability is not the focus of this paper. However, the actual outputs of both models are binary and easily understandable as they represent dropout vs. completers and intervention successes vs. failures per the above-given definitions.

(7) How will this algorithm fit into and complement current workflows? As of now, e-coaches receive general guidelines on how much time to spend with their allocated participants. Within a given RCT, these suggestions did not differ across participants. Once implemented, the models' predictions could prompt individual suggestions. For example, a stop-light system could indicate green (no risk), yellow (moderate risk), or red (high risk of dropout) (Forsell et al., 2019). With this information, therapists or e-coaches can decide or be instructed as to which participant is most in need of their time. Pedersen et al. (2019) report this approach to have been positively received by therapists in their study. Such risk profiles could also prompt automatic reminders, personalized feedback loops to identify the problem, or additional content (e.g., a module regarding motivation or goal setting).

(8) Has use of the algorithm been shown to improve patient care and outcomes? The next step to evaluating the practical value of the proposed model is the implementation within a live intervention. However, this exceeds the limits of this paper. At the same time, studies such as Forsell et al. (2019) and Pedersen et al. (2019) have empirically proven the superiority of adaptive care for both dropout and outcome predictions. In the baseline, the limited resources are currently being distributed at random. Empirical evidence shows that many patients benefit from unguided interventions (Richards & Richardson, 2012) and Forsell et al (2019) show that at-risk patients – while significantly benefiting – even with enhanced care, barely reach the same health outcomes as not at-risk counterparts.

Figure 15: Histogram failure prediction out-



The best model predicting outcome recognises 93% of intervention failures (recall) while avoiding overspending on 41% of the most likely completers (specificity). The same calculations for the slightly less balanced dropout predictions lead to 55% correctly identified dropouts while avoiding overspending on 80% of completers. These metrics can be off-traded through the threshold deciding between a dropout or failure, as exemplarily shown in Figure 15. The histograms show the intervention failure probability as predicted by the winning model for each group – intervention failures and successes. As expected, successes have a higher probability of being recognised as such (right side), failures are more prevalent in the low probabilities (left side), and there is a bulk of hard-to-identify participants in the middle. Changing the threshold from T1 (highest balance accuracy) to T2 decreases the recall to 53%; however, it avoids overspending on 75% of successes. Consequently, not much more than one-third of all participants receive enhanced care, lowering costs significantly while still addressing those most likely participants to benefit from support. The threshold can be adapted to fit the available resources and can even inform the number of participants accepted in the intervention. Considering the preventive nature of the stress intervention at hand, one application of the model could be to make the intervention available without guidance to reach as many participants as possible and only offer the available guidance to those who most need it. In the T2 scenario in Figure 15, this increases the number of participants reached by threefold.

(9) Could the algorithm cause patient harm? The purpose is to optimise resource allocation while maintaining or improving the level of care over the entire population of participants. If it were used to reduce the average care level, it could harm those incorrectly classified as completers or successes through decreased levels of care. Such as prospect is especially worrisome when working with a population with severe symptoms. Depending on the importance of avoiding such false negatives, the recall can be increased at higher costs of resources. It, thus, must always be closely considered *how* to implement such a decision-support tool in *which* setting. At the same time, considering that right now, very limited resources are available, and many sick people are not being helped at all, increasing the total number of participants treated is a factor to weigh in with individual effects.

(10) Does use of the algorithm raise ethical, legal or social concerns? Albeit the focus of early research being on establishing the overall feasibility, bias in the data must be considered early on. With primarily female participants that hold a university degree, the data at hand is - while typical for mental health interventions – not representative of the general population. Implementing such a model in routine care could disadvantage those groups with the already most extensive unmet needs and must be

adjusted to ensure the best possible care for all (Gianfrancesco et al., 2018). In addition, the ethical and legal aspects of an automated decision to change the level of care must be closely considered, especially in cases where the reason for the prediction is not transparent (Yang, 2022).

7 Conclusion

NLP methods can help make countless individual predictions based on text that would require impossible amounts of human resources to be analysed. While the use of sophisticated NLP methods on non-clinical texts is continuously advancing in Mental Health diagnostics (Bucur et al., 2021; Cohan et al., 2018; Nobles et al., 2018; Wołk et al., 2021), applications of NLP on E-Mental Health intervention text have been few and predominantly limited to simple models. In this case study, we train several ML models, considering various text representation methods and additional data inputs, to predict intervention failure and intervention dropout. For this, we use a dataset of 849 German-speaking participants of a stress intervention. By thoroughly evaluating combinations of the above-mentioned factors on our dataset, we contribute to the design choice of prediction models for intervention dropout and intervention outcome.

First, we demonstrate that *harnessing the sequential nature of text* by training deep learning models in combination with word embeddings outperforms the much simpler approach of using averaged word vectors on our test set. Thus, we complement existing research (Funk et al., 2020; Gogoulou et al., 2021) by proposing a task-specific LSTM architecture using word embeddings which successfully deals with the long input sequences and yields good results (average AUC score of 0.65) in dropout prediction. We further demonstrate the need to treat the embedding dimension as a hyperparameter rather than using the default values. Second, considering *supplementary baseline data* when predicting intervention outcome and *evaluation data* when predicting dropout yields the best-performing models. Thus, our findings support that the participants' backgrounds and attitudes towards the intervention hold additional information in combination with text data. Third, we underline the *solid performance of easy-to-implement approaches* to predict dropout (simple metadata and TF-IDF) and intervention outcome (advanced metadata and sentiment analysis). By providing the insights from our case study, we seek to facilitate the development of ML-based tools which augment e-coaches' work in extracting valuable information from the participants' intervention texts – Hence, easing the task of identifying participants in need of human attention. With these predictions, necessary steps towards a more successful intervention in light of limited resources to face growing needs can be initiated.

Considering the still relatively small data set size and high specificity of our intervention set-up, this research is only a step towards better understanding, predicting, and ultimately influencing participants' behaviour in DMHIs. Data sets such as this one can be considered the most promising approach to gathering knowledge in this research area. Yet, learning on few data points might not champion the same text representation methods and models, and more research is necessary to determine the generalisability of our findings. While we prove the potential of neural networks in this

setting, they require large datasets, long training times, and have a black-box nature. However, the investigation of such complex methods is necessary to ensure the best possible results – especially considering the astonishing results deep learning models achieve on other NLP tasks. To truly understand human language, words must be considered beyond their lexical meaning, and the specific context needs to be understood – a task which simple methods will never solve. One further way to address the problem of small datasets could be to use data augmentation methods as commonly used in computer vision, and more recently proposed for NLP tasks [99]. We suggest that employing attention-based [100] deep learning architectures can further enhance the model performance in prediction tasks such as ours. While designing task-specific network architectures like ours may be a complex and tedious task, large pre-trained text classification models can eliminate this work. To determine whether further research in applying pre-trained transformer models in this domain is auspicious, we examine the most prominent transformer model BERT and observe promising results in dropout prediction. Thus, we suggest investigating more sophisticated pre-trained transformer models (e.g., RoBERTa (Y. Liu et al., 2019) or XLNet (Yang et al., 2020)) in such settings. In addition to an optimised pre-training strategy, XLNet tends to process long sentences better than BERT, which could be advantageous in cases like ours and further improve the model performance. Besides the particular transformer model, the text corpora used for pre-training, as well as the approaches to integrating the important non-text features into the model architecture, should be investigated in more detail (e.g., Shen et al., 2020). Furthermore, multi-task models (e.g., predicting intervention failure and dropout at the same time), which are frequently employed in other NLP tasks (e.g., Hashimoto et al., 2017), can potentially improve results on both tasks. For the time being, simple feature representations such as metadata and classical statistical models should be considered as an easy-to-implement yet competitive option for predicting intervention failure and dropout. In that regard, further research must be conducted to investigate how to improve these predictions, for example, more automated ways of finding the most important TF-IDF features (Zhang et al., 2020).

Chapter VI: Explainable AI and Trust

Zantvoort, K., Bjurner, P., Forsell, E., Wallert, J., Funk, B., & Kaldo, V. (working paper) Opening the black box – Effects of decision transparency on therapists' trust in and intended use of AI-based decision support systems in ICBTs

Preliminary Note: At the time of this thesis submission, the data for this study is still being gathered. Therefore, the following is a short version focusing on a selection of ten measures covering the key concepts decision transparency, trust and clinical actionability of the DST. As such, the preliminary results included all case-based scores but did not consider the open-text answers. Regarding the other questionnaires (see (Bjurner, Zantvoort et al. (2024)), the primary group excluded from this preliminary analysis were the preference and user-friendliness of the DST versions and a pre-post analysis of the perceived competence of a DST. The selection of these questions and formulation of the hypotheses were completed before the analysis of the data. KZ and PB share first authorship.

Abstract: Clinicians' confident use of Decision Support Tools (DSTs) is crucial in realising the proposed value of Machine Learning-based precision care. However, user adaption has been hindered by the limited decision transparency and subsequent mistrust in Machine Learning predictions. Therefore, this study investigated the impact of local SHAP (SHapley Additive exPlanation) values on therapists' perceptions of a Machine Learning-based DST in Internet-based Cognitive Behavioural Therapy (ICBT). The randomised experiment included 35 Swedish ICBT therapists who were each presented with a DST with six exemplary patient cases with SHAP values and six without them. The study measured therapists' understanding, trust, and perception of clinical usefulness through self-ratings at multiple stages. The primary hypothesis that adding SHAP values increased self-reported trust was confirmed by our findings ($p=0.01$, $d=0.43$ [0.07-0.76]). Further, our results suggested increased understanding, agreeance with, and the perceived usefulness of the DST predictions through SHAP values. However, this was not reflected in a change in intended treatment support levels. Consequently, while increasing the therapists' comfort level regarding Machine Learning-based DSTs, it remains unclear if adding SHAP values yields any change in clinical practice. As such, the study at hand serves as a first exploration of the effects of local explanations in an AI-based DST in mental health care.

1 Introduction

Artificial intelligence (AI)-based personalization shows immense potential to transform healthcare (Udegbe et al., 2024). One promising example is Decision Support Tools (DST) for psychological treatments (Sajjadian et al., 2021; Zajac et al., 2023). In Internet-based Cognitive Behavioural Therapy (ICBT), scarce resources such as therapists' time could easily be personalized, allowing for optimised resource allocation and better health outcomes (Forsell et al., 2019). Ample single (e.g., (Chekroud et al., 2016; Gogoulou et al., 2021; Hornstein et al., 2021) and first meta-studies (Lee et al., 2018; Sajjadian et al., 2021; Vieira et al., 2022) showcase the feasibility of Machine Learning (ML) outcome predictions in the mental healthcare context. However, so far, most of them are limited to proof-of-concept studies, and very few systems have been implemented (Hornstein et al., 2023; Triantafyllidis & Tsanas, 2019; Yin et al., 2021).

Integrating ML predictions into clinical decision-making comes with various challenges (Zajac et al., 2023). Changing the care provided carries a high ethical burden, as, among others, underserving a sick patient is an error to be urgently avoided (Diprose et al., 2020; Duffourc & Gerke, 2023). This importance is reflected in healthcare providers' legal liability for any potential harm caused by ML-based decisions (Duffourc & Gerke, 2023). Consequently, in current applications, therapists provided with an ML-based prediction usually retain decision autonomy over using this information (Forsell et al., 2019; Lutz et al., 2023; Pedersen et al., 2019). This setting makes the therapists' confident use of the information provided paramount for success (Khairat et al., 2018; Tonekaboni et al., 2019; Watson et al., 2019).

However, even in light of good prediction accuracy, clinicians can be hesitant to use predictions when unsure how they were created (Berner & La Lande, 2016). Such lack of decision transparency is a central hurdle in adopting AI-based DSTs, further amplified when patients demand to understand decisions themselves (Diprose et al., 2020; Jacobs et al., 2021; Yang, 2022). The challenge is that the most accurate models tend to be hard – if not impossible - to understand (Duffourc & Gerke, 2023; Gunning et al., 2019). Sophisticated ML models innately rely on non-linear relationships and interaction effects, two aspects that are difficult to phrase into humanly understandable rules (Duffourc & Gerke, 2023; Gille et al., 2020; Gunning et al., 2019).

In their recent rise, explainable artificial intelligence (XAI) methods strive to consolidate the interests of accuracy with decision transparency (Kennedy et al., 2021; Vilone & Longo, 2021a). Retrospective explanatory methods can be divided into (1) information-based (global) and (2) instance-based (local) explanations. The first covers questions such as the input data, model type, and other meta-information about the data and performance. The second focuses on single predictions, hence, insights into the individual patient at hand (Yang, 2022). A key difference is that for local explanations, the same value of a feature may have different impacts depending on the patients' other characteristics. For example, a decision tree model may consider the same variable at different points and in different ways to accurately depict complex interaction effects.

SHapley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017) are a model-agnostic method for generating both global and local explanations. SHAP values are based on game theory and calculated by approximating individual predictions with and without said feature, indicating how much it contributed to the prediction. Summing over all individual predictions returns a global estimate which has been used to retrospectively explain feature importance in the ICBT context (Bremer et al., 2020; Hornstein et al., 2021; Ramos et al., 2021; Rodrigo et al., 2021; Zantvoort et al., 2023). However, so far, no use of individual (hence, local) SHAP values in the context of ICBTs is known, likely because the details of individual SHAP have limited use in retrospective analyses (Vilone & Longo, 2021a). Simultaneously, their detailed focus on a single patient makes them highly valuable when therapists must make timely and individual decisions. As Diprose et al. (2020) showed in their study, individual patient information significantly increased general practitioners' understanding, trust and confidence in using a prediction to decide the treatment course for a pulmonary embolism. However, it remains unclear if the same applies to therapists in mental health settings.

The current study, thus, aims to explore how the addition of SHAP values affects therapists' attitudes towards an ML-based DST in internet-based psychological treatment. To this end, the then hypotheses across the following three concepts are tested:

I. *Understandability*: Individual SHAP values will facilitate the therapists' understanding of predictions.

II. *Trust*: Providing therapists with SHAP values increases their trust in the DST.

III. *Clinical usefulness and actionability*: SHAP values will increase the therapists' perception of clinical usefulness and their likelihood to adapt treatment according to the prediction.

To test the hypotheses, a SHAP and a non-SHAP version of an otherwise identical DST for historic patients receiving ICBT for depression were created and presented in a randomised manner (SHAP/ Non-SHAP) to 35 Swedish ICBT therapists. Therefore, we explore and quantify the influence of local explainability methods on therapists' perception and utilisation of AI-based DSTs in internet-based CBT.

2 Hypotheses

Despite a growing body of literature (Nauta et al., 2023; Vilone & Longo, 2021b; Zhou et al., 2021), no consensus on the definition or evaluation of explainability exists. Generally, human-centred subjective evaluations (Colin et al., 2022; Hoffman et al., 2018) are differentiated from objective quantitative evaluations (Nauta et al., 2023). This human-centred evaluation study aims to gain insights into how XAI affects users along the commonly investigated concepts of trust, understandability and usefulness (Vilone & Longo, 2021b; Yang & Wibowo, 2022; Zhou et al., 2021). Therefore, in the next section, the hypotheses related to these concepts are described in more detail, including the measures used to analyse them. All individual translated questions are available in Supplementary Material 1.

2.1 Understandability

Considering their crucial role in the assumed mechanism of XAI (Diprose et al., 2020; Gille et al., 2020; Kennedy et al., 2021), the following hypotheses regarding decision transparency and understandability are evaluated:

H.1. *Decision transparency:* The scores on how well therapists understand how the respective version of the DST came up with their predictions are higher for the SHAP condition.

H.2. *Comprehensibility*: Therapists score the comprehensibility of the DST higher for the SHAP version of the DST.

2.2 Trust

The primary hypothesis of this paper is that SHAP values increase the therapists' trust in the prediction of the patient's final treatment outcome (Diprose et al., 2020). However, ample ways to define and measure trust exist (Vereschak et al., 2021; Yang & Wibowo, 2022). In this study, firstly, it is measured through the subjective, self-reported trust measures. Secondly, as an attempt at a more objective approach, we postulate that higher trust manifests in the extent to which therapists agree with the DST's outcome prediction (Yin et al., 2021). Lastly, the broader and previously validated human-computer trust scale (HCTS) is used (Gulati et al., 2019). The HCTS measures user trust on a multi-dimensional scale, i.e., concerning a) perceived risk, b) benevolence, c) competence, and d) reciprocity of a given system. Beyond filling the proposed placeholders, some of the HCTS statements were slightly reframed to account for the difference between the context-specific differentiation of the user (therapists) versus the patient.

Subjective trust

H.3. *Trust individual predictions*: The score of how much the therapists trust an individual patient DST prediction is higher when therapists are provided SHAP values than when they are not.

H.4. *Human-Computer Trust Score*: The total and the individual HCTS scores are higher for those who have seen SHAP values in the first block than those who have seen non-SHAP cases.

H.5. *Trust change*: The average score of whether the SHAP values increase trust in the DST prediction of treatment outcome will be above 4 (i.e. above "do not affect trust") in the direction of more trust.

H.6. *Comparative trust*: Similarly, the question of which version of the DST therapists trust the most will be above 4 (i.e. above "trust equally"), hence being biased toward the SHAP version.

Objective trust – disagreement

H.7. *Numeric disagreement*: The mean of absolute differences between the therapist's estimation of how probable a successful treatment outcome for a given patient is to the probability given by the DST is lower for the SHAP condition.

H.8. *Verbal disagreement:* When asked for a specific patient, therapists find it less likely that the DST missed or misjudged any factors for the SHAP cases than the non-SHAP cases.

2.3 Clinical Usefulness and Actionability

Finally, we postulate that SHAP increases the therapists' perceived clinical usefulness and actionability of the DST (Diprose et al., 2020). Again, this is measured through the subjective, self-reported measure of usefulness and the more objective, action-based measure of whether therapists are willing to change treatment based on the outcome prediction.

H.9. *Usefulness:* Therapists indicate a higher usefulness of the DST for guiding future treatment decisions for an individual patient in the SHAP condition.

H.10. *Intended support level:* When asked what treatment support level an individual patient should receive going forward, for the SHAP condition, therapists a) are more likely to deviate from the average treatment overall, and specifically b) more treatment is given for predicted failures and c) less treatment is given for predicted successes.

3 Methodology

This trial leveraged anonymized historical routine-care patient cases to evaluate a DST with and without SHAP explanations in a randomised experiment. Study participants were ICBT therapists in Sweden, and data was collected and managed between May and July 2024 using the REDCap (Research Electronic Data Capture) tool hosted at Karolinska Institutet. REDCap is a secure, web-based software platform designed for data capture in research studies (Harris et al., 2009, 2019). The following describes the creation, selection and presentation of the anonymized patient examples, including the outcome predictions and SHAP values. Next, we describe the study design, including randomisation, participant selection, and the questionnaires, followed by the analysis plan.

3.1 Predictions

The outcome predictions were based on 2,881 routine care patients with major depression disorder who received treatment from the Internet psychiatric clinic in Stockholm, Sweden, between 2008 and 2020. The intervention (Titov et al., 2018) and cohort details, data pre-processing and feature engineering steps have previously been published (Zantvoort et al., 2024). In summary, the model was trained on socio-demographics, symptoms, intervention interactions, and text metadata features from the first four intervention weeks. The individual features, their means, and standard deviations can be found in Supplementary Material 2.

To classify treatment success, the Montgomery-Åsberg Depression Rating Scale-Self (MADRS-S) for depression was used (Montgomery & Asberg, 1979; Svanborg & Asberg, 1994). A total of 2,650 patients filled out a symptom questionnaire after week eight of twelve and were included in the training data. The binary intended symptom

improvement was determined by either the final symptom score being below the remittance cutoff of 11 for MADRS-S (Fantino & Moore, 2009) or a 50% improvement since the start of treatment (Karin et al., 2018).

The model was trained via 5-fold cross validation with grid-search on 80% of data and evaluated on the 20% test set (N=530). In accordance with previous research on this data set (Hentati Isacsson et al., 2023; Zantvoort, Hentati Isacsson, et al., 2024), a Random Forest model was optimised on its number of estimators (5, 10, 25, 50, 500, 1,200), the minimum samples for a split (10, 25, 50, 100, 200), and the maximum depth of the trees (5, 10, 25, 50, 100, 500, 750). All steps were implemented in Python, using the pandas (McKinney, 2010), Numpy (Harris et al., 2020), and Scikit-learn (Pedregosa et al., 2011) libraries.

After pre-processing, the data set comprised 1,342 treatment successes and 1,308 treatment failures. The values chosen in the final model for the respective hyperparameters were 50, 50, and 50. The resulting model achieved a good (Kraemer et al., 2003) balanced accuracy of 0.70 and AUC of 0.78 on the test set. The model correctly identified 69% of failures with a precision of 71%. SHAP values were calculated for the test patients using the tree-explainer (Lundberg et al., 2020) of the Python SHAP package.

3.2 Patient Case Selection

In line with previous research (Forsell et al., 2019; Hentati Isacsson et al., 2023; Hornstein et al., 2021; Wallert et al., 2022), symptom scores were the most important source of information for symptom-based outcome prediction. To represent diverse cases, including more complex decisions, the 530 test cases were filtered based on their reliance on symptom features and the amount of information summed up under "other features". This process yielded 19 patients from the highest (>61%), 24 from the mid (38-61%) and 17 from the lowest third (≤37%) of success probabilities. Three researchers (PB, EF, JW) manually viewed the plots, successively narrowing them down to twelve cases, balanced across success probability (high, mid, low), age, and gender. All cases were anonymized before presentation by changing ages and patient background histories. All participants were shown the same twelve patient cases.

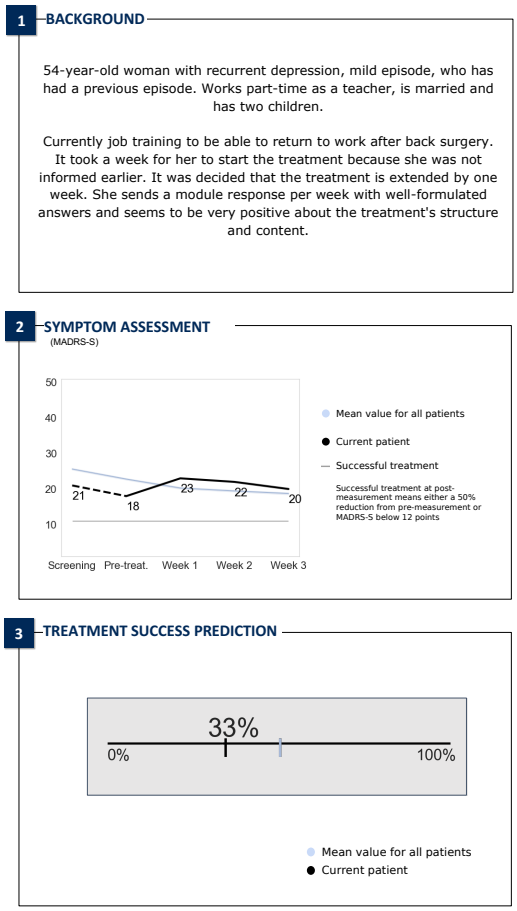
3.3 Case Presentation

The goal of the presentation of cases was to mimic a naturalistic setting of a DST in ICBT. In such a setting, a patient's baseline characteristics and history are known to a therapist. Therefore, a brief written background history (Figure 16a. and b. box 1) mainly described age and gender, primary and possible secondary diagnoses and clinical history. Further, details of the patient's occupation, household, social network, and upbringing were provided. Lastly, the progression pace, the patient's communicative style (e.g., long/short messages), attitude to the treatment, and their problems were described. A graph (Figure 16a. and b., box 2) showed the weekly symptom scores, including the cohort mean and the successful treatment cutoff, to provide the commonly available information on symptom development. The non-SHAP third box (Figure 16a., box 3) comprised only the prediction graph (0-100% success probability

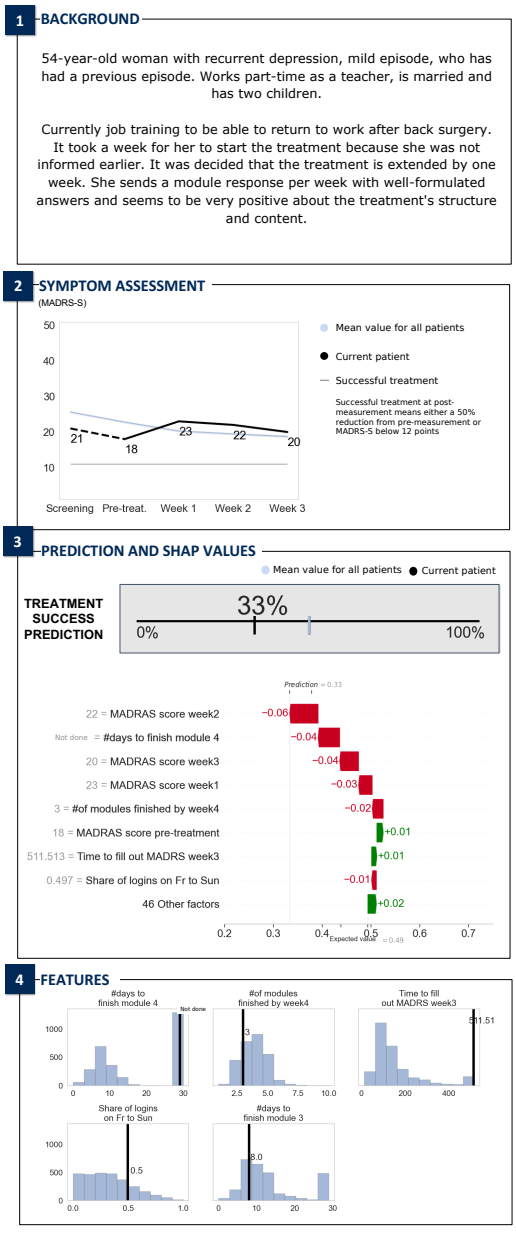
and mean cohort probability). In the SHAP condition (Figure 16b., box 3), the third box additionally showed SHAP-waterfall plots (green/red bars) for the features contributing the most to the prediction, plus a sum score of the other features. As some units of measurement (e.g. seconds) were uninformative by themselves, the SHAP cases also included histograms relating the patient's value to the cohort for the top non-symptom features (Figure 16b., box 4). Not providing symptom histograms avoided redundancy with box 2.

Figure 16: Decision Support Tool interface example case for both conditions

a) non-SHAP example



b) SHAP example



3.4 Randomisation

The same twelve cases were divided into two balanced sets, which alternated in a) SHAP or non-SHAP condition, b) in block 1 or 2, and c) in increasing or decreasing order (Figure 17). Thus, there were eight combinations and subsequent arms in total.

After inclusion, participants were randomised to one of these arms using the REDCap Randomisation Module by an independent person blind to the trial's purpose. An external statistician generated a block-randomised allocation table outside of REDCap, based on which participants received a personalized link to the specific arm.

3.5 Participants Screening and Inclusion

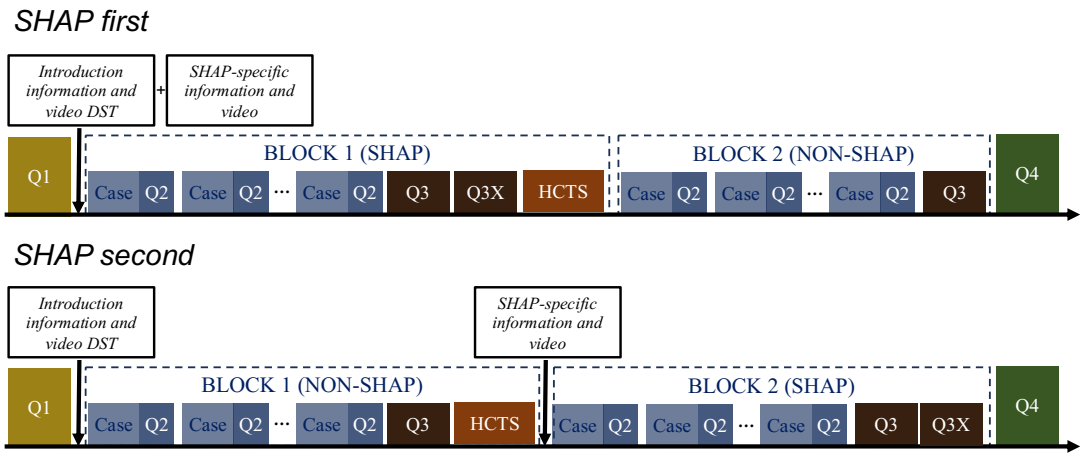
As described in the pre-registration of this study (Bjurner, Zantvoort, et al., 2024), 52 participants corresponded to a power of 80% to detect a within-group difference of Cohen's d of 0.4, i.e. a small to medium effect size. However, a multiple of eight, hence 56, was necessary to balance the conditions.

The study recruited ICBT therapists in Sweden via various channels, including networks for digital psychological treatment, social media, and direct contact via known individuals or the Swedish National Quality Register for internet-based psychological treatment (SibeR). Interested therapists signed up via the REDCap secure web application, where they were informed about the study details and data privacy, and consented to participate before filling out the screening questionnaire (Supplementary Material 3). The inclusion criteria for participation were a) having at least basic psychotherapy training in CBT in combination with an adequate care profession (i.e. nurse, physician, or mental health worker), b) having treated at least five ICBT patients or one ICBT patient plus at least 20 patients with traditional CBT, c) having treated the last ICBT-patient no more than three years ago, and d) having access to a desktop computer or laptop to conduct the trial. Each participant received a 15€ digital gift card as reimbursement upon trial completion. Email reminders, including the contact information of the study coordinator, were sent via REDCap automated invitation scheduling or manually for participants starting but not completing the different steps of the study.

3.6 Questionnaires

Analysed data comprised self-ratings at different points (Figure 17): Before seeing cases (Q1), after each patient case (Q2), after each block of cases (Q3), only after the SHAP block (Q3X) or only after the first block (HCTS), and finally, after having finished both blocks (Q4).

Figure 17: Overview of trial set-up in dependence of the SHAP condition.



Notes: *Content order differed between SHAP in the first or second block. While each block comprised six cases and all participants saw the same twelve cases, the cases differed in conditions (SHAP/non-SHAP), and order (e.g., 1-6 vs. 6-1)*

Q1 – Baseline: Upon study start, participants' general attitudes towards AI and ML and in a clinical setting were recorded. These baseline questions (Supplementary Material 4) included the understanding and expectation of performance and impact regarding AI/ML.

Information: Before showing any cases, participants were introduced to the ICBT program and DST via video and written information. The material covered the ICBT program for depression, the definition of treatment success, and the model set-up, including the features used. Following Yang's (2022) definition, the non-SHAP condition thus included global explanations of the ML model. Before starting the SHAP block, the participants watched an additional video explaining individual SHAP values and histograms and how they could be interpreted in this context. After both videos, participants answered quiz questions to measure their understanding of the information provided in the video. There were four questions after the general video and six after the SHAP video, all of which can be found in Supplementary Material 5.

Q2 – Twelve individual cases: Next, depending on the randomisation, a block of six (non-)SHAP cases was shown, followed by questions about trust in and the understanding and usefulness of the DTS version. Further, therapists provided their own prediction of treatment outcomes and opinions if the DST may have missed or misjudged any factors. Lastly, they rated the level of treatment support they thought the individual patient needed for a successful treatment compared to an average patient. These Q2 questions were repeated after each case.

HCTS – Only after the first block: After finishing the first block of cases, regardless of which condition, participants were asked to fill out the Human-Computer Trust Scale (HCTS) (Gulati et al., 2019). As the HCST comprises twelve questions, asking it only once was a trade-off between validated measurement and avoidance of too lengthy study and subsequent adherence issues.

Q3 – After each block: After each block, participants rated the previously seen DST version regarding trust and understandability. For the SHAP condition (Q3X), additional questions indicated the perceived added value of the SHAP explanations.

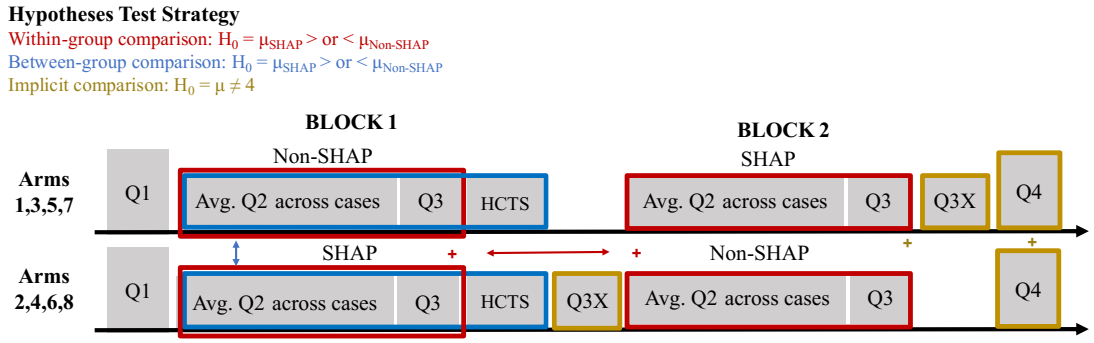
Q4 – Questions after finishing both versions: Finally, the participants answered how the two versions compared in terms of participant trust.

3.7 Statistical Analysis

In line with the study's pre-registration (Bjurner, Zantvoort et al., 2024), the primary analysis was within-group comparisons, comparing the two blocks of each participant to each other. Where applicable, the between-group comparisons compared the data between therapists who were initially (first block) exposed to SHAP and those who were not (second block). Questions specifically asking to rate the SHAP condition (implicit comparisons) were tested regarding their deviation from neutral (4) scores.

Contingent on the respective parametric assumptions, dependent and independent t-tests were used to maximize power on the moderately sized sample. If the Shapiro-Wilks test suggested the t-test's normality assumption was violated, the non-parametric Wilcoxon test was used for the within-group comparisons. In that scenario, a Mann-Whitney was used for the in-between-group data. A Welch's test was used if the equal variance assumption was violated between groups. Figure 18 shows which hypothesis was tested with which analysis.

Figure 18: Overview test strategy



Notes: Different testing strategies and comparisons are applied depending on the measures and their timing. Where both within- and between-group comparisons were possible, within-group was the primary analysis.

4 Results

In this section, first, the participant numbers that resulted in the final cohort are described. Further, the results of the statistical analysis are reported, followed by insights into the remaining questionnaires' answers.

4.1 Final Cohort

During data collection from May to July 2024, 67 therapists completed the screening questionnaire. Six of them were excluded; four due to having treated less than five ICBT patients or one ICBT patient plus less than 20 patients with traditional CBT, and

two due to having treated the last ICBT-patient more than three years ago. Of the remaining 61 randomised therapists, 35 completed the study within the set timeframe, and their descriptive statistics can be found in Table 8.

Table 8: Cohort descriptives

Dimension	**Distribution**
Gender	71% Female, 29% male
Current Position	54% Licensed psychologist 17% Healthcare profession with basic psychotherapy training 14% Licensed psychotherapist 11% Intern-Psychologist 3% Last year student psychologist/ psychotherapist program
Years of experience psychological treatment	Median 6 (95%-CI 5.30-8.56)
Number of CBT patients treated	9% 1-4 6% 5-19 20% 20-99 11% 100-199 54% >200
Number of adult ICBT patients treated	3% 1-4 17% 5-19 37% 20-99 17% 100-199 26% >200

4.2 Hypotheses Testing

Thirteen participants saw non-SHAP cases first, and 22 saw SHAP cases first. There were no missing values. The normality assumption was violated for three within-group and six between-group comparisons, whereas the equal variance assumption was violated for one between-group comparison. The key result metrics of the respective tests are shown in Table 9. The superscript letter behind the effect size indicates the test used, and bold formatting in the first column marks the primary analysis. The detailed results, including the normality and equal variance test, are available in Supplementary Material 6.

Table 9: Analysis results with descriptive and p-values

a) Within-Group Comparison		SHAP		Non-SHAP		Comparative Results		
#[1]	Variable Name	mean	SD	mean	SD	p-value	Cohen's d [95%-CI]	RBC
H2.	Comprehensibility	4.97	1.52	4.86	1.57	0.841		0.05
H3.	Case-based trust*	4.86	1.01	4.51	0.7	0.018	0.43 [0.07, 0.76] [a]	
H7.	Numeric disagreement*	8.58	3.99	10.09	3.95	0.029	-0.39 [-0.73, -0.04]	
H8.	Verbal disagreement	2.56	0.65	2.67	0.39	0.194	-0.22 [-0.56, 0.11] [a]	
H9.	Usefulness*	4.82	1.09	4.34	1.02	0.003		0.66
H10.	a) Change treatment at all	0.96	0.56	1.07	0.51	0.397	-0.15 [-0.48, 0.19] [a]	
H10.	b) More treatment failures	1.81	1.06	1.93	0.99	0.701	-0.07 [-0.40, 0.27] [a]	
H10.	c) Less treatment success	3.79	0.89	3.83	0.8	0.828	-0.04 [-0.30, 0.37] [a]	

b) Between-Group Comparison		SHAP		Non-SHAP		Comparative Results		
#[1]	Variable Name	mean	SD	mean	SD	p-value	Cohen's d [95%-CI]	RBC
H4.	HCTS Total	2.94	0.51	2.61	0.5	0.074	-0.65 [-1.35, 0.07] [a]	
H4.	a) Risk	2.26	0.94	2.11	0.72	0.596	-0.19 [-0.87, 0.50] [a]	
H4.	b) Benevolent	3.28	1.16	3.23	1.04	0.886	-0.05 [-0.74, 0.64] [a]	
H4.	c) Competence	3.28	0.64	2.89	0.73	0.086		0.35 [d]
H4.	d) Reciprocity*	2.92	0.99	2.21	0.98	0.019		0.48 [d]
H1.	Decision transparency*	5.38	1.04	4.64	1.26	0.042		0.39 [d]
H2.	Comprehensibility	4.92	1.55	5.14	1.25	0.753		0.07 [d]
H3.	Case-based trust*	5.06	1.05	4.38	0.69	0.05	-0.77 [n/a] [b]	
H7.	Numeric disagreem.***	6.90	3.11	11.62	3.48	<0.001	1.41 [0.59, 2.20] [a]	
H8.	Verbal disagreement	2.50	0.57	2.71	0.44	0.257		0.23 [d]
H9.	Usefulness*	5.05	1.08	4.26	1.11	0.047	-0.72 [1.43, 0.00] [a]	
H10	a) Change treatment at all	1.10	0.49	1.11	0.50	0.945	0.02 [-0.66, -0.71] [a]	
H10	b) More treatment failures	1.96	0.75	1.93	1.12	0.651		0.09 [d]
H10	c) Less treatment success	0.08	0.83	0.18	0.75	0.704	-0.13 [-0.82, 0.55] [a]	

c) Implicit Comparison					
#[1]	Variable Name	mean	SD	p-value	RBC
H5.	Trust change***	5.43	1.14	<0.001	0.89 [c]
H6.	Comparative trust***	6.09	1.15	<0.001	0.99 [c]

*Notes: Effect sizes are shown as Cohen's d for [a] Student's [95%-CI] and [b] Welch's t-test (no CI), and rank biserial correlation (RBC) for [c] Wilcoxon and [d] Mann-Whitney U test. * = p < 0.05, ** = p < 0.01, *** = p < 0.001. [1]Respective primary test in bold.*

Understandability: While the perceived comprehensibility of the DST was not significantly altered, adding SHAP to the DST significantly increased the perceived decision transparency. This holds true for the within-group comparison ($p<0.001$, RBC=0.69) and less but also significantly for the between-group comparison ($p=0.04$, RBC=0.39).

Trust: Therapists significantly ($p=0.01$, $d=0.43$ [0.07-0.76]) trusted the individual case predictions with SHAP (4.86) more than without (4.51). This difference was even higher (5.38 to 4.64, $p=0.05$, $d=-0.77$) for the in-between-group comparison. For the HCTS, the difference between conditions was only significant for the reciprocity subscale ($p=0.02$, RBC=0.48). Regarding the implicit comparisons, SHAP significantly and strongly increased trust in predictions for both the within (mean 5.43, $p<0.001$ RBC=0.89), and the between comparisons (mean 6.09, $p<0.001$, RBC=0.99).

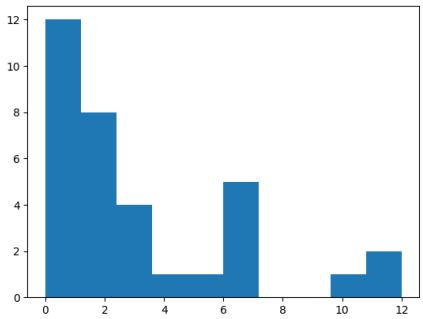
Considering agreeance as a more objective trust measure, therapists' own success probability prediction deviated significantly more from the DST for non-SHAP than SHAP cases ($p=0.03$, $d=-0.39$ [-0.73--0.04]). Again, this difference was even larger for the between-group comparison ($p<0.001$, $d=1.41$ [0.59, 2.20]).

Clinical Usefulness and Actionability: Therapists found the SHAP version of the DST significantly more useful for deciding how to proceed with the treatment in both the within ($p=0.003$, RBC=0.66) and between comparisons ($p=0.047$, $d=-0.72$ [-1.43, 0.00]). However, this did not translate into any significant changes in the intended treatment support level between the conditions.

4.3 Other Responses

In terms of baseline attitudes towards AI, the median responses on the 5-item Likert scale indicated a low perception of threat (median=2) and the probability that AI will lead to bad decisions (2), and a neutral position regarding the understanding of how AI works (3), what to expect from it (3), and its role in making healthcare better today (3). However, the median agreeance to being asked whether they perceived AI/ ML as something hopeful (4) and if it will make care better in the future (4) was higher.

Figure 19: Histogram quiz errors



Notes: Total of 24 questions

Regarding the quiz questions, most therapists had no or few mistakes, while five therapists answered five questions, and three answered up to half of the questions wrong (see Figure 19). The questions that were answered wrong most required recognising the patient's value on a histogram in relation to the mean (40% wrong) and one that required differentiating between per cent and per cent points increase (54% wrong).

5 Discussion

Personalizing mental health care through ML-based Decision Support Tools has long been discussed as a central lever to improve care efficiency and outcomes. However,

few real-life applications exist, and a key explanation for this is the conflict of the black-box nature of ML predictions with the expectations and responsibilities of healthcare providers (Yang, 2022). Therefore, the study at hand investigated if and how adding patient-specific explanatory information in the form of local SHAP values influences how ICBT therapists perceive the DST.

5.1 Evaluation of Hypotheses

Firstly, adding SHAP to the DST increased therapist's understanding of the decision rationale without making it less comprehensible. As such, we conclude that providing explanatory data to clinicians is a feasible way of narrowing the gap between AI and human cognition. Nevertheless, further research is necessary to determine if the time needed to investigate and leverage the additional data is available in routine care (Jacobs et al., 2021). One way of offering the additional information without overloading the therapists is to only show it on demand, e.g., by clicking on it (Jacobs et al., 2021; Xie et al., 2020).

Secondly, we found that therapists indicated higher trust in treatment outcome predictions when presented with SHAP. This effect was significant and moderate for single cases (H3) and large for the isolated implicit comparison (H5) and direct implicit comparison between SHAP and non-SHAP DST versions (H6). Further, this increase in trust also extended to a significant effect in numeric disagreement between therapists and DST predictions (H7). Hence, in line with previous findings (Jin et al., 2024), therapists were less likely to deviate in opinion from the DST prediction if they were provided with explanations. However, no such increase was reflected in the verbal disagreement-item or validated trust measure (HCTS). Regarding the verbal disagreement, therapists generally did not find it very likely that the DST misjudged or missed factors (Mean 2.56-2.67). This may be explained by the fact that the included factors 1) were chosen by humans and 2) were named in the introduction video, constituting global explanation measures. Regarding the HCTS, the statistical power was the lowest in the fairly unbalanced between-group comparison.

Thirdly, while explained predictions had a significant medium effect on being considered more useful in knowing how to proceed with treatment (H9), this knowledge did not translate to more intended changes in clinical support level (H10). Again, different explanations for this contradiction are plausible. On the one hand, increasing human support is just one of many ways to personalize treatment (Hornstein et al., 2023), and knowing how to proceed with treatment could refer to other factors such as content or communication. On the other hand, a paper interviewing psychiatrists described a disconnect between understanding and trusting a prediction score versus actually knowing what to do with it (Jacobs et al., 2021). As Diprose et al. (2020) also only measured the confidence to use the prediction, not its actual use, it remains unclear if and how this increased trust translates into a change of treatment. Additionally, trust is a multifaceted, complex concept that builds over time (Gille et al., 2020). Considering the therapists were only exposed to the DST for a very short time during the study, the effects of an increase in trust and confidence may yet develop.

5.2 Further Insights and Limitations

As the results from the baseline questionnaires (Q1) show, participants leaned towards considering ML a good thing to begin with, and very few participants mistrusted either version of the DST. While these findings align with related larger-scaled survey studies, positive attitudes towards AI are known to be high in Nordic countries such as Sweden (Masso et al., 2023; Scantamburlo et al., 2023). Further, a bias in participant selection through both the channels used and self-selection cannot be ruled out. Therefore, country- and context-specific differences must be considered, and further investigation is required to determine the generalisability of findings. However, the preliminary results' key limitation is the failure to meet the minimal sample size of 56 participants, particularly affecting the currently unbalanced between-group comparison. This likely also affects the 95%-confidence intervals of the effect sizes, which, for example, in the case of H3 or H7, are very large. As data gathering is still ongoing, it is yet to be seen if these preliminary results are confirmed in the final manuscript with a larger and more balanced cohort.

Notably, both versions of the DST led the therapists to propose average support for the likely successful patients (median=4/7). Exploring if this is due to risk avoidance or due to the factual belief the respective patient needed the proposed support level is out of the scope of this study. Either way, as likely failures were allocated higher support (median=6/7), this aspect must be considered when implementing a DST with the goal of optimising allocation instead of – as in this example - increasing overall support levels to above average.

Generally, SHAP values are only one of many options to add explainability to predictions (Pearson et al., 2019; Vilone & Longo, 2021a), all of which may yield different results. Additionally, SHAP values come with limitations, most notably that they are not causal (Lundberg et al., 2020) but are often mistakenly interpreted as such. So, while explainability methods can increase perceived comprehensibility, how sophisticated algorithms make decisions remains an epistemic question. This is especially important to consider and communicate if SHAP values are supposed to provide a decision base for treatment alterations. Further, data scientists commonly impute missing values as a pre-processing step, as it has been found to improve overall prediction accuracy (Bremer et al., 2020; Hentati Isacsson et al., 2023). However, in the constellation of the proposed DST, the SHAP method may place over-proportional importance on an imputed, and thus, possibly wrong, value. A last drawback of individual SHAP values is their additional constraint on the required knowledge and time to design, implement, maintain and use a DST. In that regard, while most therapists understood how to interpret the DST and SHAP cases well enough to make no or few mistakes in the quiz questions, three participants (9%) seemed to struggle with it and got half of the questions wrong, and another five (14%) answered a third of the questions wrong. It is unclear if they did not understand the concepts or if the mistakes were made due to carelessness or hurry, as no patients were at stake in this artificial setting. Either way, this finding highlights the importance of training and ensuring that all users of a DST have sufficient knowledge to interpret and use the DST information. At the same time, it also shows that short videos and informative text can suffice to introduce

these concepts for most clinicians, making this study a feasibility study for introducing SHAP values at scale.

Lastly, assuming that trust can iteratively build over time renders the question of whether explainability methods will be unnecessary once trust has been established. Even if this is the case, they may be an important tool in overcoming the adaption hurdle but would likely not be required for long within a given setting. This would change if they were used to inform how to proceed with treatment in an individual way, however, this circles back to the problem that they are not meant to be interpreted causally. Ultimately, the explanation, interpretation and understanding of predictions are just intermediate steps towards the actual goal, treatment improvement (Gille et al., 2020). Therefore, prospective randomised trials investigating the actual effect of DSTs with and without explanation on treatment outcomes are an inevitable and arguably the most important step in determining the future usefulness of XAI methods.

In summary, while SHAP values were found to increase the trust ICBT clinicians have in a DST, it remains to be seen if and how that translates into an improvement of care.

Ethical Approval

The study was conducted within an ethics application (2011/2091-31/3 with amendments 2016/21-32, 2017/2320-32, 2018/2550-32, 2019-04295 and 2021-01123) approved by the Swedish Ethical Review Authority. This application covers pilot trials of Decision Support Tools based on historical patient data from the Internet Psychiatry Clinic in Stockholm.

The study protocol for this randomised experiment was preregistered 29th of April, 2024-04-29 on the Open Science Framework (OSF) (Bjurner, Zantvoort, et al., 2024).

Chapter VII: Personalization Strategies

Hornstein, S., Zantvoort, K., Ulrike, L., Funk, B., & Kevin Hilbert. (2023). Personalization Strategies in Digital Mental Health Interventions: A Systematic Review and Conceptual Framework for Depressive Symptoms. Frontiers in Digital Health 5.

Introduction: Personalization is a much-discussed approach to improve adherence and outcomes for Digital Mental Health interventions (DMHIs). Yet, major questions remain open, such as (1) what personalization is, (2) how prevalent it is in practice, and (3) what benefits it truly has.

Methods: We address this gap by performing a systematic literature review identifying all empirical studies on DMHIs targeting depressive symptoms in adults from 2015 to September 2022. The search in Pubmed, SCOPUS and Psycinfo led to the inclusion of 138 articles, describing 94 distinct DMHIs provided to an overall sample of approximately 24,300 individuals.

Results: Our investigation results in the conceptualization of personalization as purposefully designed variation between individuals in an intervention's therapeutic elements or its structure. We propose to further differentiate personalization by what is personalized (i.e., intervention content, content order, level of guidance or communication) and the underlying mechanism [i.e., user choice, provider choice, decision rules, and machine-learning (ML) based approaches]. Applying this concept, we identified personalization in 66% of the interventions for depressive symptoms, with personalized intervention content (32% of interventions) and communication with the user (30%) being particularly popular. Personalization via decision rules (48%) and user choice (36%) were the most used mechanisms, while the utilisation of ML was rare (3%). Two-thirds of personalized interventions only tailored one dimension of the intervention.

Discussion: We conclude that future interventions could provide even more personalized experiences and especially benefit from using ML models. Finally, empirical evidence for personalization was scarce and inconclusive, making further evidence for the benefits of personalization highly needed.

1 Introduction

At an estimated lifetime prevalence of more than 10% (Hasin et al., 2018; Lim et al., 2018), major depressive disorder (MDD) is the second leading cause of years lived in disability (Ferrari et al., 2013). While this makes efficient treatments urgently needed, traditional approaches such as face-to-face psychotherapy are difficult to access for a significant part of patients (Moroz et al., 2020; Singer et al., 2022; Wang et al., 2002). However, providing treatment through digital channels such as mobile applications and online formats (Tal & Torous, 2017) is effective in reducing depressive symptoms (Königbauer et al., 2017; Moshe et al., 2021) in a cost-effective way (Donker et al., 2015). Since most of the world population has access to the internet(Martin, 2019) and/or a smartphone (O'Dea, 2021), digital mental health interventions (DMHIs) bypass barriers to traditional treatment.

Despite their potential, DMHIs inherit some of the general problems in depression treatment: Estimates for treatment dropout, as observed in RCTs, are up to 50% when considering publication bias (Torous et al., 2020). Moreover, response rates are unsatisfactory at less than 50% (Karyotaki et al., 2021). Therefore, improving outcomes and reducing dropouts in DMHIs are expected to be highly impactful in facing the burden of depression.

Luckily, DMHIs' unique delivery channel provides new opportunities to improve the treatment of those suffering from depressive symptoms. Specifically, digital applications can efficiently be individualised to improve users' experience and outcomes, as observable across many other domains, ranging from e-commerce (Kaptein & Parvinen, 2015) over e-learning (Zheng et al., 2022) towards social media (Shanahan et al., 2019). Simultaneously, the importance of accommodating patients' preferences for treatment outcomes in mental healthcare has been well established (Swift et al., 2018). Hence, the personalization of interventions to adapt treatment to individual needs is a promising approach to improving care, for depressive symptoms and beyond (Andrews & Williams, 2014; Aung et al., 2017; Chawla & Davis, 2013; D'Alfonso, 2020).

In line with that idea, a meta-analysis from 2013 showed that algorithm-based tailoring of DMHIs is associated with better outcomes (Lustria et al., 2013). A review from 2022 found that none of the 26 reviewed apps for depression used just-in-time (JIT) adaptations, a mechanism for personalizing the timing of content delivery based on the individual or the situation (Teepe et al., 2021). Another current systematic review investigated tailored interventions for workplace mental health (Moe-Byrne et al., 2022), finding benefits on several outcomes when content or feedback was tailored towards the individual. Finally, a component network analysis examined the benefits of common internet-based cognitive behavioural therapy (ICBT) packages for depression, discovering small interactions between treatment components and patient characteristics (Furukawa et al., 2021).

While these publications are unified in their call for more personalization in DMHIs, they do not add up to a satisfactory empirical and theoretical ground for it. Firstly, the

fragmented use of vocabulary fails to demarcate personalization from other distinct phenomena related to variability in DMHIs. For example, the term 'tailoring' is used across various scopes and foci (Lustria et al., 2013; Moe-Byrne et al., 2022; Ta Park et al., 2019), while similar mechanisms are elsewhere called 'individualised' (Zagorscak et al., 2020) or 'personalized' (Furukawa et al., 2021; Lau et al., 2020). This diversity in vocabulary is shared with non-digital settings, as for traditional psychotherapy, 15 different terms for the same phenomena of varying treatment between individuals were reported (Captari et al., 2018). Secondly, in contrast to the breadth of used vocabulary, the focus of mechanisms within studies seems to be relatively narrow, focusing on specific mechanisms (Lustria et al., 2013; Teepe et al., 2021) or areas (Moe-Byrne et al., 2022; Zagorscak et al., 2020) of personalization. This potentially leads to an underestimation of variability already in place. Finally, while two of the mentioned reviews investigated the benefits of personalization through direct comparisons, they did so without a specific focus on depression and with few studies being included. In conclusion, the concept, prevalence, and efficacy of personalization in DMHIs for depressive symptoms are not adequately delineated. Therefore, a disorder-specific review developing a conceptual framework for personalization and reviewing a wide span of interventions seems needed.

This systematic review aims to reduce the gap between the potential of personalization and its actual implementation by performing a comprehensive review of DMHIs for depressive symptoms with the following purposes:

I. Extract a conceptual framework that allows a clear and meaningful way of investigating, discussing, and classifying personalization.

II. Apply this framework to the available literature and report current use and mechanisms.

III. Evaluate the available evidence by identifying studies that directly compare interventions with different degrees of personalization.

2 Methods

This review was planned and reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). The protocol of this review was registered in the International Prospective Register of Systematic Reviews of the National Institute for Health Research (PROSPERO) under the ID CRD42022357408. The protocol was updated once after initial piloting to improve the alignment of the inclusion criteria and data extraction method with the scope of the review. Specifically, a new classification dimension for personalization was added that occurred in the literature and did not fit the pre-defined schema and the exclusion of e.g., prenatal depression was added to improve the comparability between included interventions. The final version of the protocol can be found in the Supplementary Material (Appendix 1).

2.1 Search Strategy

In the first step, a search was performed in three major databases (SCOPUS, PubMed, PsycInfo) to identify all published studies on DMHIs for depressive symptoms. The full search strings can be found in Appendix 2. Additionally, three related reviews (Himle et al., 2022; Karyotaki et al., 2021; Torous et al., 2020) were screened, and studies not yet included were added. Finally, papers brought to the author's awareness by being discussed in our included articles, not included yet but fulfilling our selection criteria, were added.

2.2 Selection Criteria

We included empirical studies on DMHIs specifically targeting depressive symptoms, determining the interventions target by authors' self-report. This covered both, patients with diagnosed major depressive disorder (MDD), as well as with subclinical levels of symptoms. To be considered a DMHI, interventions needed to be delivered through the internet and/or a smartphone. We included only empirical, peer-reviewed, English studies and conference articles with original data and patient cohort. To ensure a focus on the most relevant interventions for current use, we start our search from 2015 onwards.

To narrow down the focus of this review, studies on interventions targeting comorbid disorders such as anxiety were excluded. The same applied to those targeting a specific subtype of depression (e.g., prenatal depression), a single sub-symptom (e.g., rumination), or adolescent or elderly people (below 18 years or >64 years). Finally, those studies using digital technologies exclusively as a means of communication, such as one-on-one psychotherapy delivered via the web, were excluded as well.

2.3 Selection Procedure

One of the researchers (S.H.) performed an initial screening based on the title and abstract of the studies identified through the search strategy. A second researcher (K.Z.) conducted the same procedure for a randomly chosen subset of 100 studies, resulting in excellent interrater reliability (0.94). The full description of the intervention was then read by both reviewers for all remaining papers to determine the final selection, extract interventions and code the variables of interest. Disagreements on any aspect of this process were solved by discussion between the reviewers until a consensus was reached. If full texts were unavailable, they were requested from the corresponding author. This occurred 12 times, with 8 of the articles made available on request.

2.4 Development of the Conceptual Framework

During the initial screening and before the update of the PROSPERO registration, we developed the proposed framework in an iterative process, considering usability, conceptual literature, and the observed interventions. We departed from a common dictionary definition defining personalization as "*the action of designing or producing something that meets someone's individual requirement*" (Surprenant & Solomon,

1987). Based on that, we intended to classify personalization in DMHIs in a broad enough way to cover the diversity of mechanisms present in related reviews and studies. At the same time, we intended to narrow down the concept to those mechanisms affecting the therapeutic content and structure, setting it apart from superficial sources of variability. Therefore, we excluded interactivity (Deighton & Sorrell, 1996), the sole replay of user input as part of the app experience. For example, showing each patient their previously set goal might be a powerful tool, but it does not change the underlying therapeutic elements delivered. Additionally, we factored out customization (Sundar & Marathe, 2010), minor aesthetic adaptation such as users ability to change the colour of an avatar. Finally, seeing personalization as referring to the level of the individual patient, we excluded group-based variability, such as cultural adaptation of the entire intervention (Spanhel et al., 2021).

Numerous screened interventions used a structured session-based approach to deliver their intervention - a common approach among manualized mental health interventions (Luborsky, 1984). Therefore, we identified a) content (what is delivered during a session) and b) order (how sessions are ordered) as potential areas of personalization. Since c) guidance (level of human contact) is a highly relevant and variable aspect of DMHIs (Karyotaki et al., 2021) we added it as another dimension. Finally, as we discovered prompts and mechanisms targeting the timing of interventions not being sufficiently represented in these three categories, we appended d) communication as another dimension.

While, as mentioned above, we intended to exclude customization as minor user-choice-based adaptations of the intervention, we did not exclude user choice per se from our concept. This differs from the use in fields like marketing, where anything done by the user is defined as customization, not personalization (Sundar & Marathe, 2010). However, we saw the inclusion of actively designed user choice being justified for the following reasons: Firstly, user choice was a common mechanism described in the included interventions. Secondly, those mechanisms seem easily implementable and therefore highly relevant for practitioners interested in personalizing their intervention. Finally, user agency has been shown to be particularly relevant in mental healthcare (Swift et al., 2018). We also identified provider choice as another mechanism for guided and blended interventions. For data-driven personalization mechanisms, we saw rule-based and ML as distinct mechanisms applying static or learning criteria for personalization.

2.5 Data Extraction

The framework developed above was applied to all identified interventions, coding the presence of personalization for each of the four (a-d) dimensions and classifying the underlying mechanism. For this, interventions had to be extracted from the included studies, and information from several studies on the same intervention had to be merged. If more than one distinct intervention was presented in a study, they were coded separately. Intervention versions in different languages were not coded separately if not reported to be clearly distinct in their content. If more than one study was

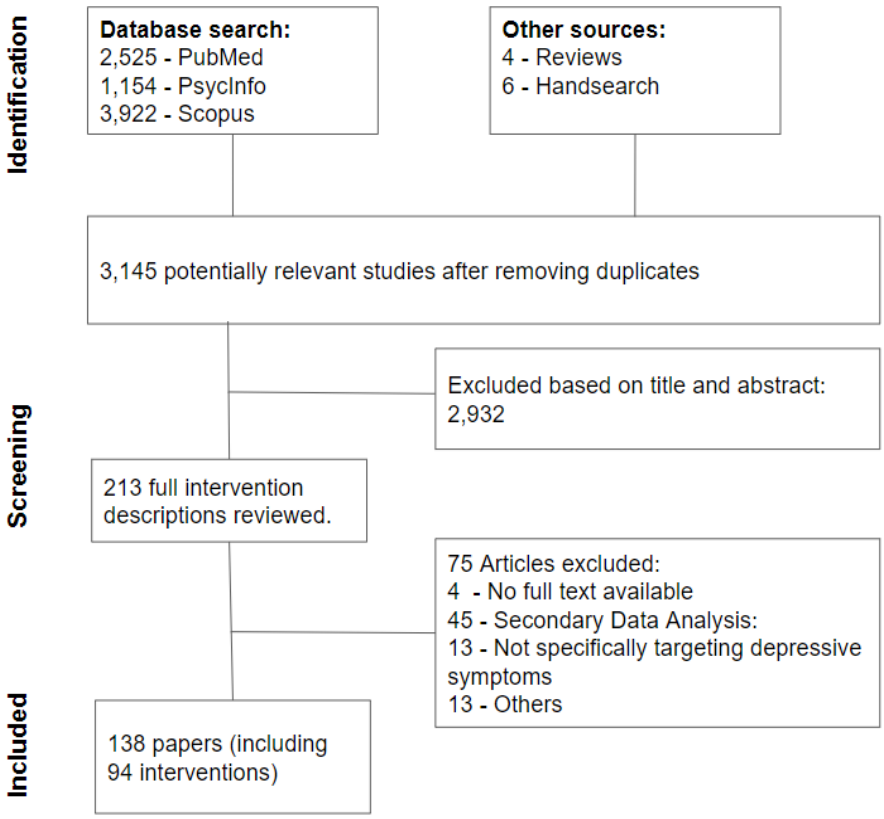
available, a single observation of personalization resulted in a positive coding, but conflicting information was noted. Additionally, cited material such as older papers, weblinks, or appendices were consulted in the refrained from additional free-hand research on the reported interventions. In case information was indicative of personalization but insufficient for our coding, we contacted the corresponding author and asked for clarification. For this, we provided a four-week response window, including one reminder. Out of the seven authors contacted, six responded by providing additional information. In the single case where authors did not respond (Burton et al., 2016) we decided to code restrictively and assume the simpler of the potential mechanisms involved (rule-based instead of ML). Finally, for evaluating the evidence for personalization, we included every study that directly compared intervention versions that differed in their degree of personalization, according to our framework. We extracted effect sizes, dependent variables, and sample sizes for those.

3 Results

3.1 Study Selection

Overall, we identified 3.143 potentially relevant publications and screened the title and abstract. For 213 of those, the full intervention description was reviewed, resulting in the final inclusion of N=138 papers describing k=94 distinct DMHIs for depressive symptoms (see Figure 20).

Figure 20: PRISMA flowchart of study inclusion



3.2 Intervention and Study Characteristics

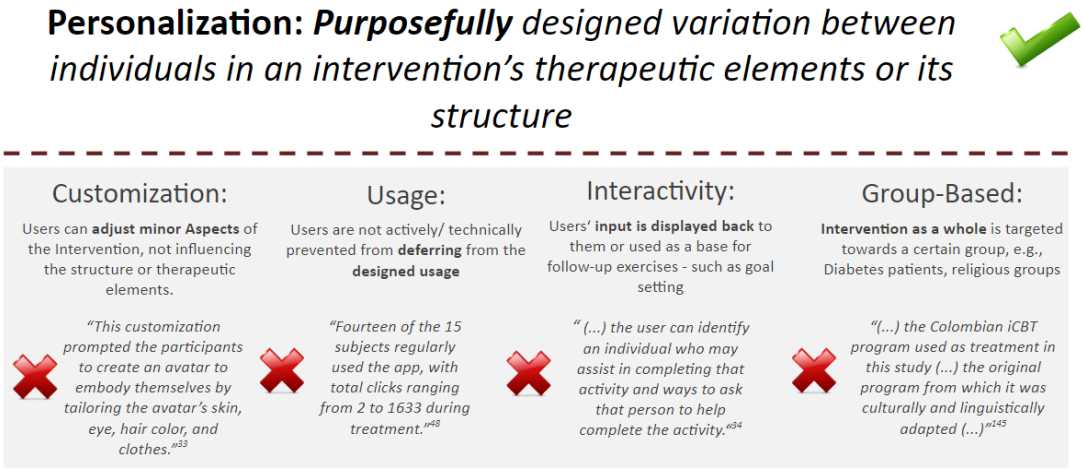
While mostly one study per intervention was included, for some up to seven publications on distinct trials were present and kept for further analysis. Across all studies, the reviewed interventions were deployed to approximately 24.300 participants, with an average sample size of 259 participants per intervention (range 1 to 2964). 75 of the interventions were used in a randomised controlled trial, with the remaining evidence coming from feasibility studies, naturalistic routine care data, and other study designs.

Most interventions had a duration between 6 and 12 weeks, and around 40 of the interventions report a structured module-/ session-based design, delivering the content in pre-defined blocks. Finally, 38 interventions were unguided (no human contact within the intervention), 32 guided (including guidance from clinician or coach), and 14 blended (combining face-to-face and digital treatment), with the remaining 10 covering more than one of those categories. An overview of all characteristics can be found in Appendix 3.

3.3 Conceptual Framework

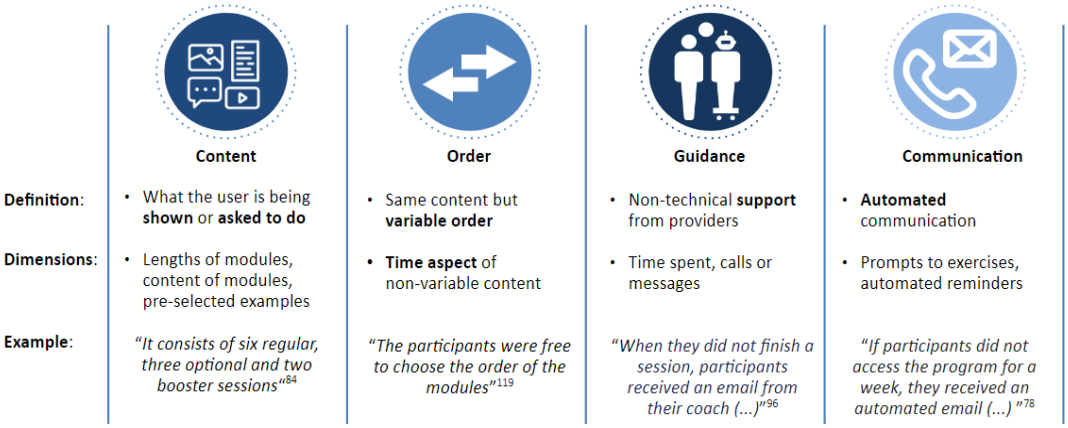
A conceptual framework of personalization in DMHIs was synthesised from the reviewed DMHIs and theoretical considerations. In summary, an understanding of personalization as *purposefully designed variation between individuals in an intervention's therapeutic elements or its structure* emerged. As such, personalization is differentiated from customization, usage, interactivity, and group-based adaptations. Customization describes minor adjustments, such as visual aspects, leaving the actual therapeutic ingredients unchanged. Usage refers to possible user-induced differences in app usage that were not actively or purposefully designed. For example, variability in the time spent on a module is usage, the offering of short and long versions of a module qualifies as personalization. Interactivity, the mere replay of user input as for example commonly used for goal-setting exercises, as this leaves the actual therapeutic elements and structure unchanged. Finally, as we understand personalization as referring to the level of the individual, we see it as being distinct from group-based variability, such as the adaptation for a particular cultural context (see Figure 21).

Figure 21: Delineation of personalization from other forms of adaptation.



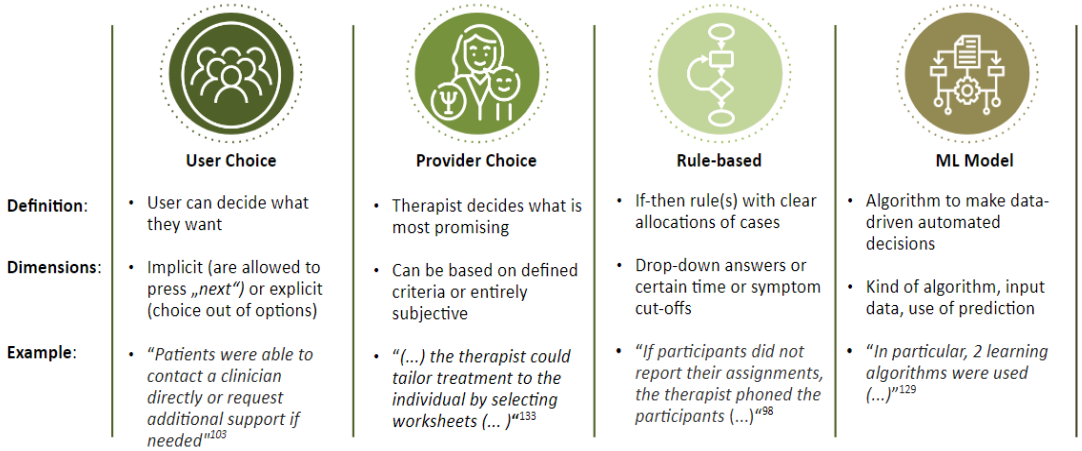
Within our definition of personalization, four personalizable intervention dimensions emerged, namely content, guidance level, order, and communication, as summarised in Figure 22. Content describes all variability in the delivered intervention material, such as exercises, psycho-educative material or topics presented. Order includes cases when patients receive the same content but in different order. Guidance refers to the extent of therapeutic support offered. Communication concerns the channel, timing, and content of actively offered information outside of the intervention's content. This primarily includes prompts or reminder messages. Mechanisms regarding the frequency and timing of the intervention, such as JIT mechanisms, also fall under communication.

Figure 22: Dimensions of personalization



Further, four different mechanisms beyond personalization emerged: user choice, provider choice, rule-based and ML-based personalization (see Figure 23). User choice covers intentionally designed personalization based on the direct choice of the participant. For provider choice, either the individual providing guidance, or the clinician involved in a blended setting makes the personalization decision. Among automated personalization mechanisms, rule-based (if-then-decision rules) from Machine Learning (decisions with "*learned*" decision criteria) personalization mechanisms are gathered.

Figure 23: Mechanisms of personalization



3.4 Results on Personalization

Applying the proposed framework for classifying variability in DMHIs, personalization was reported for 62 of the 94 interventions (66%). Most prominently, personalization mechanisms were used in the content for 30 of the interventions (32%). This was followed by personalized communication (30%), type (25%), and order (4%). 43 of the 62 (69%) interventions with a reported personalization mechanism did so for only a single dimension, while one DMHI reported a mechanism for all four subdomains of their intervention (Hatcher et al., 2018; Heim et al., 2021; Piera-Jiménez et al., 2021; Roepke et al., 2015; Rosso et al., 2017; Smith et al., 2017; Titov et al., 2015).

Across the 107 reported personalization mechanisms, rule-based was most prominent, being used in 51 cases (48%). User choice was observed in 39 cases (36%), and providers were involved in personalization 14 times (13%). The use of Machine Learning was reported three times (3%). Rule-based personalization was particularly prominent in the communication domain, accounting for 21 occurrences. Similarly, human guidance was personalized using decision rules 16 times. For content, user choice had a more prominent role, being reported 15 times. The use of personalization is summarised in Figure 24, with examples of the 3 most strategies being presented in Table 10. The share of interventions applying at least one personalization mechanism was the highest for guided interventions (72%), followed by unguided (63%) and tailed by blended (57%) interventions. Generally, the dimensions of personalization were equally spread across guidance levels. However, provider choice was nearly twice as common for blended than for guided interventions.

Figure 24: Number of studies per personalization type and mechanism

	User Choice	Provider Choice	Rule-based	ML-Model	N per Dimension
Content	15	7	13	3	38 (36%)
Order	3	1	1	0	5 (5%)
Guidance	10	6	16	0	32 (30%)
Communication	11	0	21	0	32 (30%)
N per Mechanism	39 (36%)	14 (13%)	51 (48%)	3 (3%)	107

Table 10: Personalization mechanisms per dimension in the intervention

Count	Dimension	Mechanism	Description
21	communication	rule-based	e.g., reminder for inactivity/ non-completion
16	guidance	rule-based	e.g., increased guidance/ clinician contact for symptom changes
15	content	user choice	e.g., optional content selectable for patient

3.5 Use of Automated Decisions for Personalization

Among the 55 automated mechanisms used, most were rule-based mechanisms of personalization. Here, activity data was heavily utilised, for example, for reminders in case of inactivity. Another common pattern was the use of symptom scores like the PHQ to step up care in the form of additional guidance (Callan et al., 2021) or the change from guided to blended care (Schueller & Mohr, 2015). While those approaches mostly used overall symptom severity, one exemption was the personalization based on suicide risk as e.g., in the form of additional prompts (Zagorscak et al., 2018).

We identified three clear use cases of ML techniques for personalization. Firstly, EmoRecorder (Hung et al., 2015) used an activity recommendation system based on diverse data sources like app activity, sensor data and past recommendations. However, the intervention was at an early stage, being tested on a sample of only 15 healthy individuals. Secondly, MOSS (Wahle et al., 2016) built on a JIT framework to assign intervention content depending on users' context and preferences. As such, it tested a recommender system with a sample of 126 adults. A third recommender system approach, MUBS (Rohani et al., 2019), applied a combination of ML and user choice by providing the 17 patients with a set of content recommendations.

3.6 Empirical Comparison of More and Less Personalized Interventions

Among the 138 papers in the final review, we identified two papers that included a direct comparison of a more and a less personalized version of an intervention. One study had participants fill out a questionnaire on motivational schemata and either matched them with an intervention arm to fit their motivational preference or a general one (Bücker et al., 2022). Results showed effects for one of the two included motives ('being supported') on anticipated adherence, working alliance, and satisfaction; however, the overall sample size of this trial was just 55 participants. Secondly, a study compared a program version including JIT prompts with one without those prompts, therefore, differing the personalization in the communication domain between trial arms (Everitt et al., 2021). While both versions showed significant effects compared to the waitlist, no effects were reported between the arms. Again, this should be interpreted with caution, considering the sample size of around 60 individuals per group.

4 Discussion

In recent years, personalization has been widely discussed as a promising avenue to improve DMHI adherence and outcomes. Nevertheless, it remains unclear what it entails and how it is used. In this review, we address this need for the case of depressive symptoms, by defining personalization as purposefully designed variation in intervention content, order, guidance, or communication. As possible mechanisms to operationalise personalization, we extract user choice, provider choice, decision rules, and ML. Applying this framework to 94 interventions for depressive symptoms reveals that two-thirds use at least one technique for personalization. Especially rule-based personalization of communication and guidance and user choice-based personalization of content is common. However, among interventions applying personalization, a majority does so just for one out of four dimensions of the intervention. Also, the use of ML models is scarce and limited to feasibility studies. Additionally, just two of the included studies investigated the benefits of personalization, both having small samples and just one finding supporting evidence.

Arguably, the biggest contrast between the proposed potentials in the personalization of DMHIs (Andrews & Williams, 2014; Aung et al., 2017; Chawla & Davis, 2013; D'Alfonso, 2020) and the existing literature is the lack of implemented ML mechanisms. Several of the implemented non-learning algorithms and decision rules were well designed. Yet, literature on ML in DMHIs reveals ample further promising and feasible use cases. Firstly, a notable body of research provides encouraging results in outcome (Hornstein et al., 2021; Vieira et al., 2022) and dropout (Bennemann et al., 2022; Bremer et al., 2020) predictions in DMHIs. Adapting the interventions for assumed non-responders is a low-hanging fruit and has already been successful for other disorders (Forsell et al., 2019). Secondly, a prominent algorithmic approach to personalization in digital products is recommender systems (Alamdari et al., 2020; Kulkarni et al., 2020; Z. Liu et al., 2022). While all included ML approaches were such recommender systems, they were in early stages and deployed to very small sample sizes. Finally, all included ML approaches focused on the content of the intervention. However, ML also is a promising approach to personalize guidance, communication and order.

Contrasting theory and observations in another dimension, the data used for personalization just samples a fraction of the technically possible. While app usage patterns are an obvious data option, smartphones can also measure sleep patterns (Ong & Gillespie, 2016), physical activity (Bort-Roig et al., 2014), social interactions (Boukhechba et al., 2018), and many other data points known to be relevant for depressive symptoms. Readily available toolkits like Apple's health kit (North & Chaudhry, 2016) reduce the effort for implementation significantly. However, particularly passive sensing was rarely utilised in the reviewed interventions. Notably, the potential of ML-based personalization is heavily intertwined with the quality of the data available to them. Beyond that, aspects such as ethical responsibility in health care and privacy rights must be strongly considered, especially when investigating automated decisions (Carr, 2020).

Several interventions used self-reported symptoms for the personalization of the intervention. Noticeably, these mechanisms mostly used overall symptom severity. This approach disregards that symptom profiles can vary massively between patients with the same overall score (Fried & Nesse, 2015). Some evidence points toward distinct symptom patterns being associated with different optimal treatment procedures (Boschloo et al., 2019). Therefore, while overall severity seems reasonable for varying guidance or communication, the sub-symptoms might be a promising ground for personalizing content and order.

The two included trials that manipulated personalization did so with small sample sizes and inconclusive results. Subsequently, one barrier to implementing personalization might be the lack of clear evidence for its benefits. However, RCTs investigating personalization are likely costly and require large sample sizes when assuming smaller effect sizes than for waitlist-controlled studies. Luckily, meta-analytic approaches allow summarising evidence across studies, even when personalization is rarely directly manipulated. While we mentioned one such approach investigating interactions between individuals and benefits of ICBT packages (Furukawa et al., 2021), we consider similar approaches for other personalization mechanisms as very promising. However, the identification and comparison of relevant studies in meta-analyses requires shared vocabulary and a common framework. We believe that such future work will benefit from the shared conceptual framework proposed in this article.

There are some limitations of this review that should be considered. Firstly, published studies are just one marker of what interventions are in use. While several included interventions originated in a commercial setting, those from academic settings will likely still be overrepresented in this review. Secondly, we focused on personalization within an intervention, excluding the personalization of interventions themselves. For example, past approaches investigated the data-driven personalization of therapy school (Delgadillo & Gonzalez Salas Duhne, 2020) or the decision between medication and CBT (Gunlicks-Stoessel et al., 2020). Thirdly, identifying interventions for depressive symptoms while excluding those addressing comorbid disorders, particularly anxiety, has proven challenging. One example is when anxiety was mentioned as intervention target in a cited study, but not in the original paper. While this seems understandable in light of the well-established comorbidity of depression and anxiety (Lamers et al., 2011), this resulted in several edge cases of inclusion. Fourthly, we took interactivity, customization, and group-based adaptions out of the scope of this review due to their difference in nature to personalization. This should not be misunderstood as an assumed inferiority, and we call for the further investigation of these approaches to complement or even substitute personalization. Finally, as we developed our framework exclusively with studies on depressive symptoms, it remains unclear whether there are more aspects to consider with other disorders. However, we expect this framework to provide value beyond the use case of depressive symptoms and encourage future studies to investigate personalization strategies in other domains.

In conclusion, our conceptual development and empirical evaluation holistically characterizes the current use of personalization for DMHIs for depressive symptoms. A broad conceptualization of personalization reveals that most interventions incorporate

personalization mechanisms. However, we conclude that we are barely scratching the surface of what is technically possible and already gold standard in other research and business areas. At the same time, we see the thin empirical ground as a barrier to implementation and call for more direct and meta-analytic evidence to delineate the benefits personalization has over an ‘one size fits all’-approach. Finally, as we see this question as equally pressing for other disorders, we hope for similar-minded approaches for those in the future.

Bibliography

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, *31*(1), 92–116. https://doi.org/10.1080/10503307.2020.1808729

Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry: A systematic literature review: Text Mining Applications in Psychiatry. *International Journal of Methods in Psychiatric Research*, *25*(2), 86–100. https://doi.org/10.1002/mpr.1481

Ahmed, H., & Lofstead, G. (2021). *Quantifying the Crisis in Reproducible Machine Learning.* (SAND2021-10258C). Sandia National Lab. (SNL-NM), Albuquerque, NM (United States). https://doi.org/10.2172/1883039

Aimone, J. A., Ball, S., & King-Casas, B. (2016). It's not what you see but how you see it: Using eye-tracking to study the risky decision-making process. *Journal of Neuroscience, Psychology, and Economics*, *9*(3–4), 137–144. https://doi.org/10.1037/npe0000061

Alamdari, P. M., Navimipour, N. J., Hosseinzadeh, M., Safaei, A. A., & Darwesh, A. (2020). A Systematic Study on the Recommender Systems in the E-Commerce. *IEEE Access*, *8*, 115694–115716. https://doi.org/10.1109/ACCESS.2020.3002803

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., Duan, Y., Abdullah, A., Farhan, L., Lu, Y., Gupta, A., Albu, F., Abbosh, A., & Gu, Y. (2023). A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, *10*(1), 46. https://doi.org/10.1186/s40537-023-00727-2

American Psychiatric Association. (2006). Treatment of patients with eating disorders,third edition. American Psychiatric Association. *The American Journal of Psychiatry*, *163*(7 Suppl), 4–54.

Amir, S., Coppersmith, G., Carvalho, P., Silva, M. J., & Wallace, B. C. (2017). Quantifying Mental Health from Social Media with Neural User Embeddings. *Proceedings of the 2nd Machine Learning for Healthcare Conference*, 306–321. https://proceedings.mlr.press/v68/amir17a.html

Andaur Navarro, C. L., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ*, n2281. https://doi.org/10.1136/bmj.n2281

Andersson, G., Carlbring, P., & Rozental, A. (2019). Response and Remission Rates in Internet-Based Cognitive Behavior Therapy: An Individual Patient Data Meta-Analysis. *Frontiers in Psychiatry*, *10*, 749. https://doi.org/10.3389/fpsyt.2019.00749

Andrews, G., Basu, A., Cuijpers, P., Craske, M. G., McEvoy, P., English, C. L., & Newby, J. M. (2018). Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: An updated meta-analysis. *Journal of Anxiety Disorders*, *55*, 70–78. https://doi.org/10.1016/j.janxdis.2018.01.001

Andrews, G., & Williams, A. D. (2014). INTERNET PSYCHOTHERAPY AND THE FUTURE OF PERSONALIZED TREATMENT: Commentary: Internet Psychotherapy. *Depression and Anxiety*, *31*(11), 912–915. https://doi.org/10.1002/da.22302

Antal, M., Fejér, N., & Buza, K. (2021). SapiMouse: Mouse Dynamics-based User Authentication Using Deep Feature Learning. *2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, 61–66.

Arapakis, I., & Leiva, L. A. (2016). Predicting User Engagement with Direct Displays Using Mouse Cursor Information. *Proceedings of the 39th Int. ACM SIGIR Conf. on Research and Development in Inform. Retrieval*, 599–608.

Arapakis, I., & Leiva, L. A. (2020). Learning Efficient Representations of Mouse Movements to Predict User Attention. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1309–1318. https://doi.org/10.1145/3397271.3401031

Arias, D., Saxena, S., & Verguet, S. (2022). Quantifying the global burden of mental disorders and their economic value. *eClinicalMedicine*, *54*. https://doi.org/10.1016/j.eclinm.2022.101675

Atla, A., Tada, R., Sheng, V., & Singireddy, N. (2011). Sensitivity of different machine learning algorithms to noise. *Journal of Computing Sciences in Colleges*, *26*, 96–103.

Aung, M. H., Matthews, M., & Choudhury, T. (2017). Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies: Aung et al. *Depression and Anxiety*, *34*(7), 603–609. https://doi.org/10.1002/da.22646

Aziz, M., Erbad, A., Belhaouari, S. B., Almourad, M. B., Altuwairiqi, M., & Ali, R. (2023). Who uses mHealth apps? Identifying user archetypes of mHealth apps. *DIGITAL HEALTH*, *9*. https://doi.org/10.1177/20552076231152175

Baker, S. L., Heinrichs, N., Kim, H.-J., & Hofmann, S. G. (2002). The liebowitz social anxiety scale as a self-report instrument: A preliminary psychometric analysis. *Behaviour Research and Therapy*, *40*(6), 701–715. https://doi.org/10.1016/s0005-7967(01)00060-2

Balki, I., Amirabadi, A., Levman, J., Martel, A. L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S. C., Kong, D., Moody, A. R., & Tyrrell, P. N. (2019). Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. *Canadian Association of Radiologists Journal*, *70*(4), 344–353. https://doi.org/10.1016/j.carj.2019.06.002

Barrett, M. S., Chua, W.-J., Crits-Christoph, P., Gibbons, M. B., & Thompson, D. (2008). Early withdrawal from mental health treatment: Implications for psychotherapy practice. *Psychotherapy: Theory, Research, Practice, Training*, *45*(2), 247–267. https://doi.org/10.1037/0033-3204.45.2.247

Bates, S., Hastie, T., & Tibshirani, R. (2022). *Cross-validation: What does it estimate and how well does it do it?* (arXiv:2104.00673). arXiv. http://arxiv.org/abs/2104.00673

Baumeister, H., Reichler, L., Munzinger, M., & Lin, J. (2014). The impact of guidance on Internet-based mental health interventions—A systematic review. *Internet Interventions*, *1*(4), 205–215. https://doi.org/10.1016/j.invent.2014.08.003

Beard, C., Millner, A. J., Forgeard, M. J. C., Fried, E. I., Hsu, K. J., Treadway, M. T., Leonard, C. V., Kertz, S. J., & Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine*, *46*(16), 3359–3369. https://doi.org/10.1017/S0033291716002300

Becker, D., van Breda, W., Funk, B., Hoogendoorn, M., Ruwaard, J., & Riper, H. (2018). Predictive modeling in e-mental health: A common language framework. *Internet Interventions*, *12*, 57–67. https://doi.org/10.1016/j.invent.2018.03.002

Beintner, I., Emmerich, O. L. M., Vollert, B., Taylor, C. B., & Jacobi, C. (2019). Promoting positive body image and intuitive eating in women with overweight and obesity via an online intervention: Results from a pilot feasibility study. *Eating Behaviors*, *34*, 101307. https://doi.org/10.1016/j.eatbeh.2019.101307

Beintner, I., Vollert, B., Zarski, A.-C., Bolinski, F., Musiat, P., & Jacobi, C. (2019). Adherence Reporting in Randomized Controlled Trials Examining Manualized Multisession Online Interventions: Systematic Review of Practices and Proposal for Reporting Standards. *Journal of Medical Internet Research*.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta*, *760*, 25–33. https://doi.org/10.1016/j.aca.2012.11.007

Ben-Israel, D., Jacobs, W. B., Casha, S., Lang, S., Ryu, W. H. A., de Lotbiniere-Bassett, M., & Cadotte, D. W. (2020). The impact of machine learning on patient care: A systematic review. *Artificial Intelligence in Medicine*, *103*, 101785. https://doi.org/10.1016/j.artmed.2019.101785

Bennemann, B., Schwartz, B., Giesemann, J., & Lutz, W. (2022). Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *The British Journal of Psychiatry*, *220*(4), 192–201. https://doi.org/10.1192/bjp.2022.17

Berking, M., & Whitley, B. (2014). *Affect Regulation Training: A Practitioners' Manual*. Springer. https://doi.org/10.1007/978-1-4939-1022-9

Berner, E. S., & La Lande, T. J. (2016). Overview of Clinical Decision Support Systems. In E. S. Berner (Ed.), *Clinical Decision Support Systems: Theory and Practice* (pp. 1–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-31913-1_1

Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, *105*(9), 17–24.

Bjurner, P., Isacsson, N. H., Abdesslem, F. B., Boman, M., Forsell, E., & Kaldo, V. (2024). *Study protocol for a triple-blind randomised controlled trial evaluating a machine learning-based predictive clinical decision support tool for internet-delivered cognitive behaviour therapy (ICBT) for depression and anxiety*. https://osf.io/cs4bx/

Bjurner, P., Zantvoort, K., Forsell, E., Wallert, J., Funk, B., & Kaldo, V. (2024). Effects on therapists' trust and perception of a decision support tool when adding decision transparency to explain machine learningpredictionsof internet-based cognitive behavioral therapy–Protocol of a randomized experiment. *OSF Preregistration*. https://osf.io/https://osf.io/zn5b3

Blei, D. M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1986). *Occam's Razor*. *24*(6). https://www.sciencedirect.com/science/article/abs/pii/0020019087901141

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bone, D., Lee, C.-C., Chaspari, T., Gibson, J., & Narayanan, S. (2017). Signal Processing and Machine Learning for Mental Health Research and Clinical Applications [Perspectives]. *IEEE Signal Processing Magazine*, *34*(5), 196–195. https://doi.org/10.1109/MSP.2017.2718581

Bort-Roig, J., Gilson, N. D., Puig-Ribera, A., Contreras, R. S., & Trost, S. G. (2014). Measuring and Influencing Physical Activity with Smartphone Technology: A Systematic Review. *Sports Medicine*, *44*(5), 671–686. https://doi.org/10.1007/s40279-014-0142-5

Boschloo, L., Bekhuis, E., Weitz, E. S., Reijnders, M., DeRubeis, R. J., Dimidjian, S., Dunner, D. L., Dunlop, B. W., Hegerl, U., Hollon, S. D., Jarrett, R. B., Kennedy, S. H., Miranda, J., Mohr, D. C., Simons, A. D., Parker, G., Petrak, F., Herpertz, S., Quilty, L. C., … Cuijpers, P. (2019). The symptom-specific efficacy of antidepressant medication vs. cognitive behavioral therapy in the treatment of depression: Results from an individual patient data meta-analysis. *World Psychiatry*, *18*(2), 183–191. https://doi.org/10.1002/wps.20630

Boukhechba, M., Daros, A. R., Fua, K., Chow, P. I., Teachman, B. A., & Barnes, L. E. (2018). DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones. *Smart Health*, *9–10*, 192–203. https://doi.org/10.1016/j.smhl.2018.07.005

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Breakwell, G. M., Barnett, J., & Wright, D. B. (2020). *Research Methods in Psychology* (5th ed.). Sage Publications. https://www.torrossa.com/en/resources/an/5019108

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *199–231*.

Bremer, V., Becker, D., Funk, B., & Lehr, D. (2017). PREDICTING THE INDIVIDUAL MOOD LEVEL BASED ON DIARY DATA. *Proceedings of the 25th European Conference on Information Systems*, 1161–1177.

Bremer, V., Chow, P. I., Funk, B., Thorndike, F. P., & Ritterband, L. M. (2020). Developing a Process for the Analysis of User Journeys and the Prediction of Dropout in Digital Health Interventions: Machine Learning Approach. *Journal of Medical Internet Research*, *22*(10). https://doi.org/10.2196/17738

Bricker, J., Miao, Z., Mull, K., Santiago-Torres, M., & Vock, D. M. (2023). Can a Single Variable Predict Early Dropout From Digital Health Interventions? Comparison of Predictive Models From Two Large Randomized Trials. *Journal of Medical Internet Research*, *25*, e43629. https://doi.org/10.2196/43629

Bücker, L., Berger, T., Bruhns, A., & Westermann, S. (2022). Motive-Oriented, Personalized, Internet-Based Interventions for Depression: Nonclinical Experimental Study. *JMIR Formative Research*, *6*(9), e37287. https://doi.org/10.2196/37287

Bucur, A.-M., Cosma, A., & Dinu, L. P. (2021). *Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT* (arXiv:2106.16175). arXiv. http://arxiv.org/abs/2106.16175

Burton, C., Szentagotai Tatar, A., McKinstry, B., Matheson, C., Matu, S., Moldovan, R., Macnab, M., Farrow, E., David, D., Pagliari, C., Serrano Blanco, A., Wolters, M., & for the Help4Mood Consortium. (2016). Pilot randomised controlled trial of Help4Mood, an embodied virtual

agent-based system to support treatment of depression. *Journal of Telemedicine and Telecare*, *22*(6), 348–355. https://doi.org/10.1177/1357633X15609793

Bush, K., Kivlahan, D. R., McDonell, M. B., Fihn, S. D., Bradley, K. A., & for the Ambulatory Care Quality Improvement Project (ACQUIP). (1998). The AUDIT Alcohol Consumption Questions (AUDIT-C): An Effective Brief Screening Test for Problem Drinking. *Archives of Internal Medicine*, *158*(16), 1789–1795. https://doi.org/10.1001/archinte.158.16.1789

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*, 01643933. https://doi.org/10.1016/j.bpsc.2017.11.007

Cabitza, F., & Campagner, A. (2021). The need to separate the wheat from the chaff in medical informatics. *International Journal of Medical Informatics*, *153*. https://doi.org/10.1016/j.ijmedinf.2021.104510

Callan, J. A., Dunbar Jacob, J., Siegle, G. J., Dey, A., Thase, M. E., DeVito Dabbs, A., Kazantzis, N., Rotondi, A., Tamres, L., Van Slyke, A., & Sereika, S. (2021). CBT MobileWork©: User-Centered Development and Testing of a Mobile Mental Health Application for Depression. *Cognitive Therapy and Research*, *45*(2), 287–302. https://doi.org/10.1007/s10608-020-10159-4

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, *23*(5), 649–685. https://doi.org/10.1017/S1351324916000383

Camacho-Collados, J., & Pilehvar, M. T. (2018). On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. In T. Linzen, G. Chrupała, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 40–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5406

Captari, L. E., Hook, J. N., Hoyt, W., Davis, D. E., McElroy-Heltzel, S. E., & Worthington, E. L. (2018). Integrating clients' religion and spirituality within psychotherapy: A comprehensive meta-analysis. *Journal of Clinical Psychology*, *74*(11), 1938–1951. https://doi.org/10.1002/jclp.22681

Carey, K. B., Neal, D. J., & Collins, S. E. (2004). A psychometric analysis of the self-regulation questionnaire. *Addictive Behaviors*, *29*(2), 253–260. https://doi.org/10.1016/j.addbeh.2003.08.001

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: An updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*, *47*(1), 1–18. https://doi.org/10.1080/16506073.2017.1401115

Carr, S. (2020). 'AI gone mental': Engagement and ethics in data-driven technology for mental health. *Journal of Mental Health*, *29*(2), 125–130. https://doi.org/10.1080/09638237.2020.1714011

Cawley, G. C., & Talbot, N. L. C. (2017). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Cepeda, C., Dias, M. C., Rindlisbacher, D., Gamboa, H., & Cheetham, M. (2021). Knowledge extraction from pointer movements and its application to detect uncertainty. *Heliyon*, *7*(1), e05873. https://doi.org/10.1016/j.heliyon.2020.e05873

Cepoiu, M., McCusker, J., Cole, M. G., Sewitch, M., Belzile, E., & Ciampi, A. (2008). Recognition of Depression by Non-psychiatric Physicians—A Systematic Literature Review and Meta-analysis. *Journal of General Internal Medicine*, *23*(1), 25–36. https://doi.org/10.1007/s11606-007-0428-5

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chawla, N. V., & Davis, D. A. (2013). Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. *Journal of General Internal Medicine*, *28*(S3), 660–665. https://doi.org/10.1007/s11606-013-2455-8

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science (New York, N.Y.)*, *383*(6679), 164–167. https://doi.org/10.1126/science.adg8538

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X

Chen, M. C., Anderson, J. R., & Sohn, M. H. (2001). What can a mouse cursor tell us more? Correlation of eye/mouse movements on web browsing. *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, 281–282. https://doi.org/10.1145/634067.634234

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chien, I., Enrique, A., Palacios, J., Regan, T., Keegan, D., Carter, D., Tschiatschek, S., Nori, A., Thieme, A., Richards, D., Doherty, G., & Belgrave, D. (2020). A Machine Learning Approach to Understanding Patterns of Engagement With Internet-Delivered Mental Health Interventions. *JAMA Network Open*, *3*(7), e2010791. https://doi.org/10.1001/jamanetworkopen.2020.10791

Chong, P., Elovici, Y., & Binder, A. (2019). User Authentication Based on Mouse Dynamics Using Deep Neural Networks: A Comprehensive Study. *IEEE Transactions on Information Forensics and Security*, *15*, 1086–1101.

Chong, P., Tan, Y. X. M., Guarnizo, J., Elovici, Y., & Binder, A. (2018). Mouse Authentication Without the Temporal Aspect – What Does a 2D-CNN Learn? *2018 IEEE Security and Privacy Workshops (SPW)*, 15–21.

Christensen, H., Griffiths, K. M., & Farrer, L. (2009). Adherence in Internet Interventions for Anxiety and Depression: Systematic Review. *Journal of Medical Internet Research*, *11*(2), e1194. https://doi.org/10.2196/jmir.1194

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., & Goharian, N. (2018). SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. *Proceedings of the 27th International Conference on Computational Linguistics*, 1485–1497.

Colin, J., Fel, T., Cadène, R., & Serre, T. (2022). What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems*, *35*, 2832–2845.

Cook, B. L., Progovac, A. M., Chen, P., Mullin, B., Hou, S., & Baca-Garcia, E. (2016). Novel Use of Natural Language Processing (NLP) to Predict Suicidal Ideation and Psychiatric Symptoms in a Text-Based Mental Health Intervention in Madrid. *Computational and Mathematical Methods in Medicine*, *2016*, 1–8. https://doi.org/10.1155/2016/8708434

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cote-Allard, U., Pham, M. H., Schultz, A. K., Nordgreen, T., & Torresen, J. (2022). Adherence Forecasting for Guided Internet-Delivered Cognitive Behavioral Therapy: A Minimally Data-Sensitive Approach. *IEEE Journal of Biomedical and Health Informatics*, 1–12. https://doi.org/10.1109/JBHI.2022.3204737

Cruz Rivera, S., Liu, X., Hughes, S. E., Dunster, H., Manna, E., Denniston, A. K., & Calvert, M. J. (2023). Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. *The Lancet Digital Health*, *5*(3), e168–e173. https://doi.org/10.1016/S2589-7500(22)00252-7

Cuijpers, P., Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D., & van Straten, A. (2014). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: A meta-analysis. *Journal of Affective Disorders*, *159*, 118–126. https://doi.org/10.1016/j.jad.2014.02.026

D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology*, *36*, 112–117. https://doi.org/10.1016/j.copsyc.2020.04.005

Deighton, J., & Sorrell, M. (1996). The future of interactive marketing. *Harvard Business Review*, *74*(6), 151–160.

Delgadillo, J., & Gonzalez Salas Duhne, P. (2020). Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology*, *88*(1), 14–24. https://doi.org/10.1037/ccp0000476

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, *44*(3), 837–845. https://doi.org/10.2307/2531595

DeMasi, O., Kording, K., & Recht, B. (2017). Meaningless comparisons lead to false optimism in medical machine learning. *PLOS ONE*, *12*(9). https://doi.org/10.1371/journal.pone.0184604

Devaraj, S., Sharma, S. K., Fausto, D. J., Viernes, S., & Kharrazi, H. (2014). Barriers and Facilitators to Clinical Decision Support Systems Adoption: A Systematic Review. *Journal of Business Administration Research*, *3*(2), Article 2. https://doi.org/10.5430/jbar.v3n2p36

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*.

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Diprose, W. K., Buist, N., Hua, N., Thurier, Q., Shand, G., & Robinson, R. (2020). Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, *27*(4), 592–600. https://doi.org/10.1093/jamia/ocz229

Domhardt, M., Letsch, J., Kybelka, J., Koenigbauer, J., Doebler, P., & Baumeister, H. (2020). Are Internet- and mobile-based interventions effective in adults with diagnosed panic disorder and/or agoraphobia? A systematic review and meta-analysis. *Journal of Affective Disorders*, *276*, 169–182. https://doi.org/10.1016/j.jad.2020.06.059

Donker, T., Blankers, M., Hedman, E., Ljótsson, B., Petrie, K., & Christensen, H. (2015). Economic evaluations of Internet interventions for mental health: A systematic review. *Psychological Medicine*, *45*(16), 3357–3376. https://doi.org/10.1017/S0033291715001427

Donkin, L., Christensen, H., Naismith, S. L., Neal, B., Hickie, I. B., & Glozier, N. (2011). A Systematic Review of the Impact of Adherence on the Effectiveness of e-Therapies. *Journal of Medical Internet Research*, *13*(3). https://doi.org/10.2196/jmir.1772

Donkin, L., Hickie, I. B., Christensen, H., Naismith, S. L., Neal, B., Cockayne, N. L., & Glozier, N. (2013). Rethinking the Dose-Response Relationship Between Usage and Outcome in an Online Intervention for Depression: Randomized Controlled Trial. *Journal of Medical Internet Research*, *15*(10), e231. https://doi.org/10.2196/jmir.2771

dos Santos, T., & Santana, V. (2022). Identifying Distractors for People with Computer Anxiety Based on Mouse Fixations. *Interacting with Computers*, *35*. https://doi.org/10.1093/iwc/iwac025

Duffourc, M. N., & Gerke, S. (2023). The proposed EU Directives for AI liability leave worrying gaps likely to impact medical AI. *Npj Digital Medicine*, *6*(1), Article 1. https://doi.org/10.1038/s41746-023-00823-w

D'Zurilla, T. J., & Nezu, A. M. (2001). Problem-solving therapies. In K. S. Dobson (Ed.), *Handbook of cognitnve-behavioral therapies* (2nd ed., pp. 211–245). Guilford.

Ebert, D. D., Franke, M., Zarski, A.-C., Berking, M., Riper, H., Cuijpers, P., Funk, B., & Lehr, D. (2021). Effectiveness and Moderators of an Internet-Based Mobile-Supported Stress Management Intervention as a Universal Prevention Approach: Randomized Controlled Trial. *Journal of Medical Internet Research*, *23*(12), e22107. https://doi.org/10.2196/22107

Ebert, D. D., Harrer, M., Apolinário-Hagen, J., & Baumeister, H. (2019). Digital Interventions for Mental Disorders: Key Features, Efficacy, and Potential for Artificial Intelligence Applications. In Y.-K. Kim (Ed.), *Frontiers in Psychiatry* (Vol. 1192, pp. 583–627). Springer Singapore. https://doi.org/10.1007/978-981-32-9721-0_29

Ebert, D. D., Heber, E., Berking, M., Riper, H., Cuijpers, P., Funk, B., & Lehr, D. (2016). Self-guided internet-based and mobile-based stress management for employees: Results of a randomised controlled trial. *Occupational and Environmental Medicine*, *73*(5), 315–323. https://doi.org/10.1136/oemed-2015-103269

Ebert, D. D., Lehr, D., Heber, E., Riper, H., Cuijpers, P., & Berking, M. (2016). Internet- and mobile-based stress management for employees with adherence-focused guidance: Efficacy and mechanism of change. *Scandinavian Journal of Work, Environment & Health*, *42*(5), 382–394. https://doi.org/10.5271/sjweh.3573

El Alaoui, S., Hedman, E., Kaldo, V., Hesser, H., Kraepelien, M., Andersson, E., Rück, C., Andersson, G., Ljótsson, B., & Lindefors, N. (2015). Effectiveness of Internet-based cognitive-behavior therapy for social anxiety disorder in clinical psychiatry. *Journal of Consulting and Clinical Psychology*, *83*(5), 902–914. https://doi.org/10.1037/a0039198

Eloranta, S., & Boman, M. (2022). Predictive models for clinical decision making: Deep dives in practical machine learning. *Journal of Internal Medicine*, *292*(2), 278–295. https://doi.org/10.1111/joim.13483

Englhardt, Z., Ma, C., Morris, M. E., Chang, C.-C., Xu, X. "Orson," Qin, L., McDuff, D., Liu, X., Patel, S., & Iyer, V. (2024). From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, *8*(2), 56:1-56:25. https://doi.org/10.1145/3659604

Everitt, N., Broadbent, J., Richardson, B., Smyth, J. M., Heron, K., Teague, S., & Fuller-Tyszkiewicz, M. (2021). Exploring the features of an app-based just-in-time intervention for depression. *Journal of Affective Disorders*, *291*, 279–287. https://doi.org/10.1016/j.jad.2021.05.021

Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2020). Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. *JAMA Psychiatry*, *77*(1), 35–43. https://doi.org/10.1001/jamapsychiatry.2019.2664

Eysenbach, G. (2005). The Law of Attrition. *Journal of Medical Internet Research*, *7*(1), e11. https://doi.org/10.2196/jmir.7.1.e11

Fairburn, C. G., & Beglin, S. J. (2008). Eating Disorder Examination Questionnaire. In *Cognitive Behavior Therapy and Eating Disorders*. Guildford Press.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, *9*(61), 1871–1874.

Fantino, B., & Moore, N. (2009). The self-reported Montgomery-Asberg Depression Rating Scale is a useful evaluative tool in Major Depressive Disorder. *BMC Psychiatry*, *9*, 26. https://doi.org/10.1186/1471-244X-9-26

Feher, C., Elovici, Y., Moskovitch, R., Rokach, L., & Schclar, A. (2012). User identity verification via mouse dynamics. *Inf. Sci.*, *201*, 19–36.

Fehle, J., Schmidt, T., & Wolff, C. (2021). *Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques*. Konferenz zur Verarbeitung natürlicher Sprache, Düsseldorf.

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). *Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?*

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J. L., Vos, T., & Whiteford, H. A. (2013). Burden of Depressive Disorders by Country, Sex, Age, and Year: Findings from the Global Burden of Disease Study 2010. *PLoS Medicine*, *10*(11), e1001547. https://doi.org/10.1371/journal.pmed.1001547

Forsell, E., Isacsson, N., Blom, K., Jernelöv, S., Ben Abdesslem, F., Lindefors, N., Boman, M., & Kaldo, V. (2020). Predicting treatment failure in regular care Internet-Delivered Cognitive Behavior Therapy for depression and anxiety using only weekly symptom measures. *Journal of Consulting and Clinical Psychology*, *88*(4), 311–321. https://doi.org/10.1037/ccp0000462

Forsell, E., Jernelöv, S., Blom, K., & Kaldo, V. (2022). Clinically sufficient classification accuracy and key predictors of treatment failure in a randomized controlled trial of Internet-delivered Cognitive Behavior Therapy for Insomnia. *Internet Interventions*, *29*. https://doi.org/10.1016/j.invent.2022.100554

Forsell, E., Jernelöv, S., Blom, K., Kraepelien, M., Svanborg, C., Andersson, G., Lindefors, N., & Kaldo, V. (2019). Proof of Concept for an Adaptive Treatment Strategy to Prevent Failures in Internet-Delivered CBT: A Single-Blind Randomized Clinical Trial With Insomnia Patients. *American Journal of Psychiatry*, *176*(4), 315–323. https://doi.org/10.1176/appi.ajp.2018.18060699

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 72. https://doi.org/10.1186/s12916-015-0325-4

Fu, E. Y., Kwok, T. C. K., Wu, E. Y., Leong, H. V., Ngai, G., & Chan, S. C. F. (2017). Your Mouse Reveals Your Next Activity: Towards Predicting User Intention from Mouse Interaction. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, *1*, 869–874. https://doi.org/10.1109/COMPSAC.2017.270

Funk, B., Sadeh-Sharvit, S., Fitzsimmons-Craft, E. E., Trockel, M. T., Monterubio, G. E., Goel, N. J., Balantekin, K. N., Eichen, D. M., Flatt, R. E., Firebaugh, M.-L., Jacobi, C., Graham, A. K., Hoogendoorn, M., Wilfley, D. E., & Taylor, C. B. (2020). A Framework for Applying Natural Language Processing in Digital Health Interventions. *Journal of Medical Internet Research*, *22*(2), e13855. https://doi.org/10.2196/13855

Furukawa, T. A., Katherine Shear, M., Barlow, D. H., Gorman, J. M., Woods, S. W., Money, R., Etschel, E., Engel, R. R., & Leucht, S. (2009). Evidence-based guidelines for interpretation of the Panic Disorder Severity Scale. *Depression and Anxiety*, *26*(10), 922–929. https://doi.org/10.1002/da.20532

Furukawa, T. A., Suganuma, A., Ostinelli, E. G., Andersson, G., Beevers, C. G., Shumake, J., Berger, T., Boele, F. W., Buntrock, C., Carlbring, P., Choi, I., Christensen, H., Mackinnon, A., Dahne, J., Huibers, M. J. H., Ebert, D. D., Farrer, L., Forand, N. R., Strunk, D. R., … Cuijpers, P. (2021). Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: A systematic review and component network meta-analysis using individual participant data. *The Lancet Psychiatry*, *8*(6), 500–511. https://doi.org/10.1016/S2215-0366(21)00077-8

Galmiche, M., Déchelotte, P., Lambert, G., & Tavolacci, M. P. (2019). Prevalence of eating disorders over the 2000–2018 period: A systematic literature review. *The American Journal of Clinical Nutrition*, *109*(5), 1402–1413. https://doi.org/10.1093/ajcn/nqy342

Gamboa, H., & Fred, A. (2004). A behavioral biometric system based on human-computer interaction. *Proc SPIE*, *5404*, 381–392. https://doi.org/10.1117/12.542625

Gan, D. Z. Q., McGillivray, L., Han, J., Christensen, H., & Torok, M. (2021). Effect of Engagement With Digital Interventions on Mental Health Outcomes: A Systematic Review and Meta-Analysis. *Frontiers in Digital Health*, *3*, 764079. https://doi.org/10.3389/fdgth.2021.764079

Gan, D. Z. Q., McGillivray, L., Larsen, M. E., Christensen, H., & Torok, M. (2022). Technology-supported strategies for promoting user engagement with digital mental health interventions: A systematic review. *DIGITAL HEALTH*, *8*, 205520762210982. https://doi.org/10.1177/20552076221098268

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, *178*(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Gidlöf, K., Wallin, A., Dewhurst, R., & Holmqvist, K. (2013). Using Eye Tracking to Trace a Cognitive Process: Gaze Behaviour During Decision Making in a Natural Environment. *Journal of Eye Movement Research*, *6*(1), Article 1. https://doi.org/10.16910/jemr.6.1.3

Giesemann, J., Delgadillo, J., Schwartz, B., Bennemann, B., & Lutz, W. (2023). Predicting dropout from psychological treatment using different machine learning algorithms, resampling methods, and sample sizes. *Psychotherapy Research*, *33*(6), 683–695. https://doi.org/10.1080/10503307.2022.2161432

Gille, F., Jobin, A., & Ienca, M. (2020). What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine*, *1–2*, 100001. https://doi.org/10.1016/j.ibmed.2020.100001

Gledson, A., Apaolaza, A., Barthold, S., Günther, F., Yu, H., & Vigo, M. (2021). Characterising Student Engagement Modes through Low-Level Activity Patterns. *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 88–97. https://doi.org/10.1145/3450613.3456818

Gogoulou, E., Boman, M., Ben Abdesslem, F., Hentati Isacsson, N., Kaldo, V., & Sahlgren, M. (2021). Predicting Treatment Outcome from Patient Texts:The Case of Internet-Based Cognitive Behavioural Therapy. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 575–580. https://doi.org/10.18653/v1/2021.eacl-main.46

Gonzalez Salas Duhne, P., Delgadillo, J., & Lutz, W. (2022). Predicting early dropout in online versus face-to-face guided self-help: A machine learning approach. *Behaviour Research and Therapy*, *159*, 104200. https://doi.org/10.1016/j.brat.2022.104200

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, *38*(10), 1004–1015. https://doi.org/10.1080/0144929X.2019.1656779

Gunlicks-Stoessel, M., Klimes-Dougan, B., VanZomeren, A., & Ma, S. (2020). Developing a data-driven algorithm for guiding selection between cognitive behavioral therapy, fluoxetine, and combination treatment for adolescent depression. *Translational Psychiatry*, *10*(1), 321. https://doi.org/10.1038/s41398-020-01005-y

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Günther, F., Yau, C., Elison-Davies, S., & Wong, D. (2023). On the Difficulty of Predicting Engagement with Digital Health for Substance Use. *Studies in Health Technology and Informatics*, *302*, 967–971. https://doi.org/10.3233/SHTI230319

Guyon, I., Saffari, A., Dror, G., & Cawley, G. (2011). Model Selection: Beyond the Bayesian/Frequentist Divide. *Journal of Machine Learning Research*, 61–87.

Haller, K., Becker, P., Niemeyer, H., & Boettcher, J. (2023). Who benefits from guided internet-based interventions? A systematic review of predictors and moderators of treatment outcome. *Internet Interventions*, *33*, 100635. https://doi.org/10.1016/j.invent.2023.100635

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hashimoto, K., Xiong, C., Tsuruoka, Y., & Socher, R. (2017). A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1923–1933). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1206

Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, W. J., Stohl, M., & Grant, B. F. (2018). Epidemiology of Adult *DSM-5* Major Depressive Disorder and Its Specifiers in the United States. *JAMA Psychiatry*, *75*(4), 336. https://doi.org/10.1001/jamapsychiatry.2017.4602

Hassan Hosny, H. A., Ibrahim, A. A., Elmesalawy, M. M., & Abd El-Haleem, A. M. (2022). An Intelligent Approach for Fair Assessment of Online Laboratory Examinations in Laboratory Learning Systems Based on Student's Mouse Interaction Behavior. *Applied Sciences*, *12*(22), Article 22. https://doi.org/10.3390/app122211416

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

Hatcher, S., Whittaker, R., Patton, M., Miles, W. S., Ralph, N., Kercher, K., & Sharon, C. (2018). Web-based Therapy Plus Support by a Coach in Depressed Patients Referred to Secondary Mental Health Care: Randomized Controlled Trial. *JMIR Mental Health*, *5*(1), e5. https://doi.org/10.2196/mental.8510

Heber, E., Ebert, D. D., Lehr, D., Cuijpers, P., Berking, M., Nobis, S., & Riper, H. (2017). The Benefit of Web- and Computer-Based Interventions for Stress: A Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *19*(2), e32. https://doi.org/10.2196/jmir.5774

Heber, E., Lehr, D., Ebert, D. D., Berking, M., & Riper, H. (2016). Web-Based and Mobile Stress Management Intervention for Employees: A Randomized Controlled Trial. *Journal of Medical Internet Research*, *18*(1), e5112. https://doi.org/10.2196/jmir.5112

Hedman, E., Ljótsson, B., Kaldo, V., Hesser, H., El Alaoui, S., Kraepelien, M., Andersson, E., Rück, C., Svanborg, C., Andersson, G., & Lindefors, N. (2014). Effectiveness of Internet-based cognitive behaviour therapy for depression in routine psychiatric care. *Journal of Affective Disorders*, *155*, 49–58. https://doi.org/10.1016/j.jad.2013.10.023

Hedman, E., Ljótsson, B., Rück, C., Bergström, J., Andersson, G., Kaldo, V., Jansson, L., Andersson, E., Andersson, E., Blom, K., El Alaoui, S., Falk, L., Ivarsson, J., Nasri, B., Rydh, S., & Lindefors, N. (2013). Effectiveness of Internet-based cognitive behaviour therapy for panic disorder in routine psychiatric care. *Acta Psychiatrica Scandinavica*, *128*(6), 457–467. https://doi.org/10.1111/acps.12079

Heim, E., Ramia, J. A., Hana, R. A., Burchert, S., Carswell, K., Cornelisz, I., Cuijpers, P., El Chammay, R., Noun, P., Van Klaveren, C., Van Ommeren, M., Zoghbi, E., & Van'T Hof, E. (2021). Step-by-step: Feasibility randomised controlled trial of a mobile-based intervention for depression among populations affected by adversity in Lebanon. *Internet Interventions*, *24*, 100380. https://doi.org/10.1016/j.invent.2021.100380

Hentati Isacsson, N., Abdesslem, F. B., Forsell, E., Boman, M., & Kaldo, V. (2023). *Machine learning predictions of outcome in Internet-based cognitive behavioral therapy: Methodological choices and clinical usefulness*. https://doi.org/10.21203/rs.3.rs-2751455/v1

Higgins, O., Short, B. L., Chalup, S. K., & Wilson, R. L. (2023). Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review. *International Journal of Mental Health Nursing*, *32*(4), 966–978. https://doi.org/10.1111/inm.13114

Hilbert, K., Böhnlein, J., Meinke, C., Chavanne, A., Langhammer, T., Stumpe, L., Winter, N., Leenings, R., Adolph, D., Arolt, V., Bischoff, S., Cwik, J., Deckert, J., Domschke, K., Fydrich, T., Gathmann, B., Hamm, A., Heinig, I., Herrmann, M., & Lueken, U. (2024). Lack of evidence for predictive utility from resting state fMRI data for individual exposure-based cognitive behavioral therapy outcomes: A machine learning study in two large multi-site samples in anxiety disorders. *NeuroImage*, *295*, 120639. https://doi.org/10.1016/j.neuroimage.2024.120639

Hilvert-Bruce, Z., Rossouw, P. J., Wong, N., Sunderland, M., & Andrews, G. (2012). Adherence as a determinant of effectiveness of internet cognitive behavioural therapy for anxiety and depressive disorders. *Behaviour Research and Therapy*, *50*(7–8), 463–468. https://doi.org/10.1016/j.brat.2012.04.001

Himle, J. A., Weaver, A., Zhang, A., & Xiang, X. (2022). Digital Mental Health Interventions for Depression. *Cognitive and Behavioral Practice*, *29*(1), 50–59. https://doi.org/10.1016/j.cbpra.2020.12.009

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv Preprint arXiv:1812.04608*.

Hoogendoorn, M., Berger, T., Schulz, A., Stolz, T., & Szolovits, P. (2017). Predicting Social Anxiety Treatment Outcome Based on Therapeutic Email Conversations. *IEEE Journal of Biomedical and Health Informatics*, *21*(5), 1449–1459. https://doi.org/10.1109/JBHI.2016.2601123

Hornstein, S., Forman-Hoffman, V., Nazander, A., Ranta, K., & Hilbert, K. (2021). Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach. *DIGITAL HEALTH*, *7*, 205520762110606. https://doi.org/10.1177/20552076211060659

Hornstein, S., Seiler, M., Hoffman, V., Nelson, B., Aschbacher, K., Ritter, K., & Hilbert, K. (2022). *Association of Depressive Symptoms with Resting Heart Rate Variability recorded from a Wearable Device under Naturalistic Conditions: A Machine Learning Study*. https://doi.org/10.31234/osf.io/9z3pr

Hornstein, S., Zantvoort, K., Ulrike, L., Funk, B., & Kevin Hilbert. (2023). Personalization Strategies in Digital Mental Health Interventions: A Systematic Review and Conceptual Framework for Depressive Symptoms. *Front. Digit. Health*, *5*. https://doi.org/10.3389/fdgth.2023.1170002

Houck, P. R., Spiegel, D. A., Shear, M. K., & Rucci, P. (2002). Reliability of the self-report version of the panic disorder severity scale. *Depression and Anxiety*, *15*(4), 183–185. https://doi.org/10.1002/da.10049

Howes, C., Purver, M., & McCabe, R. (2014). Linguistic Indicators of Severity and Progress in Online Text-based Therapy for Depression. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 7–16. https://doi.org/10.3115/v1/W14-3202

Hucko, M., Gazo, L., Simun, P., Valky, M., Moro, R., Simko, J., & Bielikova, M. (2019). YesElf: Personalized Onboarding for Web Applications. *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 39–44. https://doi.org/10.1145/3314183.3324978

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The Worst of Both Worlds: A Comparative Analysis of Errors in Learning from Data in Psychology and Machine Learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 335–348. https://doi.org/10.1145/3514094.3534196

Hung, G. C.-L., Yang, P.-C., Wang, C.-Y., & Chiang, J.-H. (2015). A Smartphone-Based Personalized Activity Recommender System for Patients with Depression. *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare - "Transforming Healthcare through Innovations in Mobile and Wireless Technologies."* 5th EAI International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies," London, Great Britain. https://doi.org/10.4108/eai.14-10-2015.2261655

Jacobi, C., Morris, L., Beckers, C., Bronisch-Holtze, J., Winter, J., Winzelberg, A. J., & Taylor, C. B. (2007). Maintenance of internet-based prevention: A randomized controlled trial. *International Journal of Eating Disorders*, *40*(2), 114–119. https://doi.org/10.1002/eat.20344

Jacobi, C., Völker, U., Trockel, M. T., & Taylor, C. B. (2012). Effects of an Internet-based intervention for subthreshold eating disorders: A randomized controlled trial. *Behaviour Research and Therapy*, *50*(2), 93–99. https://doi.org/10.1016/j.brat.2011.09.013

Jacobi, C., Vollert, B., Hütter, K., Bloh, P. von, Eiterich, N., Görlich, D., & Taylor, C. B. (2022). Indicated Web-Based Prevention for Women With Anorexia Nervosa Symptoms: Randomized Controlled Efficacy Trial. *Journal of Medical Internet Research*, *24*(6), e35947. https://doi.org/10.2196/35947

Jacobs, M., He, J., F. Pradier, M., Lam, B., Ahn, A. C., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445385

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021). *MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare* (arXiv:2110.15621). arXiv. https://doi.org/10.48550/arXiv.2110.15621

Jin, W., Fatehi Hassanabad, M., Guo, R., & Hamarneh, G. (2024). Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. *Artificial Intelligence in Medicine*, *148*, 102751. https://doi.org/10.1016/j.artmed.2023.102751

Kaltenthaler, E., Parry, G., Beverley, C., & Ferriter, M. (2008). Computerised cognitive-behavioural therapy for depression: Systematic review. *British Journal of Psychiatry*, *193*(3), 181–184. https://doi.org/10.1192/bjp.bp.106.025981

Kaltenthaler, E., Sutcliffe, P., Parry, G., Beverley, C., Rees, A., & Ferriter, M. (2008). The acceptability to patients of computerized cognitive behaviour therapy for depression: A systematic review. *Psychological Medicine*, *38*(11), 1521–1530. https://doi.org/10.1017/S0033291707002607

Kaptein, M., & Parvinen, P. (2015). Advancing E-Commerce Personalization: Process Framework and Case Study. *International Journal of Electronic Commerce*, *19*(3), 7–33. https://doi.org/10.1080/10864415.2015.1000216

Karin, E., Dear, B. F., Heller, G. Z., Gandy, M., & Titov, N. (2018). Measurement of Symptom Change Following Web-Based Psychotherapy: Statistical Characteristics and Analytical Methods for Measuring and Interpreting Change. *JMIR Mental Health*, *5*(3), e10200. https://doi.org/10.2196/10200

Karyotaki, E., Efthimiou, O., Miguel, C., Bermpohl, F. M. genannt, Furukawa, T. A., Cuijpers, P., & Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration. (2021). Internet-Based Cognitive Behavioral Therapy for Depression: A Systematic Review and Individual Patient Data Network Meta-analysis. *JAMA Psychiatry*, *78*(4), 361–371. https://doi.org/10.1001/jamapsychiatry.2020.4364

Karyotaki, E., Kleiboer, A., Smit, F., Turner, D. T., Pastor, A. M., Andersson, G., Berger, T., Botella, C., Breton, J. M., Carlbring, P., Christensen, H., de Graaf, E., Griffiths, K., Donker, T., Farrer, L., Huibers, M. J. H., Lenndin, J., Mackinnon, A., Meyer, B., … Cuijpers, P. (2015). Predictors of treatment dropout in self-guided web-based interventions for depression: An 'individual patient data' meta-analysis. *Psychological Medicine*, *45*(13), 2717–2726. https://doi.org/10.1017/S0033291715000665

Kennedy, B., Reimer, N. K., & Dehghani, M. (2021). *Explaining Explainability: Interpretable machine learning for the behavioral sciences* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9h6qr

Khairat, S., Marc, D., Crosby, W., & Sanousi, A. A. (2018). Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. *JMIR Medical Informatics*, *6*(2), e8912. https://doi.org/10.2196/medinform.8912

Killen, J. D., Taylor, C., Hayward, C., Wilson, D. M., Haydel, K. F., Hammer, L. D., Simmonds, B., Robinson, T. N., Litt, I., & Varady, A. (1994). Pursuit of thinness and onset of eating disorder symptoms in a community sample of adolescent girls: A three-year prospective analysis. *The International Journal of Eating Disorders*, *16*(3). https://doi.org/10.1002/1098-108x(199411)16:3<227::aid-eat2260160303>3.0.co;2-l

Kim, M., Yang, J., Ahn, W.-Y., & Choi, H. J. (2021). Machine Learning Analysis to Identify Digital Behavioral Phenotypes for Engagement and Health Outcome Efficacy of an mHealth Intervention for Obesity: Randomized Controlled Trial. *Journal of Medical Internet Research*, *23*(6), e27218. https://doi.org/10.2196/27218

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. https://doi.org/10.1177/2053951714528481

Königbauer, J., Letsch, J., Doebler, P., Ebert, D., & Baumeister, H. (2017). Internet- and mobile-based depression interventions for people with diagnosed depression: A systematic review and meta-analysis. *Journal of Affective Disorders*, *223*, 28–40. https://doi.org/10.1016/j.jad.2017.07.021

Kotu, V., & Deshpande, B. (2014). *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner* (1st ed.). Morgan Kaufmann Publishers Inc.

Kraemer, H., Morgan, G., Leech, N., Gliner, J., Vaske, J., & Harmon, R. (2003). Measures of Clinical Significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 1524–1529. https://doi.org/10.1097/00004583-200312000-00022

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kulkarni, P. V., Rai, S., & Kale, R. (2020). Recommender System in eLearning: A Survey. In S. Bhalla, P. Kwan, M. Bedekar, R. Phalnikar, & S. Sirsikar (Eds.), *Proceeding of International Conference on Computational Science and Applications* (pp. 119–126). Springer Singapore. https://doi.org/10.1007/978-981-15-0790-8_13

Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40*(5), 1847–1857. https://doi.org/10.1016/j.eswa.2012.09.017

Lamers, F., Van Oppen, P., Comijs, H. C., Smit, J. H., Spinhoven, P., Van Balkom, A. J. L. M., Nolen, W. A., Zitman, F. G., Beekman, A. T. F., & Penninx, B. W. J. H. (2011). Comorbidity Patterns of Anxiety and Depressive Disorders in a Large Cohort Study: The Netherlands Study of

Depression and Anxiety (NESDA). *The Journal of Clinical Psychiatry*, *72*(03), 341–348. https://doi.org/10.4088/JCP.10m06176blu

Lamo, Y., Mukhiya, S. K., Rabbi, F., Aminifar, A., Lillehaug, S. I., Tørresen, J., H Pham, M., Côté-Allard, U., Noori, F. M., Guribye, F., Inal, Y., Flobakk, E., Wake, J. D., Myklebost, S., Lundervold, A. J., Hammar, A., Nordby, E., Kahlon, S., Kenter, R., … Nordgreen, T. (2022). Towards adaptive technology in routine mental health care. *DIGITAL HEALTH*, *8*. https://doi.org/10.1177/20552076221128678

Lateh, M. A., Kamilah Muda, A., Yusof, Z. I. M., Azilah Muda, N., & Sanusi Azmi, M. (2017). *Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review*. *892*. Journal of Physics Conference Series. https://doi.org/10.1088/1742-6596/892/1/012016

Lau, Y., Chee, D. G. H., Chow, X. P., Cheng, L. J., & Wong, S. N. (2020). Personalised eHealth interventions in adults with overweight and obesity: A systematic review and meta-analysis of randomised controlled trials. *Preventive Medicine*, *132*, 106001. https://doi.org/10.1016/j.ypmed.2020.106001

Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2021). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research*, *23*(5), e15708. https://doi.org/10.2196/15708

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning*, 1188–1196. https://proceedings.mlr.press/v32/le14.html

Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C.-H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Leiva, L., & Arapakis, I. (2020). The Attentive Cursor Dataset. *Frontiers in Human Neuroscience*, *14*. https://doi.org/10.3389/fnhum.2020.565664

Lepora, N. F., & Pezzulo, G. (2015). Embodied Choice: How Action Influences Perceptual Decision Making. *PLOS Computational Biology*, *11*(4), e1004110. https://doi.org/10.1371/journal.pcbi.1004110

Lim, G. Y., Tam, W. W., Lu, Y., Ho, C. S., Zhang, M. W., & Ho, R. C. (2018). Prevalence of Depression in the Community from 30 Countries between 1994 and 2014. *Scientific Reports*, *8*(1), 2861. https://doi.org/10.1038/s41598-018-21243-x

Linardon, J., Fuller-Tyszkiewicz, M., Shatte, A., & Greenwood, C. J. (2022). An exploratory application of machine learning methods to optimize prediction of responsiveness to digital interventions for eating disorder symptoms. *International Journal of Eating Disorders*, *55*(6), 845–850. https://doi.org/10.1002/eat.23733

Linardon, J., Shatte, A., Messer, M., Firth, J., & Fuller-Tyszkiewicz, M. (2020). E-mental health interventions for the treatment and prevention of eating disorders: An updated systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *88*(11), 994–1007. https://doi.org/10.1037/ccp0000575

Linnet, J., Hertz, S. P. T., Jensen, E. S., Runge, E., Tarp, K. H. H., Holmberg, T. T., Mathiasen, K., & Lichtenstein, M. B. (2023). Days between sessions predict attrition in text-based internet intervention of Binge Eating Disorder. *Internet Interventions*, *31*, 100607. https://doi.org/10.1016/j.invent.2023.100607

Lipschitz, J., Pike, C., Hogan, T., Murphy, S., & Burdick, K. (2023). The Engagement Problem: A Review of Engagement with Digital Mental Health Interventions and Recommendations for a Path Forward. *Current Treatment Options in Psychiatry*, *10*. https://doi.org/10.1007/s40501-023-00297-3

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Liu, Z., Zou, L., Zou, X., Wang, C., Zhang, B., Tang, D., Zhu, B., Zhu, Y., Wu, P., Wang, K., & Cheng, Y. (2022). *Monolith: Real Time Recommendation System With Collisionless Embedding Table* (arXiv:2209.07663). arXiv. http://arxiv.org/abs/2209.07663

Loftus, T. J., Ruppert, M. M., Shickel, B., Ozrazgat-Baslanti, T., Balch, J. A., Efron, P. A., Upchurch, G. R., Rashidi, P., Tignanelli, C., Bian, J., & Bihorac, A. (2022). Federated learning for preserving data privacy in collaborative healthcare research. *DIGITAL HEALTH*, *8*. https://doi.org/10.1177/20552076221134455

Luborsky, L. (1984). The use of psychotherapy treatment manuals: A small revolution in psychotherapy research style. *Clinical Psychology Review*, *4*(1), 5–14. https://doi.org/10.1016/0272-7358(84)90034-5

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. pp 4768-477.

Lustria, M. L. A., Noar, S. M., Cortese, J., Van Stee, S. K., Glueckauf, R. L., & Lee, J. (2013). A Meta-Analysis of Web-Delivered Tailored Health Behavior Change Interventions. *Journal of Health Communication*, *18*(9), 1039–1069. https://doi.org/10.1080/10810730.2013.768727

Lutz, W., Deisenhofer, A.-K., Rubel, J., Bennemann, B., Giesemann, J., Poster, K., & Schwartz, B. (2022). Prospective evaluation of a clinical decision support system in psychological therapy. *Journal of Consulting and Clinical Psychology*, *90*(1), 90–106. https://doi.org/10.1037/ccp0000642

Lutz, W., Schaffrath, J., Eberhardt, S., Hehlmann, M., Schwartz, B., Deisenhofer, A.-K., Vehlen, A., Vaccarezza Schürmann, S., Uhl, J., & Moggia, D. (2023). Precision Mental Health and Data-Informed Decision Support in Psychological Therapy: An Example. *Administration and Policy in Mental Health and Mental Health Services Research*, 1–12. https://doi.org/10.1007/s10488-023-01330-6

Maldonado, M., Dunbar, E., & Chemla, E. (2019). Mouse tracking as a window into decision making. *Behavior Research Methods*, *51*(3), 1085–1101. https://doi.org/10.3758/s13428-018-01194-x

Marcus, M. D., & Wildes, J. E. (2012). Obesity in DSM-5. *Psychiatric Annals*, *42*(11), 431–435. https://doi.org/10.3928/00485713-20121105-10

Martin, D. (2019). *11. International Telecommunication Union. Measuring digital development Facts and figures. ITU. (2019). Https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ [Accessed Dec 15, 2022]*. ITUPublications.

Martín-Albo, D., Leiva, L. A., Huang, J., & Plamondon, R. (2016). Strokes of insight: User intent detection and kinematic compression of mouse cursor trails. *Information Processing & Management*, *52*(6), 989–1003. https://doi.org/10.1016/j.ipm.2016.04.005

Masino, A. J., Forsyth, D., & Fiks, A. G. (2018). Detecting Adverse Drug Reactions on Twitter with Convolutional Neural Networks and Word Embedding Features. *Journal of Healthcare Informatics Research*, *2*(1–2), 25–43. https://doi.org/10.1007/s41666-018-0018-9

Maslej, M., Kloiber, S., Ghassemi, M., Yu, J., & Hill, S. (2023). Out with AI, in with the psychiatrist: A preference for human-derived clinical decision support in depression care. *Translational Psychiatry*, *13*. https://doi.org/10.1038/s41398-023-02509-z

Masso, A., Kaun, A., & van Noordt, C. (2023). Basic values in artificial intelligence: Comparative factor analysis in Estonia, Germany, and Sweden. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01750-w

Matthiesen, J., & Holte, M. B. (2019). Simple Mouse Attribute Analysis: HCIBGO: 6th International Conference on HCI in Business, Government and Organizations under HCI International 2019. *HCI in Business, Government and Organizations. Information Systems and Analytics*, 95–113. https://doi.org/10.1007/978-3-030-22338-0_8

Matthiesen, J. J., & Brefeld, U. (2020). Assessing User Behavior by Mouse Movements. *22nd International Conference HCI International*, 68–75.

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Mendoza, J. E., & Foundas, A. L. (2008). *Clinical Neuroanatomy: A Neurobehavioral Approach*. Springer. https://dokumen.pub/clinical-neuroanatomy-a-neurobehavioral-approach-9780387366012-0387366016.html

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). *Advances in Pre-Training Distributed Word Representations*.

Milisavljevic, A., Abate, F., Le Bras, T., Gosselin, B., Mancas, M., & Doré-Mazars, K. (2021). Similarities and Differences Between Eye and Mouse Dynamics During Web Pages Exploration. *Frontiers in Psychology*, *12*, 398. https://doi.org/10.3389/fpsyg.2021.554595

Mink, J. (2008). The basal ganglia. In L. R. Squire (Ed.), *Fundamental neuroscience* (3rd ed). Elsevier / Academic Press.

Moe-Byrne, T., Shepherd, J., Merecz-Kot, D., Sinokki, M., Naumanen, P., Hakkaart-van Roijen, L., & Van Der Feltz-Cornelis, C. (2022). Effectiveness of tailored digital health interventions for mental health at the workplace: A systematic review of randomised controlled trials. *PLOS Digital Health*, *1*(10), e0000123. https://doi.org/10.1371/journal.pdig.0000123

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry: The Journal of Mental Science*, *134*, 382–389. https://doi.org/10.1192/bjp.134.4.382

Moroz, N., Moroz, I., & D'Angelo, M. S. (2020). Mental health services in Canada: Barriers and cost-effective solutions to increase access. *Healthcare Management Forum*, *33*(6), 282–287. https://doi.org/10.1177/0840470420933911

Moshe, I., Terhorst, Y., Paganini, S., Schlicker, S., Pulkki-Råback, L., Baumeister, H., Sander, L. B., & Ebert, D. D. (2022). Predictors of Dropout in a Digital Intervention for the Prevention and Treatment of Depression in Patients With Chronic Back Pain: Secondary Analysis of Two Randomized Controlled Trials. *Journal of Medical Internet Research*, *24*(8), e38261. https://doi.org/10.2196/38261

Moshe, I., Terhorst, Y., Philippi, P., Domhardt, M., Cuijpers, P., Cristea, I., Pulkki-Råback, L., Baumeister, H., & Sander, L. B. (2021). Digital interventions for the treatment of depression: A meta-analytic review. *Psychological Bulletin*, *147*(8), 749–786. https://doi.org/10.1037/bul0000334

Nacke, B., Beintner, I., Görlich, D., Vollert, B., Schmidt-Hantke, J., Hütter, K., Taylor, C. B., & Jacobi, C. (2019). everyBody–Tailored online health promotion and eating disorder prevention for women: Study protocol of a dissemination trial. *Internet Interventions*, *16*, 20–25. https://doi.org/10.1016/j.invent.2018.02.008

Nacke, B., Görlich, D., Beitner, I., Vollert, B., Schmidt-Hantke, J., Taylor, C. B., & Jacobi, C. (2024). *Tailored online eating disorder prevention and health promotion for women: Results of a dissemination trial*. In preparation.

Naegelin, M., Weibel, R. P., Kerr, J. I., Schinazi, V. R., La Marca, R., von Wangenheim, F., Hoelscher, C., & Ferrario, A. (2023). An interpretable machine learning approach to multimodal stress detection in a simulated office environment. *Journal of Biomedical Informatics*, *139*, 104299. https://doi.org/10.1016/j.jbi.2023.104299

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, *55*(13s), 1–42.

Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, *33*(4), 275–306. https://doi.org/10.1007/s10462-010-9156-z

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics*, *9*(1), 12. https://doi.org/10.1186/s13326-018-0179-8

Nixon, P., Ebert, D., Boß, L., Angerer, P., Dragano, N., & Lehr, D. (n.d.). *Web-based stress management intervention for employees experiencing effort-reward imbalance at work: A randomized controlled trial*. https://doi.org/10.2196/40488

Nixon, P., Ebert, D., Boß, L., Angerer, P., Dragano, N., & Lehr, D. (2022). Efficacy of a web-based stress management intervention for employees experiencing adverse working conditions and occupational self-efficacy as mediator: A randomized controlled trial (Preprint). *Journal of Medical Internet Research*, *24*. https://doi.org/10.2196/40488

Nobles, A. L., Glenn, J. J., Kowsari, K., Teachman, B. A., & Barnes, L. E. (2018). Identification of Imminent Suicide Risk Among Young Adults using Text Messages. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3173574.3173987

North, F., & Chaudhry, R. (2016). Apple HealthKit and Health App: Patient Uptake and Barriers in Primary Care. *Telemedicine and E-Health*, *22*(7), 608–613. https://doi.org/10.1089/tmj.2015.0106

O'Dea, S. (2021). *Number of smartphone users worldwide from 2016 to 2021*.

Oesterreich, T., Fitte, C., Behne, A., & Teuteberg, F. (2020). *Understanding the Role of Predictive and Prescriptive Analytics in Healthcare: A Multi-Stakeholder Approach*.

Olczak, J., Pavlopoulos, J., Prijs, J., Ijpma, F. F. A., Doornberg, J. N., Lundström, C., Hedlund, J., & Gordon, M. (2021). Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: An introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta Orthopaedica*, *92*(5), 513. https://doi.org/10.1080/17453674.2021.1918389

Ong, A. A., & Gillespie, M. B. (2016). Overview of smartphone applications for sleep analysis. *World Journal of Otorhinolaryngology - Head and Neck Surgery*, *2*(1), 45–49. https://doi.org/10.1016/j.wjorl.2016.02.001

Pasini, A. (2015). Artificial neural networks for small dataset analysis. *Journal of Thoracic Disease*, *7*(5), 953–960. https://doi.org/10.3978/j.issn.2072-1439.2015.04.61

Patel, V., Araya, R., Chowdhary, N., King, M., Kirkwood, B., Nayak, S., Simon, G., & Weiss, H. (2008). Detecting common mental disorders in primary care in India: A comparison of five screening questionnaires. *Psychological Medicine*, *38*, 221–228. https://doi.org/10.1017/S0033291707002334

Paul, A., Liao, W.-K., Choudhary, A., & Agrawal, A. (2021). Harnessing Psycho-lingual and Crowd-Sourced Dictionaries for Predicting Taboos in Written Emotional Disclosure in Anonymous Confession Boards. *Journal of Healthcare Informatics Research*, *5*(3), 319–341. https://doi.org/10.1007/s41666-021-00092-w

Pearson, R., Pisner, D., Meyer, B., Shumake, J., & Beevers, C. G. (2019). A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychological Medicine*, *49*(14), 2330–2341. https://doi.org/10.1017/S003329171800315X

Pedersen, D. H., Mansourvar, M., Sortsø, C., & Schmidt, T. (2019). Predicting Dropouts From an Electronic Health Platform for Lifestyle Interventions: Analysis of Methods and Predictors. *Journal of Medical Internet Research*, *21*(9). https://doi.org/10.2196/13617

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pennebaker, J., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin.

Pepa, L., Sabatelli, A., Ciabattoni, L., Monteriù, A., Lamberti, F., & Morra, L. (2021). Stress Detection in Computer Users From Keyboard and Mouse Dynamics. *IEEE Transactions on Consumer Electronics*, *67*(1), 12–19. IEEE Transactions on Consumer Electronics. https://doi.org/10.1109/TCE.2020.3045228

Perlich, C., Provost, F., & Simonof, J. S. (2004). Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *Journal of Machine Learning Research*.

Piccialli, F., Somma, V. D., Giampaolo, F., Cuomo, S., & Fortino, G. (2021). A survey on deep learning in medicine: Why, how and when? *Information Fusion*, *66*, 111–137. https://doi.org/10.1016/j.inffus.2020.09.006

Piera-Jiménez, J., Etzelmueller, A., Kolovos, S., Folkvord, F., & Lupiáñez-Villanueva, F. (2021). Guided Internet-Based Cognitive Behavioral Therapy for Depression: Implementation Cost-

Effectiveness Study. *Journal of Medical Internet Research*, *23*(5), e27410. https://doi.org/10.2196/27410

Popescu, C., Golden, G., Benrimoh, D., Tanguay-Sela, M., Slowey, D., Lundrigan, E., Williams, J., Desormeau, B., Kardani, D., Perez, T., Rollins, C., Israel, S., Perlman, K., Armstrong, C., Baxter, J., Whitmore, K., Fradette, M.-J., Felcarek-Hope, K., Soufi, G., … Turecki, G. (2021). Evaluating the Clinical Feasibility of an Artificial Intelligence–Powered, Web-Based Clinical Decision Support System for the Treatment of Depression in Adults: Longitudinal Feasibility Study. *JMIR Formative Research*, *5*(10), e31862. https://doi.org/10.2196/31862

Prasad, N., Chien, I., Regan, T., Enrique, A., Palacios, J., Keegan, D., Munir, U., Tanno, R., Richardson, H., Nori, A., Richards, D., Doherty, G., Belgrave, D., & Thieme, A. (2023). Deep learning for the prediction of clinical outcomes in internet-delivered CBT for depression and anxiety. *PLOS ONE*, *18*, e0272685. https://doi.org/10.1371/journal.pone.0272685

Rammstedt, B., Kemper, C., Klein, M., Beierlein, C., & Kovaleva, A. (2012). *Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: Big-Five-Inventory-10 (BFI-10)*.

Ramos, L. A., Blankers, M., van Wingen, G., de Bruijn, T., Pauws, S. C., & Goudriaan, A. E. (2021). Predicting Success of a Digital Self-Help Intervention for Alcohol and Substance Use With Machine Learning. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.734633

Regenthal, J. (2022). *Comparative Analysis of Feature Extraction Methods to Predict Dropout based on Text Data in Digital Mental Health Interventions*. Leuphana University.

Reins, J. A., Boß, L., Lehr, D., Berking, M., & Ebert, D. D. (2019). The more I got, the less I need? Efficacy of Internet-based guided self-help compared to online psychoeducation for major depressive disorder. *Jou. Aff. Diss*, *246*, 695–705.

Reins, J. A., Buntrock, C., Zimmermann, J., Grund, S., Harrer, M., Lehr, D., Baumeister, H., Weisel, K., Domhardt, M., Imamura, K., Kawakami, N., Spek, V., Nobis, S., Snoek, F., Cuijpers, P., Klein, J. P., Moritz, S., & Ebert, D. D. (2020). Efficacy and Moderators of Internet-Based Interventions in Adults with Subthreshold Depression: An Individual Participant Data Meta-Analysis of Randomized Controlled Trials. *Psychotherapy and Psychosomatics*, *90*(2), 94–106. https://doi.org/10.1159/000507819

Richards, D., & Richardson, T. (2012). Computer-based psychological treatments for depression: A systematic review and meta-analysis. *Clinical Psychology Review*, *32*(4), 329–342. https://doi.org/10.1016/j.cpr.2012.02.004

Richter, D., Wall, A., Bruen, A., & Whittington, R. (2019). Is the global prevalence rate of adult mental illness increasing? Systematic review and meta-analysis. *Acta Psychiatrica Scandinavica*, *140*(5), 393–407. https://doi.org/10.1111/acps.13083

Roderick, J. A., & Rubin, D. (2002). *Statistical Analysis with Missing Data* (2nd ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119013563.fmatter

Rodrigo, H., Beukes, E. W., Andersson, G., & Manchaiah, V. (2021). Exploratory Data Mining Techniques (Decision Tree Models) for Examining the Impact of Internet-Based Cognitive Behavioral Therapy for Tinnitus: Machine Learning Approach. *Journal of Medical Internet Research*, *23*(11), e28999. https://doi.org/10.2196/28999

Rodriguez-Galiano, V. F., & Chica-Rivas, M. (2014). Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models. *International Journal of Digital Earth*. https://www.tandfonline.com/doi/abs/10.1080/17538947.2012.748848

Roepke, A. M., Jaffee, S. R., Riffle, O. M., McGonigal, J., Broome, R., & Maxwell, B. (2015). Randomized Controlled Trial of SuperBetter, a Smartphone-Based/Internet-Based Self-Help Tool to Reduce Depressive Symptoms. *Games for Health Journal*, *4*(3), 235–246. https://doi.org/10.1089/g4h.2014.0046

Rohani, D. A., Tuxen, N., Lopategui, A. Q., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2019). Personalizing Mental Health: A Feasibility Study of a Mobile Behavioral Activation Tool for Depressed Patients. *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 282–291. https://doi.org/10.1145/3329189.3329214

Rommel, A., Bretschneider, J., Kroll, L. E., Prütz, F., & Thom, J. (2017). Inanspruchnahme psychiatrischer und psychotherapeutischer Leis- tungen – Individuelle Determinanten und regionale Unterschiede. *Journal of Health Monitoring*, *3–23*.

Rosenberg, M. (1979). Society and the Adolescent Self-Image. In *Society and the Adolescent Self-Image*. Princeton University Press. https://doi.org/10.1515/9781400876136

Rosso, I. M., Killgore, W. D. S., Olson, E. A., Webb, C. A., Fukunaga, R., Auerbach, R. P., Gogel, H., Buchholz, J. L., & Rauch, S. L. (2017). Internet-based cognitive behavior therapy for major depressive disorder: A randomized controlled trial. *Depression and Anxiety*, *34*(3), 236–245. https://doi.org/10.1002/da.22590

Sadasivam, R. S., Borglund, E. M., Adams, R., Marlin, B. M., & Houston, T. K. (2016). Impact of a Collective Intelligence Tailored Messaging System on Smoking Cessation: The Perspect Randomized Experiment. *Journal of Medical Internet Research*, *18*(11), e6465. https://doi.org/10.2196/jmir.6465

Sajjadian, M., Lam, R. W., Milev, R., Rotzinger, S., Frey, B. N., Soares, C. N., Parikh, S. V., Foster, J. A., Turecki, G., Müller, D. J., Strother, S. C., Farzan, F., Kennedy, S. H., & Uher, R. (2021). Machine learning in the prediction of depression treatment outcomes: A systematic review and meta-analysis. *Psychological Medicine*, *51*(16), 2742–2751. https://doi.org/10.1017/S0033291721003871

Sander, J., Moessner, M., & Bauer, S. (2021). Depression, Anxiety and Eating Disorder-Related Impairment: Moderators in Female Adolescents and Young Adults. *International Journal of Environmental Research and Public Health*, *18*(5), Article 5. https://doi.org/10.3390/ijerph18052779

Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Abbafati, C., Adolph, C., Amlag, J. O., Aravkin, A. Y., Bang-Jensen, B. L., Bertolacci, G. J., Bloom, S. S., Castellano, R., Castro, E., Chakrabarti, S., Chattopadhyay, J., Cogen, R. M., Collins, J. K., … Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, *398*(10312), 1700–1712. https://doi.org/10.1016/S0140-6736(21)02143-7

Saseendran, A., Setia, L., Chhabria, V., Chakraborty, D., & Barman Roy, A. (2019). *Impact of Noise in Dataset on Machine Learning Algorithms*. https://doi.org/10.13140/RG.2.2.25669.91369

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. *2011 31st International Conference on Distributed Computing Systems Workshops*, 166–171. https://doi.org/10.1109/ICDCSW.2011.20

Scantamburlo, T., Cortés, A., Foffano, F., Barrué, C., Distefano, V., Pham, L., & Fabris, A. (2023). *Artificial Intelligence across Europe: A Study on Awareness, Attitude and Trust*.

Schapire, R. E. (2013). Explaining AdaBoost. In B. Schölkopf, Z. Luo, & V. Vovk (Eds.), *Empirical Inference* (pp. 37–52). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_5

Schneider, E. E., Schönfelder, S., Domke-Wolf, M., & Wessa, M. (2020). Measuring stress in clinical and nonclinical subjects using a German adaptation of the Perceived Stress Scale. *International Journal of Clinical and Health Psychology: IJCHP*, *20*(2), 173–181. https://doi.org/10.1016/j.ijchp.2020.03.004

Schueller, S., & Mohr, D. (2015). Initial Field Trial of a Coach-Supported Web-Based Depression Treatment. *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*. 9th International Conference on Pervasive Computing Technologies for Healthcare, Istanbul, Turkey. https://doi.org/10.4108/icst.pervasivehealth.2015.260115

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. IEEE Transactions on Signal Processing. https://doi.org/10.1109/78.650093

Scott, I., Carter, S., & Coiera, E. (2021). Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health & Care Informatics*, *28*(1), e100251. https://doi.org/10.1136/bmjhci-2020-100251

Shanahan, T., Tran, T. P., & Taylor, E. C. (2019). Getting to know you: Social media personalization as a means of enhancing brand loyalty and perceived quality. *Journal of Retailing and Consumer Services*, *47*, 57–65. https://doi.org/10.1016/j.jretconser.2018.10.007

Shatte, A., Hutchinson, D., & Teague, S. (2018). *Machine learning in mental health: A systematic scoping review of methods and applications* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/hjrw8

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59 Suppl 20*, 22-33;quiz 34-57.

Shen, J. X., Ma, M. D., Xiang, R., Lu, Q., Vallejos, E. P., XU, G., HUANG, C. R., & LONG, Y. (2020). Dual memory network model for sentiment analysis of review text. *Knowledge-Based Systems*, *188*(105004). https://doi.org/10.1016/j.knosys.2019.105004

Shrestha, A., & Mahmood, A. (2019). Review of Deep Learning Algorithms and Architectures. *IEEE Access*, *7*, 53040–53065. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2912200

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Singer, S., Engesser, D., Wirp, B., Lang, K., Paserat, A., Kobes, J., Porsch, U., Mittag, M., Taylor, K., Gianicolo, E., & Maier, L. (2022). Effects of a statutory reform on waiting times for outpatient psychotherapy: A multicentre cohort study. *Counselling and Psychotherapy Research*, *22*(4), 982–997. https://doi.org/10.1002/capr.12581

Smeden, M. v, Moons, K. G., Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. In *Statistical Methods in Medical Research* (pp. 2455–2474).

Smink, W. A. C., Sools, A. M., Postel, M. G., Tjong Kim Sang, E., Elfrink, A., Libbertz-Mohr, L. B., Veldkamp, B. P., & Westerhof, G. J. (2021). Analysis of the Emails From the Dutch Web-Based Intervention "Alcohol de Baas": Assessment of Early Indications of Drop-Out in an Online Alcohol Abuse Intervention. *Frontiers in Psychiatry*, *12*, 575931. https://doi.org/10.3389/fpsyt.2021.575931

Smith, J., Newby, J. M., Burston, N., Murphy, M. J., Michael, S., Mackenzie, A., Kiln, F., Loughnan, S. A., O'Moore, K. A., Allard, B. J., Williams, A. D., & Andrews, G. (2017). Help from home for depression: A randomised controlled trial comparing internet-delivered cognitive behaviour therapy with bibliotherapy for depression. *Internet Interventions*, *9*, 25–37. https://doi.org/10.1016/j.invent.2017.05.001

Spanhel, K., Balci, S., Feldhahn, F., Bengel, J., Baumeister, H., & Sander, L. B. (2021). Cultural adaptation of internet- and mobile-based interventions for mental disorders: A systematic review. *Npj Digital Medicine*, *4*(1), 128. https://doi.org/10.1038/s41746-021-00498-1

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*.

Spineli, L. M., Pandis, N., & Salanti, G. (2015). Reporting and handling missing outcome data in mental health: A systematic review of Cochrane systematic reviews and meta-analyses. *Research Synthesis Methods*, *6*(2), 175–187. https://doi.org/10.1002/jrsm.1131

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry: A survey of current progress in depression detection, diagnosis and treatment. *Brain Informatics*, *10*(1), 10. https://doi.org/10.1186/s40708-023-00188-6

Sun, X., & Xu, W. (2014). Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters*, *21*(11), 1389–1393. IEEE Signal Processing Letters. https://doi.org/10.1109/LSP.2014.2337313

Sundar, S. S., & Marathe, S. S. (2010). Personalization versus Customization: The Importance of Agency, Privacy, and Power Usage. *Human Communication Research*, *36*(3), 298–322. https://doi.org/10.1111/j.1468-2958.2010.01377.x

Surprenant, C. F., & Solomon, M. R. (1987). Predictability and Personalization in the Service Encounter. *Journal of Marketing*, *51*(2), 86–96. https://doi.org/10.1177/002224298705100207

Svanborg, P., & Asberg, M. (1994). A new self-rating scale for depression and anxiety states based on the Comprehensive Psychopathological Rating Scale. *Acta Psychiatrica Scandinavica*, *89*(1), 21–28. https://doi.org/10.1111/j.1600-0447.1994.tb01480.x

Swift, J. K., Callahan, J. L., Cooper, M., & Parkin, S. R. (2018). The impact of accommodating client preference in psychotherapy: A meta-analysis. *Journal of Clinical Psychology*, *74*(11), 1924–1937. https://doi.org/10.1002/jclp.22680

Symons, M., Feeney, G. F. X., Gallagher, M. R., Young, R. McD., & Connor, J. P. (2019). Machine learning vs addiction therapists: A pilot study predicting alcohol dependence treatment outcome from patient data in behavior therapy with adjunctive medication. *Journal of Substance Abuse Treatment*, *99*, 156–162. https://doi.org/10.1016/j.jsat.2019.01.020

Ta Park, V. M., Ton, V., Yeo, G., Tiet, Q. Q., Vuong, Q., & Gallagher-Thompson, D. (2019). Vietnamese American Dementia Caregivers' Perceptions and Experiences of a Culturally Tailored, Evidence-Based Program to Reduce Stress and Depression. *Journal of Gerontological Nursing*, *45*(9), 39–50. https://doi.org/10.3928/00989134-20190813-05

Tal, A., & Torous, J. (2017). The digital mental health revolution: Opportunities and risks. *Psychiatric Rehabilitation Journal*, *40*(3), 263–265. https://doi.org/10.1037/prj0000285

Teepe, G. W., Da Fonseca, A., Kleim, B., Jacobson, N. C., Salamanca Sanabria, A., Tudor Car, L., Fleisch, E., & Kowatsch, T. (2021). Just-in-Time Adaptive Mechanisms of Popular Mobile Apps for Individuals With Depression: Systematic App Search and Literature Review. *Journal of Medical Internet Research*, *23*(9), e29412. https://doi.org/10.2196/29412

Terhorst, Y., Knauer, J., Philippi, P., & Baumeister, H. (2023). The Relation between passively collected GPS mobility metrics and depressive symptoms: A systematic review and meta-analysis. (Preprint). *Journal of Medical Internet Research*. https://doi.org/10.2196/51875

The Lancet Global Health. (2020, November). *Mental health matters*. https://doi.org/10.1016/S2214-109X(20)30432-0

Titov, N., Dear, B. F., Staples, L. G., Terides, M. D., Karin, E., Sheehan, J., Johnston, L., Gandy, M., Fogliati, V. J., Wootton, B. M., & McEvoy, P. M. (2015). Disorder-specific versus transdiagnostic and clinician-guided versus self-guided treatment for major depressive disorder and comorbid anxiety disorders: A randomized controlled trial. *Journal of Anxiety Disorders*, *35*, 88–102. https://doi.org/10.1016/j.janxdis.2015.08.002

Titov, N., Dear, B., Nielssen, O., Staples, L., Hadjistavropoulos, H., Nugent, M., Adlam, K., Nordgreen, T., Bruvik, K. H., Hovland, A., Repål, A., Mathiasen, K., Kraepelien, M., Blom, K., Svanborg, C., Lindefors, N., & Kaldo, V. (2018). ICBT in routine care: A descriptive analysis of successful clinics in five countries. *Internet Interventions*, *13*, 108–115. https://doi.org/10.1016/j.invent.2018.07.006

Tomitaka, S., & Furukawa, T. A. (2021). The GAD-7 and the PHQ-8 exhibit the same mathematical pattern of item responses in the general population: Analysis of data from the National Health Interview Survey. *BMC Psychology*, *9*, 149. https://doi.org/10.1186/s40359-021-00657-9

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). *What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use* (arXiv:1905.05134). arXiv. https://doi.org/10.48550/arXiv.1905.05134

Torous, J., Lipschitz, J., Ng, M., & Firth, J. (2020). Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis. *Journal of Affective Disorders*, *263*, 413–419. https://doi.org/10.1016/j.jad.2019.11.167

Triantafyllidis, A. K., & Tsanas, A. (2019). Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature. *Journal of Medical Internet Research*, *21*(4), e12286. https://doi.org/10.2196/12286

Tsang, E. W. K. (2014). Case studies and generalization in information systems research: A critical realist perspective. *The Journal of Strategic Information Systems*, *23*(2), 174–186. https://doi.org/10.1016/j.jsis.2013.09.002

Tylka, T. L. (2006). Development and psychometric evaluation of a measure of intuitive eating. *Journal of Counseling Psychology*, *53*(2), 226–240. https://doi.org/10.1037/0022-0167.53.2.226

Tzafilkou, K., & Protogeros, N. (2018). Mouse Behavioral Patterns and Keystroke Dynamics in End-User Development: What can they tell us? In *Computers in Human Behavior* (Vol. 83, pp. 288–305).

Udegbe, F. C., Ebulue, O. R., Ebulue, C. C., & Ekesiobi, C. S. (2024). AI's impact on personalized medicine: Tailoring treatments for improved health outcomes. *Engineering Science & Technology Journal*, *5*(4), 1386–1394.

Vachon, F., & Tremblay, S. (2019). What Eye Tracking Can Reveal About Dynamic Decision Making. In *Advances in Cognitive Engineering and Neuroergonomics*. AHFE International (USA).

van Berkel, N., Goncalves, J., Hosio, S., Sarsenbayeva, Z., Velloso, E., & Kostakos, V. (2020). Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies*, *134*, 1–12. https://doi.org/10.1016/j.ijhcs.2019.10.003

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, *28*(8), 2455–2474. https://doi.org/10.1177/0962280218784726

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2). https://doi.org/10.1145/3476068

Vieira, S., Liang, X., Guiomar, R., & Mechelli, A. (2022). Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clinical Psychology Review*, *97*, 102193. https://doi.org/10.1016/j.cpr.2022.102193

Vilone, G., & Longo, L. (2021a). Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction*, *3*, 615. https://doi.org/10.3390/make3030032

Vilone, G., & Longo, L. (2021b). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106.

von Glischinski, M., Willutzki, U., Stangier, U., Hiller, W., Hoyer, J., Leibing, E., Leichsenring, F., & Hirschfeld, G. (2018). Liebowitz Social Anxiety Scale (LSAS): Optimal cut points for remission and response in a German sample. *Clinical Psychology & Psychotherapy*, *25*(3), 465–473. https://doi.org/10.1002/cpp.2179

Wahle, F., Kowatsch, T., Fleisch, E., Rufer, M., & Weidt, S. (2016). Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. *JMIR mHealth and uHealth*, *4*(3), e111. https://doi.org/10.2196/mhealth.5960

Wallert, J., Boberg, J., Kaldo, V., Mataix-Cols, D., Flygare, O., Crowley, J. J., Halvorsen, M., Ben Abdesslem, F., Boman, M., Andersson, E., Hentati Isacsson, N., Ivanova, E., & Rück, C. (2022). Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data. *Translational Psychiatry*, *12*(1), Article 1. https://doi.org/10.1038/s41398-022-02133-3

Wallert, J., Gustafson, E., Held, C., Madison, G., Norlund, F., von Essen, L., & Olsson, E. M. G. (2018). Predicting Adherence to Internet-Delivered Psychotherapy for Symptoms of Depression and Anxiety After Myocardial Infarction: Machine Learning Insights From the U-CARE Heart Randomized Controlled Trial. *Journal of Medical Internet Research*, *20*(10), e10754. https://doi.org/10.2196/10754

Wang, P. S., Demler, O., & Kessler, R. C. (2002). Adequacy of Treatment for Serious Mental Illness in the United States. *American Journal of Public Health*, *92*(1), 92–98. https://doi.org/10.2105/AJPH.92.1.92

Wang, P. S., Lane, M., Olfson, M., Pincus, H. A., Wells, K. B., & Kessler, R. C. (2005). Twelve-Month Use of Mental Health Services in the United States: Results From the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 629. https://doi.org/10.1001/archpsyc.62.6.629

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: Beyond the black box. *BMJ*, l886. https://doi.org/10.1136/bmj.l886

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association: JAMIA*, *20*(1), 144–151. https://doi.org/10.1136/amiajnl-2011-000681

W.H.O. (2022). *World Mental Health Report—Transforming mental health for all*. World Health Organization.

Wichmann, T., & DeLong, T. (2014). The Basal Ganglia. In *Principles of Neural Science* (Fifth Edition). McGraw-Hill Education. neurology.mhmedical.com/content.aspx?aid=1101682080

Wołk, A., Chlasta, K., & Holas, P. (2021). *Hybrid approach to detecting symptoms of depression in social media entries.*

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2021). Deep Learning With Anaphora Resolution for the Detection of Tweeters With Depression: Algorithm Development and Validation Study. *JMIR Mental Health*, *8*(8), e19824. https://doi.org/10.2196/19824

Wu, M. S., Chen, S.-Y., Wickham, R. E., Leykin, Y., Varra, A., Chen, C., & Lungu, A. (2022). Predicting non-initiation of care and dropout in a blended care CBT intervention: Impact of early digital engagement, sociodemographic, and clinical factors. *DIGITAL HEALTH*, *8*. https://doi.org/10.1177/20552076221133760

Xie, Y., Chen, M., Kao, D., Gao, G., & Chen, X. "Anthony." (2020). CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376807

Yamauchi, T. (2013). Mouse Trajectories and State Anxiety: Feature Selection with Random Forest. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 399–404. https://doi.org/10.1109/ACII.2013.72

Yamauchi, T., & Xiao., K. (2018). Reading emotion from mouse cursor motions: A effective computing approach. *Cognitive Science*, *42*(3), 711–819. https://doi.org/10.1111/cogs.12557

Yamauchi, T., & Xiao, K. (2018). *Reading Emotion From Mouse Cursor Motions: Affective Computing Approach*. https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.12557

Yang, C. C. (2022). Explainable Artificial Intelligence for Predictive Modeling in Healthcare. *Journal of Healthcare Informatics Research*, *6*(2), 228–239. https://doi.org/10.1007/s41666-022-00114-1

Yang, R., & Wibowo, S. (2022). User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets*, *32*(4), 2053–2077.

Yang, W., Wang, M., Zou, S., Peng, J., & Xu, G. (2021). An Implicit Identity Authentication Method Based on Deep Connected Attention CNN for Wild Environment. *Proceedings of the 2021 9th International Conference on Communications and Broadband Networking*, 94–100. https://doi.org/10.1145/3456415.3457222

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *XLNet: Generalized Autoregressive Pretraining for Language Understanding* (arXiv:1906.08237). arXiv. https://doi.org/10.48550/arXiv.1906.08237

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yeruva, V. K., Junaid, S., & Lee, Y. (2019). Contextual Word Embeddings and Topic Modeling in Healthy Dieting and Obesity. *Journal of Healthcare Informatics Research*, *3*(2), 159–183. https://doi.org/10.1007/s41666-019-00052-5

Yin, J., Ngiam, K. Y., & Teo, H. H. (2021). Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. *Journal of Medical Internet Research*, *23*(4), e25759. https://doi.org/10.2196/25759

Yusupova, O. (2021). *Mathematical Model For Defining Academic Interests Of Students In Personalized Education* (p. 427). https://doi.org/10.15405/epsbs.2021.12.02.52

Zagorscak, P., Heinrich, M., Bohn, J., Stein, J., & Knaevelsrud, C. (2020). How individuals change during internet-based interventions for depression: A randomized controlled trial comparing standardized and individualized feedback. *Brain and Behavior*, *10*(1), e01484. https://doi.org/10.1002/brb3.1484

Zagorscak, P., Heinrich, M., Sommer, D., Wagner, B., & Knaevelsrud, C. (2018). Benefits of Individualized Feedback in Internet-Based Interventions for Depression: A Randomized Controlled Trial. *Psychotherapy and Psychosomatics*, *87*(1), 32–45. https://doi.org/10.1159/000481515

Zajac, H., Li, D., Dai, X., Carlsen, J., Kensing, F., & Andersen, T. (2023). Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Transactions on Computer-Human Interaction*, *30*. https://doi.org/10.1145/3582430

Zantvoort, K., Hentati Isacsson, N., Funk, B., & Kaldo, V. (2024). Data set size vs homogeneity – A Machine Learning study on pooling intervention data in E-Mental Health dropout predictions. *SAGE Digital Health*, *10*, 1–11. https://doi.org/DOI: 10.1177/20552076241248920

Zantvoort, K., Nacke, B., Görlich, D., Hornstein, S., Jacobi, C., & Funk, B. (2024). *Predictive Power, Variance and Generalizability – A Machine Learning Case Study on Minimal Necessary Data*

Sets Sizes in Digital Mental Health Intervention Predictions. https://doi.org/10.21203/rs.3.rs-4616728/v1

Zantvoort, K., Scharfenberger, J., Boß, L., Lehr, D., & Funk, B. (2023). Finding the Best Match—A Case Study on the (Text-)Feature and Model Choice in Digital Mental Health Interventions. *Journal of Healthcare Informatics Research*, *7*(4), 447–479. https://doi.org/10.1007/s41666-023-00148-z

Zhang, H. (2004). The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, *2*.

Zhang, Y., Zhou, Y., & Yao, J. (2020). Feature Extraction with TF-IDF and Game-Theoretic Shadowed Sets. *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, *1237*, 722–733. https://doi.org/10.1007/978-3-030-50146-4_53

Zheng, L., Long, M., Zhong, L., & Gyasi, J. F. (2022). The effectiveness of technology-facilitated personalized learning on learning achievements and learning perceptions: A meta-analysis. *Education and Information Technologies*, *27*(8), 11807–11830. https://doi.org/10.1007/s10639-022-11092-7

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593.

Zimmermann, P., Guttormsen, S., Danuser, B., & Gomez, P. (2003). Affective Computing—A Rationale for Measuring Mood With Mouse and Keyboard. *International Journal of Occupational Safety and Ergonomics : JOSE*, *9*, 539–551. https://doi.org/10.1080/10803548.2003.11076589

Appendix

Paper	N	Features	Models	Task	Results
Wallert et al. (2018)	90	25 baseline and 7 linguistic	Random Forrest	First homework submission predicts if three or more will be submitted	Accuracy = 0.64 (52% dropout)
Pedersen, Mansourvar, Sortsø, & Schmidt (2019)	2,684	>11 baseline and user behaviour	Logistic regression, Decision Trees, and Random Forest	Inactivity for at least four weeks within 12 months	AUC = 0.92
Bremer et al. (2020)	101-151	83-401 baseline and intervention behaviour	Boosted decision trees	Data from module 1-6 predicts completed 7/7 modules	AUC = 0.6-0.9 in dependence of time used to gather data
Smink et al. (2021)	770	12 baseline and 2 types of linguistic	Logistic regression, XGBoost, Mixed Effect Random Forest, Multi-Layer Perceptron Model	Data from first four emails to predict therapist's dropout decision	Accuracy = 0.58-0.61 (55% dropout)
Moshe et al. (2021)	253	8 baseline and 4 intervention behaviour	Logistic regression	completing <6/9 modules	AUC = 0.72
Kim et al. (2021)	45	40 baseline	Linear regression	Number of log ins	$R^2 = 0.42$
Côté-Allard et al. (2022)	342	2 intervention behaviour x 7-42 days (sequential form)	Self-Attention Network	7-42 days 8 logins or more over a period of at least 56 days	Balanced accuracy = 64-79% for 7-42 days
Linardon, Fuller-Tyszkiewicz, Shatte, & Greenwood (2022)	Baseline: 826 Interven. Behav.: 340	36 baseline and 110 intervention behaviour	Linear regression, Support Vector Machines, K-Nearest Neighbor, CART Decision Trees, and Random Forest	accessing less than half of the available material	AUC = 0.48-0.52 for questionnaire data only, AUC = 0.62-0.93 for behaviour data
Gonzalez Salas Duhne et al. (2022)	1,611	36 baseline features	Logistic regression	three or fewer sessions of treatment	AUC = 0.57
Bricker et al. (2023)	4,301	23 baseline and 7 intervention behaviour features	Logistic regression, Decision Tree, Support Vector Machine, and undefined neural network	First 7 days of data to predict stopping usage after 7 days	AUC = 0.6-0.94
Günther et al. (2023)	22,796	62 baseline features	XGBoost	Data at baseline predicts completion of 8 sessions	AUC = 0.57
Linnet et al. (2023)	164	24 baseline and 4 intervention behaviour features	Logistic regression	Time between sessions 1-4 predicts completing 12/12 modules	Balanced accuracy 0.61-0.73 for session 1 to 4

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01360-w

Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions

Kirsten Zantvoort[1] ✉, Barbara Nacke[2], Dennis Görlich[3], Silvan Hornstein[4], Corinna Jacobi[2] & Burkhardt Funk[1]

Artificial intelligence promises to revolutionize mental health care, but small dataset sizes and lack of robust methods raise concerns about result generalizability. To provide insights on minimal necessary data set sizes, we explore domain-specific learning curves for digital intervention dropout predictions based on 3654 users from a single study (ISRCTN13716228, 26/02/2016). Prediction performance is analyzed based on dataset size (N = 100–3654), feature groups (F = 2–129), and algorithm choice (from Naive Bayes to Neural Networks). The results substantiate the concern that small datasets ($N \leq 300$) overestimate predictive power. For uninformative feature groups, in-sample prediction performance was negatively correlated with dataset size. Sophisticated models overfitted in small datasets but maximized holdout test results in larger datasets. While N = 500 mitigated overfitting, performance did not converge until N = 750–1500. Consequently, we propose minimum dataset sizes of N = 500–1000. As such, this study offers an empirical reference for researchers designing or interpreting AI studies on Digital Mental Health Intervention data.

The rapid advancement of artificial intelligence (AI) in various industries has spurred great anticipation for its transformative power in health care[1,2]. One area that particularly stands to benefit from AI-based improvements is mental health[3,4]. With 16% of global disability-adjusted life years attributed to them and staggering economic costs, mental disorders are immensely burdensome for individuals and societies alike[5]. Further, mental disorders are heterogeneous in their treatment needs, and AI promises a resource-efficient way to personalize, scale and improve mental health care[4,6–8]. However, among the central challenges in realizing AI's envisioned potential within mental health interventions (MHIs) is the limitation of data set sizes[4,6,8–10].

In contrast to diagnostics or public health data[3], median data set sizes of machine learning (ML) application studies with MHI data barely exceed 100–150 patients[4,8,9,11]. Digital mental health interventions (DMHIs) allow for an easier collection of datasets than face-to-face (f2f) therapy[7,12], but median data set sizes are still only 155–350[7,13,14]. This is problematic because prediction power is notoriously known to be overestimated in such small data set sizes[15–17].

Sajjadian et al.[9] found that MHI studies with small data set sizes reported significantly higher performance metrics than methodologically sound studies ($p = 0.005$). Further, they reported that 71% of the 59 investigated studies lacked an appropriate validation method and instead reported single test set or cross-validation (CV) results. Zantvoort et al.[13] reported that DMHI dropout prediction models trained on small data sets produced the highest CV results but performed worst on the larger test set. As a result, several authors caution the interpretation of the current state of results and warn about possible consequences. Deploying an ungeneralizable model risks suboptimal care, deteriorating patient outcomes, wasted resources, and, thus, ultimately leads to the opposite of the intended effects[6,9,13,18,19].

Despite their undebatable relevance, minimal necessary sample sizes, as they are standard in classical statistical settings, are uncommon in ML applications[20]. While no all-encompassing solution is available, a key approach for better understanding them are learning curves[20–22]. A recent study by Giesemann et al.[21] produced such learning curves for dropout predictions in f2f psychotherapy and suggested 300 data points as a minimal necessary sample size. However, they only used eight patient-reported features and did not investigate overfitting or result variance. Further, only minimal insights are available into the interaction effect of sample sizes, model types and the number and type of features in DMHI data. Flexible models approximate realities' complexity well, however, they risk overfitting, especially on small

[1]Institute of Information Systems, Leuphana University, Lüneburg, Germany. [2]Department of Clinical Psychology and Psychotherapy, Faculty of Psychology, Technische Universität Dresden, Dresden, Germany. [3]Institute of Biostatistics and Clinical Research, University Münster, Münster, Germany. [4]Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany. ✉e-mail: kirsten.zantvoort@leuphana.de

data sets[10,13,23]. Simple models tend to produce more stable results but risk disregarding valuable information[24–26].

Additionally, the effectiveness of any model significantly depends on the nature and number of predictors[22,24]. Especially for DMHIs, feature numbers can quickly grow into hundreds of variables[12,27]. At the same time, data protection and adherence concerns call for a data minimalism approach[12,28,29]. Moreover, several papers have reported that fewer hand-crafted variables improved their results[12,30,31].

In conclusion, the key questions repeatedly arising in ML studies in DMHIs are (1) how the dataset size influences the results[9,12,21], (2) which of the ample algorithms to implement[21,26–28,30,32], and (3) which of the abundant possible variables to use[12,27,30]. The current study aims to investigate the interdependence of these questions by analyzing the learning curves for dropout predictions across (1) six models with varying levels of flexibility and (2) six feature groups differing in their predictive power and extent. Beyond test set performance levels, the results will be investigated regarding their variance, generalizability from the training to test set, and convergence trajectory to derive insights into minimal necessary data set sizes. To this end, we leverage 3,654 users' data from digital eating disorder prevention interventions provided to the general public in Germany[33]. Eating disorders are highly prevalent[34] and associated with immense levels of suffering[35]. While DMHIs are effective in preventing and treating EDs, intervention dropout is a substantial issue among them[36]. Measures such as guidance can mitigate dropout but are costly[37,38]. Using AI to identify users at risk of dropping out allows for optimizing resource allocation and improving outcomes regardless of the availability of final symptom scores[30,37,38]. As such, within the limits of a single-dataset case study, this paper seeks to provide insights to improve the design and interpretation of ML studies on DMHI data.

Results

Final Values

The final data set comprised 3654 users, of whom 63% were classified as dropouts. Feature groups ranged from 2 features (F) (simple questionnaire), over 7 (simple behavior), 13 (selected behavior), 51 (extended questionnaire), and 64 (mixed) to a maximum of 129 features (extended behavior) in addition to the intervention information. The descriptive statistics, including for the training and test set, can be found in Supplementary Table 1.

Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machines, (SVM) Random Forest (RF), adaBoost and shallow Multilayer Perceptron Neural Network (NN) models were trained with 10-fold CV on 80% of the assumed data set sizes between 100 and 3,654 users and evaluated on the test set of 731 users. Hyperparameters differed across settings (e.g., regularization for 7 vs 129 features), and are published in this study's GitHub repository. All detailed result metrics are published in Supplementary Table 2. Supplementary Table 3 holds the p-values for the DeLong tests.

Predictive Power of Feature Groups

Approximating the prediction performance in terms of the area under the curve score (AUC) via the best model on $N = 3{,}654$, the assumed predictive power across feature types was confirmed. As shown in Fig. 1a, b, there was no information in the simple (0.53 test AUC) and only moderate (0.66 AUC) in the extended questionnaire data. The simple behavior data (Fig. 1c) already achieved an AUC of 0.72, which was increased to 0.77 for the extended (Fig. 1d) and 0.80 for the selected (Fig. 1d) behavior data. From there, the mixed features (Fig. 1f) only slightly increased results to 0.81 AUC. Since the simple questionnaire data had no predictive power, its results will only be discussed in the context of overfitting.

Overfitting on Small Data Set Sizes

Overfitting was a substantial problem for the small data set sizes ($N \leq 300$), such that the CV results exceeded the test results by up to 0.12 in AUC (*on average* 0.05, see Supplementary Table 2). With increasing data set sizes ($N \geq 500$), overfitting was substantially reduced for all features (mean 0.02, max. 0.06 AUC), except the simple questionnaire (mean 0.05, max. 0.07 AUC). Both the extent of overfitting on small data set sizes and its reduction with increased data set sizes varied across the (1) feature types and numbers and (2) model types.

Firstly, in terms of feature types, low-information feature groups (simple and extended questionnaire, Fig. 1a, b) were the most likely to overfit. For data set sizes of $N \leq 300$, their avg. difference between the training and test scores without NB was −0.07 (max. −0.12 AUC). Choosing the winning model based on CV scores for the simple questionnaire data led to up to 70% of the results being >0.61 AUC despite a useless model (Table 1). Further, for these two feature types, up to $N = 300$ training results got worse with increasing data set size (avg. −0.03, max. −0.06 AUC) as seen in Fig. 1a, b. The same was visible in the simple behavioral data (Fig. 1c) but less severe and only for RF and SVM (avg. and max. −0.02 AUC for $N \leq 500$).

For the extended behavior, selected behavior and mixed data, gaps between training and test set performance for $N \leq 300$ were also prevalent but less severe (avg. −0.05, max. −0.09 AUC). For these three most informative feature groups, both training and test results increased with data set sizes (Fig. 1d–f), and the models winning in the training scores consistently also produced the highest test scores. Hence, the extent of overfitting in the results decreased as the information value of the features increased.

In terms of the number of features, the very small groups (simple questionnaire with 2, and simple behavior with 7 features) overfitted slightly more than their larger counterparts (14, 51 and 129 features). However, this effect was slightly reversed when increasing from selected behavior (13 features, mean 0.04 AUC, max. 0.06) to extended behavior (129 features, mean 0.05, max. 0.09) or mixed features (64 features, mean 0.05, max. 0.09).

Secondly, regarding model types, simpler models were less likely to overfit. As reported in Table 1, at $N = 100$, the share of CV results with at least +0.10 higher AUC than the test results was by far the lowest for NB (avg. 13%). On the other end of the spectrum, the tree-based models overestimated mode performance by at least +0.10 AUC in 42% (adaBoost) and 45% (RF) of the cases. However, across all models, these shares dropped substantially (avg. 7-8%, Table 1) for $N = 300$ and to mostly 0% by $N = 500$ (Table 1). Thus, the effect that more sophisticated models overfit more than simple models diminished with increasing data set size.

Variance of Results

As shown in Fig. 2, the prediction results of the individual validation folds were highly unstable for small data set sizes. The AUCs' standard deviation (S.D.) averaged across runs was by far the highest for $N = 100$ at 0.20 AUC. As such, the variability of AUC results spanned across a large part of the AUCs scale of 0-1, with the expectation to be between 0.5 (no information value) and 1 (perfect score). This variability steeply declined as the data set size increased as it had already halved by N = 400 (S.D. 0.10, Fig. 2). After that, it continued to drop, with the lowest average value in our results being S.D. 0.03 AUC at $N = 3{,}654$. As such, one can expect stable, thus similar, results for repeated calculations on large data, however, the results can largely differ when using small data sets.

Parallel to the observations in overfitting, the result variance was highest for the unpredictive feature groups. The single validation folds of $N = 100$ in the simple questionnaire data covered the entire AUC score range from very bad to very good (AUC mean 0.60, ±S.D. 0.37–0.83, min. 0.00, max. 1.00). Variance was lowest but still very high for the selected behavior data (AUC mean 0.70, ±S.D. 0.52–0.94, min. 0.10, max. 1.00).

Performance Convergence per Model

The convergence points of the test set performance differed per model type and are shown in Fig. 3. The simple questionnaire results are shown in the graphs but ignored in the calculations as there was no predictive power to converge towards.

The simpler models NB, LR and SVMs (Fig. 3a–c) all had a median convergence point of $N = 750$. The more sophisticated tree-based models converged later at $N = 1{,}000$ (Fig. 3d, e), followed by the NN at $N = 1500$

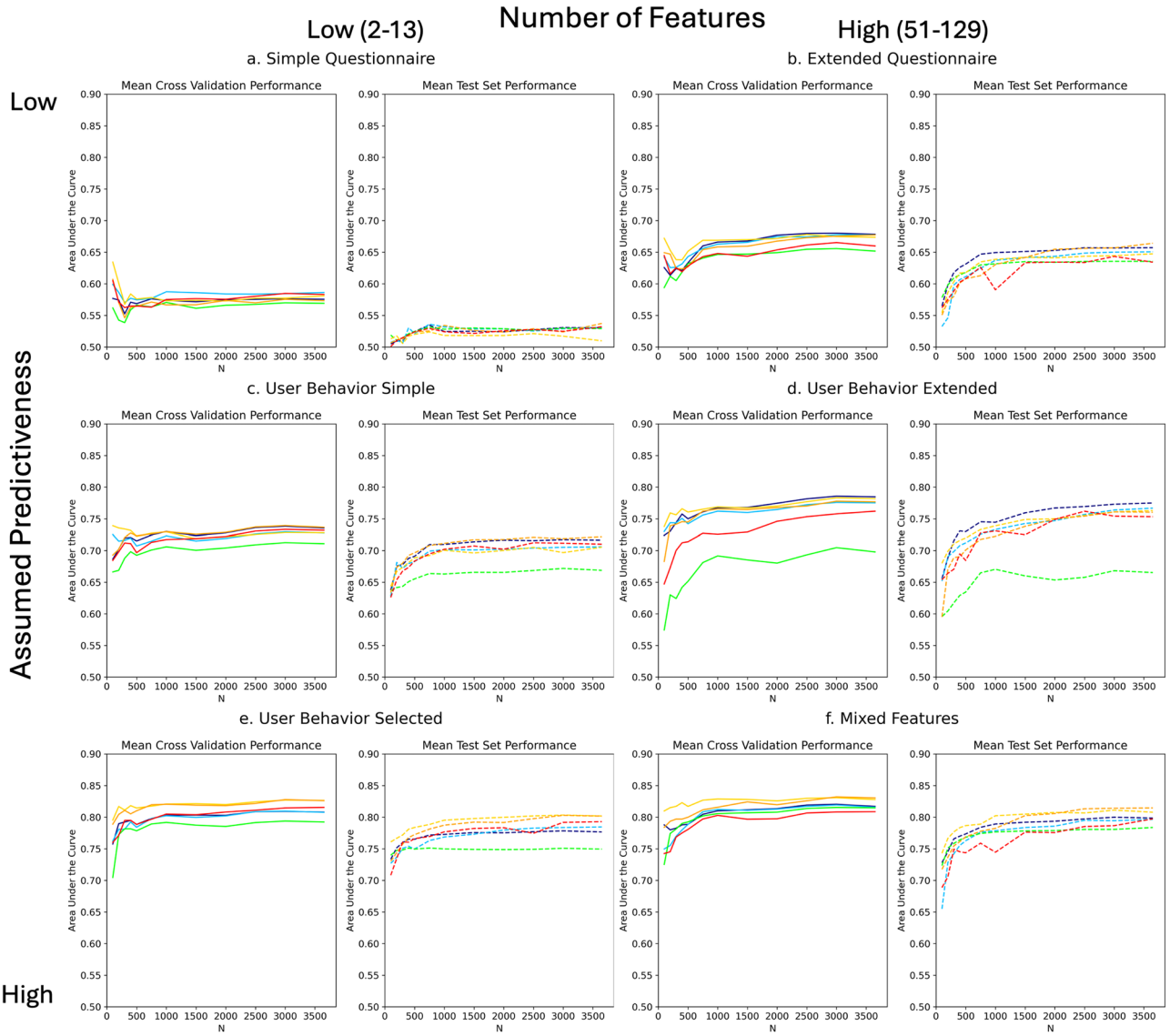


Fig. 1 | Training and test learning curves per feature type. Learning curves for (**a**) simple questionnaire, (**b**) extended questionnaire, (**c**) simple user behavior, (**d**) extended user behavior, (**e**) selected user behavior, and (**f**) mixed features. Each panels shows the respective mean AUC score for the Cross-Validation on the training data (solid line) on the respective left and mean test data performance (dotted line) on the right side. The colors of the lines represent the model types, i.e., Logistic Regression (dark blue), Support Vector Machines (light blue), Naïve Bayes (green), Random Forest (yellow), adaBoost (orange) and Neural Network (red).

(Fig. 3f). The NB (Fig. 3a) had no performance improvement (+0% AUC) when provided with large data set sizes ($N = 3{,}654$ instead of 750), whereas LR (Fig. 3b) and RF (Fig. 3d), on average, grew +2%. SVMs (Fig. 3c) and NNs (Fig. 3f) could slightly better leverage the largest data set (+3%) but were surpassed by adaBoost (Fig. 3e) on average increasing the AUC between $N = 750$ and 3,654 by +5%.

It is noticeable that the NN (Fig. 3f) showed oscillation and larger variability in the results for much longer than the other models, where this only occurred for very small data sets. Training it on the small data sets partly (< 20% of runs) gave convergence warnings.

Marginal Value and Convergence of Additional Features

The marginal benefit of complex features was highest for large data set sizes, and more predictive feature groups tended to converge at higher data set sizes (Fig. 3).

Adding the extended questionnaire features to the simple ones (F = 51, dark and light blue lines in Fig. 3) continuously improved results as the data set size grew (avg. 0.51–0.53 versus 0.55–0.66 test AUC for $N = 100$–3,654). The same was the case for increasing the simple to the extended behavior data (avg. 0.63–0.70 versus 0.64–0.75 test AUC, turquoise and light green lines in Fig. 3). Due to overfitting on $N = 100$, the best CV results for the simple behavior were equal to those of the extended behavior data (AUC = 0.74), despite being lower on the test set (0.64 vs 0.68 AUC), as shown in Fig. 4. This effect faded with increasing data set size, and at N = 500, even the test set performance of the extended group surpassed the simple one's CV scores.

Similarly, using selected instead of extended behavioral data was most beneficial on the small data sets (+0.08–0.03 test AUC difference at N = 100–3,654, Fig. 1e, d). Generally, for all models but the NB, the extended behavior data curve (light green in Fig. 3) was the steepest after $N = 1000$, such that it was closing the gap to the selected behavior features. For LR (Fig. 3b), it even had already matched the selected behavioral data's performance at $N = 3654$.

Adding more than 50 questionnaire features to the selected behavior data for the mixed data set (yellow in Fig. 3) first led to slightly less ($N \leq 200$, avg. difference in test AUC −0.02), then equal ($N = 300$–500, 0.00), and ultimately slightly better performance ($N > 500$, +0.01). As the only exception, using selected (F = 13) instead of simple (F = 7) behavioral data

Table 1 | Overfitting as share of training Cross-Validation results (in %) that are at least +0.10 AUC higher than the respective test results per model and feature type

	LR	SVM	NB	RF	adaBoost	NN
***N* = 100**						
Simple Questionnaire	0.40	0.40	0.30	0.70	0.60	0.40
Extended Questionnaire	0.50	0.60	0.10	0.60	0.50	0.50
Simple Behavior	0.40	0.40	0.30	0.50	0.40	0.20
Extended Behavior	0.30	0.30	0.00	0.50	0.50	0.10
Selected Behavior	0.20	0.10	0.00	0.10	0.30	0.10
Mixed Features	0.20	0.50	0.10	0.30	0.20	0.20
***N* = 300**						
Simple Questionnaire	0.20	0.20	0.20	0.20	0.10	0.20
Extended Questionnaire	0.10	0.10	0.00	0.10	0.10	0.20
Simple Behavior	0.10	0.10	0.20	0.20	0.10	0.10
Extended Behavior	0.10	0.00	0.00	0.10	0.10	0.00
Selected Behavior	0.00	0.00	0.00	0.00	0.00	0.00
Mixed Features	0.00	0.00	0.00	0.00	0.00	0.00
***N* = 500**						
Simple Questionnaire	0.10	0.10	0.10	0.10	0.10	0.10
Extended Questionnaire	0.00	0.00	0.00	0.00	0.00	0.00
Simple Behavior	0.00	0.00	0.00	0.00	0.00	0.00
Extended Behavior	0.00	0.00	0.10	0.00	0.00	0.00
Selected Behavior	0.00	0.00	0.00	0.00	0.00	0.00

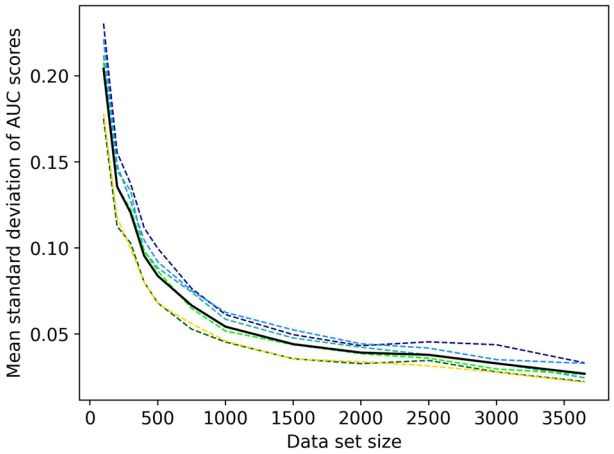


Fig. 2 | Cross-validation result variance per feature type. Mean standard deviation of the single folds' area under the curve score as dotted lines in different colors per feature type i.e., simple baseline (dark blue), extended baseline (light blue), simple behavior (turquoise), extended behavior (light green), selected behavior (dark green), mixed features (yellow). Mean across all features in black solid line.

was always beneficial, but most so on the small data sets (avg. +0.12–0.08 test AUC difference for $N = 100$–3,654).

Model and Feature Combinations

Naive Bayes (NB, green in Fig. 1) obtained competitive test results (top3 models) for smaller data set sizes, specifically for the extended questionnaire ($N \leq 750$), mixed features ($N \leq 400$), selected behavior ($N \leq 200$), and simple behavior ($N = 100$). However, NB never outperformed the respective other top3 models ($p > 0.05$). Furthermore, as shown in Fig. 1c–e, NB significantly underperformed compared to the other models for behavior data, particularly for extended features and larger data set sizes ($p < 0.05$).

Logistic Regression (LR, dark blue in Fig. 1), on the other hand, performed very well in almost all settings. It consistently outperformed most models for the extended questionnaire data for $N = 200$–500 ($p < 0.05$). For $N > 500$, LR continued performing well but was first matched by RF and later ($N > 2500$) by adaBoost. In the extended behavior data, LR was below or equal to RF for $N \leq 200$ but significantly outperformed all models ($p < 0.05$) with few exceptions after that.

Support Vector Machines (SVMs, light blue in Fig. 1) mainly performed in the mid-field but were most competitive with a linear kernel in the two extended feature types. As such, they performed similarly to the top model LR on extended behavior data for $N > 2500$ ($p = 0.06$–0.08) and regularly outperformed ($p < 0.05$) NB, NN and adaBoost.

Similarly to LR, Random Forest models (RFs, yellow in Fig. 1) performed very well, especially for the highest information feature types. They consistently outperformed all models for selected behavior and mixed features, with the only regular exception being adaBoost for $N > 750$ in selected behavior and $N > 1000$ in mixed features.

adaBoost (orange in Fig. 1) tended to perform better with larger data set sizes. For the highest information features, it progressively caught up to RF as of $N > 400$. Additionally, adaBoost performed very well in the simple behavior data ($N > 100$) and the extended questionnaire data ($N > 1500$).

Multilayer Perceptron Neural Networks (NN, red in Fig. 1) were among the top3 models for simple behavior ($N > 750$) and selected behavior ($N > 200$) data and occasionally performed well for extended behavior data. NN's most competitive results were for data set sizes of 1500 or more, where it was most likely to outperform NB, LR, or SVMs.

Recall-Precision Tradeoff

While the detailed results are discussed on the AUC only for clarity and brevity, the following reports the noteworthy tendencies for recall and precision at the default threshold of 0.5. All detailed metrics, including balanced accuracy, f1-score, precision, and recall, are published in Supplementary Table 2.

adaBoost generally achieved among the highest recall scores across runs, but in the case of the respective simple and extended versions of the questionnaire and behavior data, it was at the expense of precision. A similar pattern was observed for the NB model, which either had high recall or high precision but never a winning balance. For selected behavior and mixed features, the NN and adaBoost models achieved the most balanced result between recall and precision. However, as reported above, they were outperformed by the RF model in terms of AUC, which—at the default threshold—achieved higher precision than recall.

Discussion

Sophisticated ML models promise to disrupt mental healthcare through resource optimization and personalization[7,9], for example by lowering dropout[38] and improving health outcomes[39]. However, in DMHI settings, median data set sizes barely reach 155–350[7,9,27]. Such data set sizes are known to overfit and have been proven to not suffice for many sophisticated models[15–17]. However, very limited insights are available as of which data set size these problems are mitigated in DMHI settings. Therefore, the current study leveraged a dataset 10–24-times as big as the reported medians to evaluate performance levels, internal generalizability and variance across different feature groups (i.e., low to high predictive power with F = 2–129) and six model types (Naïve Bayes, Logistic Regression, Support Vector Machines, Random Forest, adaBoost, and Multilayer Perceptron Neural Network models).

Our first key finding confirms that CV results on small, thus most common, data set sizes overestimate the prediction performance. Especially worrisome is that the effect was exacerbated for uninformative features, such that a useless model had up to a 70% likelihood of returning seemingly good CV scores. Further, we reproduced the negative correlation between data set sizes and CV results[9,13] for $N \leq 300$ and partly $N \leq 500$ for the least predictive features. In these settings, such high training results were associated with the worst test results[13,18]. While overfitting was also prevalent in $N \leq 300$ for the

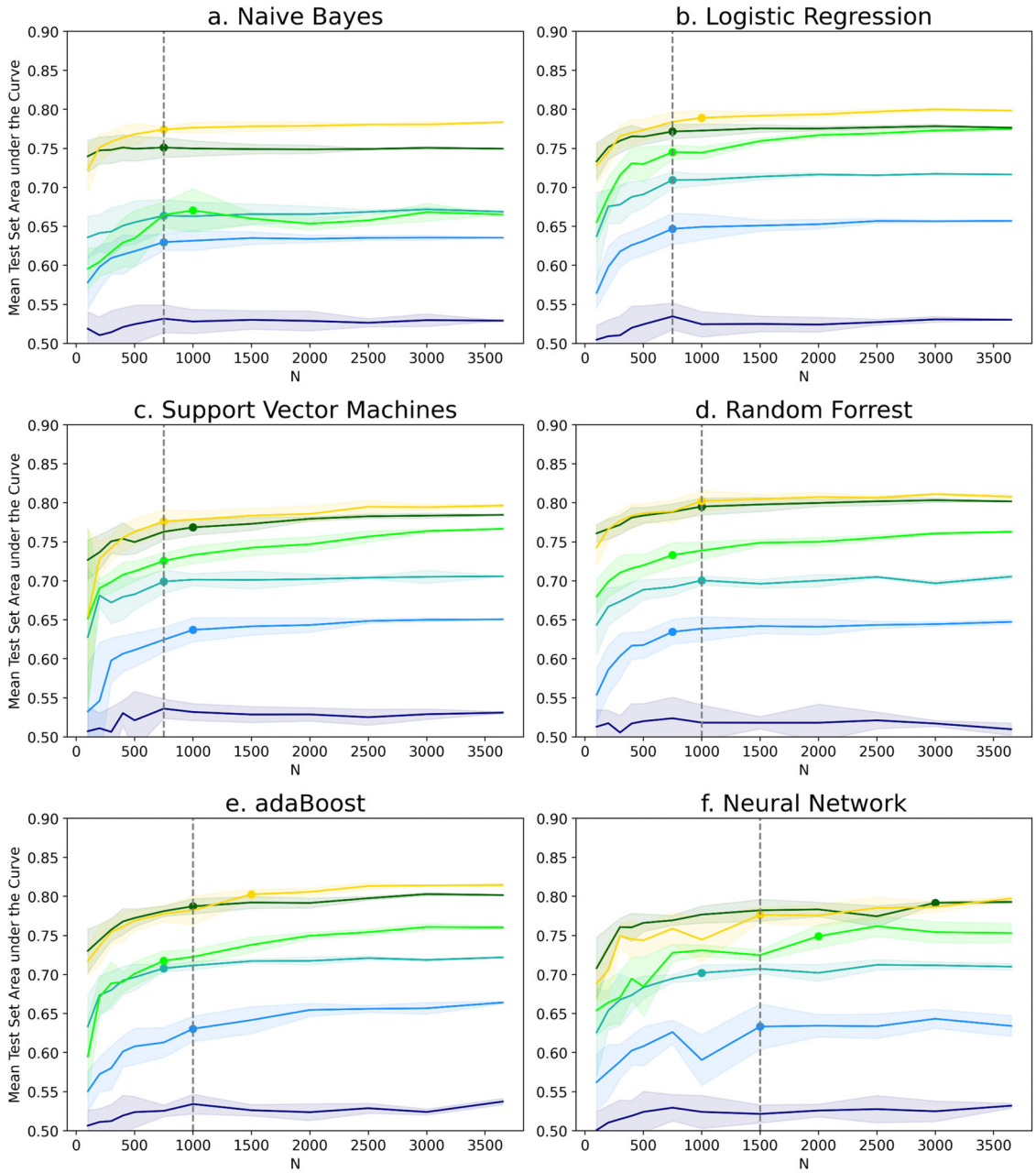


Fig. 3 | Test learning curve and convergence curves per model type. Learning curves on the test data per model: (**a**) Naive Bayes, (**b**) Logistic Regression, (**c**) Support Vector Machines, (**d**) Random Forest, (**e**) adaBoost, and (**f**) Multilayer Perceptron Neural Network. The colors indicate the different feature types, i.e., simple questionnaire (dark blue), extended questionnaire (light blue), simple behavior (turquoise), extended behavior (light green), selected behavior (dark green), mixed features (yellow). The respective mean area under the curve score is shown as sold horizontally plotted line and their S.D. as shaded area around it. Knee points indicate point of performance convergence as colored circles for the individual and grey dotted line as median across feature types. Knee points are not shown for simple questionnaire due to lack of predictive power.

more predictive features, it was lower, and the best training translated to the best test results. Further, among all features, the individual validation scores were highly unstable for $N \leq 300$ (S.D. 0.13–0.20 AUC). Evaluating on a single fold is common[9,18], and publication bias risks an overrepresentation of the higher end of that variance in published studies[8,40]. Thus, we conclude that results from data set sizes of $N \leq 300$ imply a substantial risk of being inflationary and ungeneralizable, especially for features with low predictive power.

A second, closely related key result is that CV scores on small data sets risk underestimating the superiority of complex versus simple features. This is caused by, firstly, large data being necessary to leverage additional features and, secondly, simple features overfitting more. For the largest feature group (F = 129), our data set size even may have been too small as it continued catching up to the already converged selected feature's performance. However, more research on larger data sets is necessary to investigate this hypothesis. Therefore, we tentatively confirm previous findings[12,31] that hand-crafted and theoretically driven selected features are preferable, especially for small data sets.

The third key result confirms that simpler models are less likely to overfit but converge earlier and are less competitive for higher data set sizes. More flexible models, on the other hand, heavily overfit small data sets but produce the best results on the high information features, especially for large data set sizes. Consistent with theory and empirical evidence[25,41], particularly NB gave robust results but was not very competitive overall. On the other

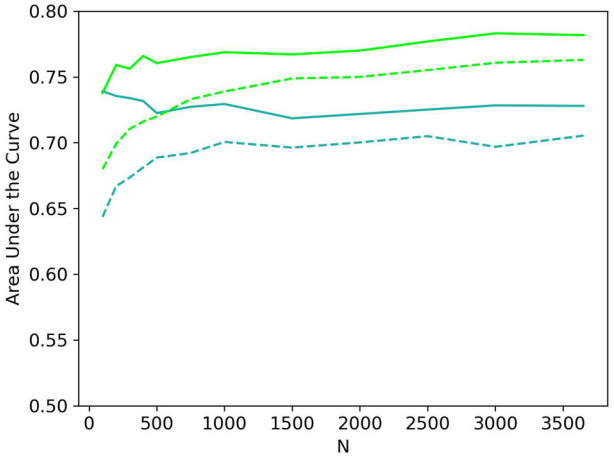


Fig. 4 | Random Forest simple versus extended behavior features. Random Forest area under the curve for Cross-Validation on training data (solid lines) and test data (dotted lines) learning curve for simple behavior data (F = 7) in green and extended behavior data (F = 129) in turquoise.

end of the spectrum, especially RF and SVMs seemed very competitive on noisy and small data sets but actually overfitted[25,42,43]. adaBoost performed badly on small but was most effective in leveraging large data sets. RF was one of the two most competitive algorithms across settings, already efficiently leveraging mid-sized data sets for predictive features[25]. LR was the second competitive algorithm, confirming its balance of overfitting less on small data sets[43] but only partly being outperformed in large data sets. The fact that LR is easier to interpret and faster to train than the tree-based models emphasizes its essential role as a staple baseline model to beat[6,27,31].

The fourth key finding is that prediction performance in our study did not converge until $N = 750$ for simpler, and 1000–1500 for more sophisticated models. Both are substantially above Giesemann et al.'s[21] findings that their results stopped improving at $N = 300$. A possible explanation is that their study on f2f-therapy investigated only eight features, which all fall in our extended baseline definition. As a result, their maximum test AUC score ($N = 10{,}000$) was 0.62, which our extended baseline data also achieved at $N = 300$. Further, in our data, more predictive features partly converged later than those with less information value. One possible hypothesis could, therefore, be that their earlier convergence point may be due to the limit of available predictive information in the features used. Thus, we conclude that more sophisticated models paired with larger data set sizes ($N > 750$) are necessary to approximate the true potential for the common feature groups in DMHIs.

Beyond the potentially still too-small sample size of 3654, this paper has several limitations. Firstly, it is only one case study, and while concurring with previous knowledge, this study per design does not suffice to reliably differentiate between setting-specific and generalizable tendencies. Further, the study at hand only focuses on internal generalizability and does not evaluate the models on an external data set. As models already overfitting the internal validation are unlikely to generalize to new data sets, our study constitutes a first step in the improvement of generalizability in this research area[18]. Regarding sample bias, the interventions considered are preventative and the sample only comprises self-referred female participants. Additionally, the five study arms were heterogeneous in their content, lengths, and user symptom strength[33]. As pooling interventions already mitigates overfitting[13], results may differ if repeated on a single intervention. However, this also implies that overfitting in this study may be underestimated, making the proposed increase of minimal data set sizes even more critical. Hence, the current study presents first insights, but more research is necessary to confirm the proposed minimal data set sizes. As a second limitation, while the operationalization of the outcome and feature groups was empirically and theoretically founded, many other options[12,27,44] are possible and may influence results. We proposed six different feature groups representing low to high predictiveness for intervention dropout, but they would, for example, differ in health outcome predictions[27,45]. Further, although recent works substantiate the assumption that our findings still apply[9,11,19], features such as neuroimaging or biological data are not considered in the current study. The same limitation applies to pre-processing steps and model choice, including more sophisticated Neural Networks than the shallow MLP used. Fourthly, while using the elbow method allows an analytical approach to determining convergence, it does not consider the trade-off of the cost that additional data points induce. Further, oscillations can influence elbow points, though mitigated by choosing the global instead of local elbow point.

In terms of recommendations, we, firstly, strongly discourage mistaking CV or, even less so, single test set results for suitable performance measures on small data set sizes ($N = 100$–300). Doing so exacerbates publication bias and causes ungeneralizable result expectations[13,18,19,40]. A key step against overfitting is separating the validation set for the hyperparameter decision from the model choice, for example, through nested CV[15]. Ideally, models should be validated on external data sets in addition to the internal validation methods in order to ensure broader generalizability[18]. Further, especially for complex features or ones with unknown or low information value, having a reasonably sized test set is indispensable[18,46]. Based on our results and previous suggestions[46], we, therefore, propose a minimal data set size of $N = 500$ for predictions in DMHIs to mitigate overfitting.

Secondly, even though $N = 500$ started producing internally reliable results, it did not suffice to approximate many of our feature groups maximum predictive power. Performance did not converge until $N = 750$ for LR, SVM and NB, and for the more flexible models, it even required $N = 1000$–1500. Further, the predictive power of additional and mixed features increased in higher data set sizes. We, therefore, suggest $N = 1000$ as a minimal data set size when comparing simple to more complex feature groups.

Lastly, and closely related to the other points, we recommend being mindful of the interaction between the nature and number of features, data set sizes and models. While ML methods can theoretically handle many features, for small data set sizes, the noise of additional features and the models' ability to overfit it must be considered[24,25,41,43]. Further, the hyperparameters, especially those concerning regularization, need to be chosen accordingly. To determine the adequateness of the set-up, we suggest implementing and reporting a learning curve approach leading up to the maximum available data set size. On the one hand, a downward CV trajectory suggests substantial overfitting. On the other hand, a continuously steep upward trajectory of both CV and test results suggests an underestimation of the predictive power due to a lack of data.

In summary, this paper contributes to the field of research by providing insights to aid the design and interpretation of predictions in DMHI settings. As such, it aims to combat unrealistic result expectations and the consequent disenchantment in a field where AI can be of great value but is only gradually gaining a foothold.

Methods

Case Study Background—everyBody Study

The everyBody dissemination study (ISRCTN13716228) provided evidence-based eating disorder (ED) prevention and health promotion programs[47–50] in Germany[33]. Participants ($N = 3654$) were adult women without full-syndrome EDs recruited from the general population between November 2016 and May 2019. All participants gave informed consent to participate in the study, and participation was anonymous. This primary study was a stratified, nonrandomized, parallel-group interventional design where intervention content matched risk and symptom levels. From the total sample, 452 users were allocated to the Basic intervention, 397 to Original, 1386 to Plus, 80 to AN, and 1339 to Fit. The interventions comprised 4 to 12 weekly online sessions (20 to 60 min) based on cognitive-behavioral principles, including psychoeducation, exercises to promote body image and balanced eating, and—if applicable—to reduce ED symptoms. Four out of five interventions were supplemented with daily or weekly online diaries. Four interventions had access to moderated peer group discussions, and two included weekly coach feedback messages.

Questionnaires were completed at screening, baseline, mid-intervention, post-intervention, 6-month, and 12-month follow-up. Analysis of pre-post changes of weight-related concerns within the completer subset revealed notable decreases in weight-related concerns across four of the five study arms (effect sizes d = −0.45 to d = −0.94)[51].

The screening and allocation process, individual intervention design and data generation is described in detail in Supplementary Note 1. Additional information can be found in the pre-registration protocol of the study[33] and its primary publication[51]. The trial was approved by the ethics board of Technische Univeristät Dresden (EK 83032016) and pre-registered at ISRCTN (No. 13716228, 26/02/2016). All participants gave informed consent to participate in the study, and participation was anonymous.

Definition of Outcome

Session completion was chosen to operationalize dropout, as it was found to be the most closely connected to intervention outcome[52]. While the different interventions had variable numbers of sessions (4–12), they presented similar dropout patterns, as seen in Fig. 5.

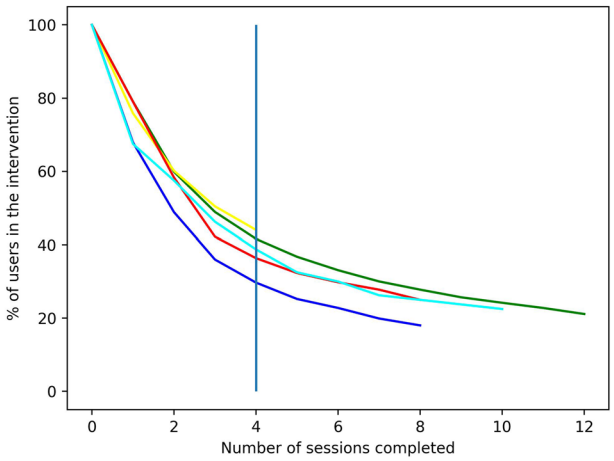


Fig. 5 | Dropout curves per intervention arm. Dropout curves defined by the share of users that finished each number of sessions across interventions, i.e., Fit (green), basic (yellow), Plus (dark blue), original (red), AN (turquoise). Vertical blue line indicates the cutoff point of four models, such that patients leaving on the left are categorize as dropouts and those on the right as completers.

Therefore, completing less than four sessions was defined as dropout to account for the minimum length of four weeks in the shortest intervention. This definition led to 56% dropout in the Basic intervention, 64% in Original, 70% in Plus, 61% in AN, and 58% in Fit. While many other dropout definitions are possible[53], this operationalization presents the possibility of identifying the users most at risk of leaving across interventions while ensuring sufficient time left to intervene[39].

Feature Groups and Pre-Processing

The most common overarching categories for dropout predictors are questionnaire data and intervention user behavior data[12,27]. For the current study, feature groups were categorized based on the number of features and their empirically proposed predictive power regarding dropout. The categories considered and their key details are shown in the overview in Table 2 and briefly described in the text below. Across all feature groups, the basic information of which intervention the user participated in, its lengths in weeks, and the starting year was also added.

The translated original questions and units can be found in Supplementary Table 1. An overview of the almost 200 features' description including their number of missing values is provided in Supplementary Note 2. All data processing was done in Python, primarily relying on the NumPy[54] and Pandas[55] libraries. Missing values were imputed with a multivariate iterative imputer[56] using the training sets questionnaire and weekly aggregated user behavior variables described below.

Regarding the questionnaire data, for the primary dissemination trial, various items were collected before intervention start, ranging from the standard primary symptom data up to less common measures such as personality scores. As pre-intervention questionnaire data has limited predictive power regarding dropout by itself[27,30,32,57], it was used to investigate a low predictive power setting. For the simple questionnaire data, only the screening and baseline primary symptom questionnaires (Weight Concern Scale[58]) were used. For the extended questionnaire data another 49 measures on psychological symptoms and characteristics, socio-demography, and user expectations were chosen based on their availability from the primary study and assumed usefulness. As a result, missing data was minimal, with five variables with <1.5% and six variables with <15% missing entries. The six latter were voluntary, and most users either answered all or none. Therefore, an additional variable was added to indicate this choice.

For the intervention user behavior data, log files and user submissions were aggregated into a set of simple, selected, and extended features. Only data from the first week of the intervention was used to leave sufficient time to intervene against dropout.

Table 2 | Overview Feature Groups

Name	Description	Key Aspects	#
Simple Questionnaire	Primary symptom scores (WCS)[58] at screening and baseline	Assumed low predictive power[27,30,32,57], available before intervention start	2
Extended Questionnaire	Variety of self-report questionnaires incl. WCS[58] and further eating disorder[67,68], depression[69], and anxiety[70] symptoms and behavior patterns, personality[71], self-regulation[72] and self-esteem scores[73], psychiatric and weight loss history, alcohol use[74], socio-demographic information, and user expectations.	Assumed low predictive power[27,30,32,57], theoretically available before intervention start but with high time-invest from users	51
Simple User Online Behavior	Sum of logins per day of the first week	Assumed high predictive power[28,30], very simple to obtain	7
Selected User Online Behavior	Single aggregation for the first week of time to complete sessions, seconds spent, number of logins, number and length of answers, diary entries and messages to coaches and group chat	Assumed high predictive power[12,13,27,31] with effort into researching and choosing most promising options and aggregation measures	13
Extended User Online Behavior	Variables from log files aggregated per day of the first week, incl. sessions completed, seconds spent, log ins, time spent in beginning/mid/end of the week and morning/day/evening, session completion, count and number of characters of diaries, group, and coach messages, exercises, answers to the sixteen most common closed questions as mean, min and max	High predictive power but possible loss due to complexity[12,31,38], automatically collected during first week of intervention with limited time invest	129
Mixed Features	Extended questionnaire + selected user online behavior	Mixed, with reported increase of predictive value[27]	64

The simple behavior data followed related work on generalizable features in DMHIs and counted the users' number of logins per day for the first week of the intervention[28,30]. For the selected user behavior, features were selected based on the related work[12,13,27,31,45] and aggregated per week, mitigating sparsity, multicollinearity, and complexity. For the extended user behavior, the same raw data instead was separately aggregated per day and included additional less known or theoretically less informative features as well as more aggregation forms (e.g., mean, minimum and maximum).

For the mixed features, the two types of features (selected behavior and extended questionnaire data) were added together for the last group to consider possible interaction effects[27].

Algorithms

Six common ML algorithms[16,26] were included in a trade-off of investigating different models while maintaining a reasonable computational load and ability to present results. For the simple algorithms, Naïve Bayes (NB)[59], Logistic Regression (LR), and Support Vector Machines (SVMs)[60] with a linear and radial kernel option and classifier were trained. In terms of more sophisticated tree-based models, first, Random Forest (RF) models were used due to their high flexibility and good performance in similar settings[13,26,27]. Second, to leverage the upsides of sequentially combining several tree learners, adaBoost decision trees were included. Lastly, a Multilayer Perceptron covered the family of Neural Networks (NNs). Considering the simplicity and small data set sizes at hand, a shallow architecture with a single hidden layer was chosen. All of these model types have been extensively discussed in various sources[16,61] and will, therefore, not be further detailed here.

Learning Curves and Training Set up

To estimate training performance, 10-fold cross-validation (CV) with grid search was implemented. The best resulting estimator was re-trained on the entire training dataset and evaluated on the previously set aside test set of 20% of the data. A standard scaler was incorporated into the pipeline. Regarding the hyperparameter ranges, default values were expanded upon if the outermost values appeared insufficient or excessive within the training data results.

Following authors such as Giesemann et al.[21], Balki et al.[20], and Perlich et al.[23], learning curves were used to provide insights into the effect of sample size on prediction performance. For the data set sizes, the space of 100, 200, 300, 400, 500, 750, 1000, 1500, 2000, 2500, 3000, and 3654 was explored to balance a comprehensive investigation with computational costs. The models were trained on 80% of the respective N to represent the data set sizes. The test set was stratified for dropout and each of the samples was stratified across the fives interventions. Further, training was repeated on samples drawn with different seeds ten times for small data set sizes (≤ 500), five times for the mid data set sizes (≤ 2000), and three times for the remaining large dataset sizes[21]. The model training was implemented with the scikit-learn[62] library in Python, and the code is publicly available in this paper's GitHub repository.

Evaluation and Result Analysis

The area under the curve (AUC) score was used to compare results across all settings without depending on a threshold. In terms of evaluation, the scores were classified into no (0.50–0–0.56 AUCs), low (0.57–0.64), moderate (0.65–70), good (0.71–0.75) and very good (>0.75) predictive power[63]. Predictive power per feature group was approximated through the test score for the model type with the highest training scores at $N = 3654$. A two-tailed DeLong test[64,65] with a significance threshold of $\alpha = 0.05$ was used to compare the test AUCs between models. The DeLong test was chosen because it is non-parametric, aimed at comparing AUCs and sufficiently computationally efficient[64]. The test returns the p-value for the null hypothesis of equal performance, hence the assumption that no model performs better than the other. Failing to reject the null hypothesis ($p > 0.05$) leads to possible differences in the AUC being assumed to be due to random chance.

The variability of results was determined through the standard deviation of single validation results across repetitions. To determine overfitting, first, the difference between the mean training and test score was considered. Next, the percentage of CV scores at least +0.10 AUC higher than the mean test set were investigated. The threshold 0.10 was chosen as it is a step that definitively jumped one results categorization introduced above, meaning, for example, a "low" score would become "good". Performance convergence was investigated by considering the diminishing marginal benefit of adding more data through the so-called elbow method. To this end, the kneed algorithm[66] Python implementation was used and set to find the global convergence point.

Data availability

The data used in this study is not publicly available due to legal restriction caused by the limitations in the data usage agreements and participants consent. However, qualified researchers can apply for data access through contacting the authors of this paper. The primary study's pre-registration is published[33].

Code availability

The code for the learning curves can be accessed through the following GitHub repository without restrictions: https://github.com/KiraZant/everybodylearningcurves.

Received: 21 June 2024; Accepted: 25 November 2024;
Published online: 18 December 2024

References

1. Cruz Rivera, S. et al. Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. *Lancet Digit. Health* **5**, e168–e173 (2023).
2. Ben-Israel, D. et al. The impact of machine learning on patient care: a systematic review. *Artif. Intell. Med.* **103**, 101785 (2020).
3. Shatte, A., Hutchinson, D. & Teague, S. *Machine Learning in Mental Health: A Systematic Scoping Review of Methods and Applications*. https://osf.io/hjrw8 (2018).
4. Aafjes-van Doorn, K., Kamsteeg, C., Bate, J. & Aafjes, M. A scoping review of machine learning in psychotherapy research. *Psychother. Res.* **31**, 92–116 (2021).
5. Arias, D., Saxena, S. & Verguet, S. Quantifying the global burden of mental disorders and their economic value. *eClinicalMedicine* **54**, 101675 (2022).
6. DeMasi, O., Kording, K. & Recht, B. Meaningless comparisons lead to false optimism in medical machine learning. *PLoS ONE* **12**, e0184604 (2017).
7. Hornstein, S., Zantvoort, K., Lueken, U., Funk, B. & Hilbert, K. Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms. *Front Digit Health* **5**, 1170002 (2023).
8. Squires, M. et al. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Inf.* **10**, 10 (2023).
9. Sajjadian, M. et al. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. *Psychol. Med.* **51**, 2742–2751 (2021).
10. Bzdok, D. & Meyer-Lindenberg, A. Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **3**, 223–230 (2018).
11. Vieira, S., Liang, X., Guiomar, R. & Mechelli, A. Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clin. Psych. Rev.* **97**, 102193 (2022).
12. Bremer, V., Chow, P. I., Funk, B., Thorndike, F. P. & Ritterband, L. M. Developing a process for the analysis of user journeys and the

prediction of dropout in digital health interventions: machine learning approach. *J. Med. Internet Res.* **22**, e17738 (2020).
13. Zantvoort, K., Hentati Isacsson, N., Funk, B. & Kaldo, V. Data set size vs homogeneity – A Machine Learning study on pooling intervention data in E-Mental Health dropout predictions. *SAGE Digit. Health* **10**, 20552076241248920 (2024).
14. Karyotaki, E. et al. Internet-Based Cognitive Behavioral Therapy for Depression: A Systematic Review and Individual Patient Data Network Meta-analysis. *JAMA Psychiatry* **78**, 361–371 (2021).
15. Bates, S., Hastie, T. & Tibshirani, R. *Cross-Valid.: what does it Estim. how well does it do it? arXiv* **119**, 1434–1445 (2024).
16. Hastie, T., Tibshirani, R. & Friedman, J. H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. (Springer, New York, NY, 2017).
17. Lateh, M. A., Kamilah Muda, A., Yusof, Z. I. M., Azilah Muda, N. & Sanusi Azmi, M. Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review. **892**, (2017).
18. Chekroud, A. M. et al. Illusory generalizability of clinical prediction models. *Science* **383**, 164–167 (2024).
19. Hilbert, K. et al. Lack of evidence for predictive utility from resting state fMRI data for individual exposure-based cognitive behavioral therapy outcomes: A machine learning study in two large multi-site samples in anxiety disorders. *NeuroImage* **295**, 120639 (2024).
20. Balki, I. et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can. Assoc. Radiol. J.* **70**, 344–353 (2019).
21. Giesemann, J., Delgadillo, J., Schwartz, B., Bennemann, B. & Lutz, W. Predicting dropout from psychological treatment using different machine learning algorithms, resampling methods, and sample sizes. *Psychother. Res.* **33**, 683–695 (2023).
22. van Smeden, M. & Moons, K. G. et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat. Methods Med. Res.* **28**, 2455–2474 (2019).
23. Perlich, C., Provost, F. & Simonof, J. S. Tree induction vs. logistic regression: a learning-curve analysis. *J. Mach. Learn. Res*. (2004).
24. Kwon, O. & Sim, J. M. Effects of data set features on the performances of classification algorithms. *Expert Syst. Appl.* **40**, 1847–1857 (2013).
25. Atla, A., Tada, R., Sheng, V. & Singireddy, N. Sensitivity of different machine learning algorithms to noise. *J. Comput. Sci. Coll.* **26**, 96–103 (2011).
26. Fernandez-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? (2014).
27. Zantvoort, K., Scharfenberger, J., Boß, L., Lehr, D. & Funk, B. Finding the Best Match—a Case Study on the (Text-)Feature and Model Choice in Digital Mental Health Interventions. *J. Healthc. Inform. Res.* **7**, 447–479 (2023).
28. Cote-Allard, U., Pham, M. H., Schultz, A. K., Nordgreen, T. & Torresen, J. Adherence Forecasting for Guided Internet-Delivered Cognitive Behavioral Therapy: A Minimally Data-Sensitive Approach. *IEEE J. Biomed. Health Inform*. 1–12 https://doi.org/10.1109/JBHI.2022.3204737 (2022).
29. Forsell, E. et al. Predicting treatment failure in regular care Internet-Delivered Cognitive Behavior Therapy for depression and anxiety using only weekly symptom measures. *J. Consult. Clin. Psychol.* **88**, 311–321 (2020).
30. Bricker, J., Miao, Z., Mull, K., Santiago-Torres, M. & Vock, D. M. Can a single variable predict early dropout from digital health interventions? Comparison of predictive models from two large randomized trials. *J. Med. Internet Res.* **25**, e43629 (2023).
31. Hentati, I. N., Forsell, E., Boman, M. & Kaldo, V. Methodological choices and clinical usefulness for machine learning predictions of outcome in Internet-based cognitive behavioural therapy. *Commun. Med.* **4**, https://doi.org/10.1038/s43856-024-00626-4 (2024).
32. Linardon, J., Fuller-Tyszkiewicz, M., Shatte, A. & Greenwood, C. J. An exploratory application of machine learning methods to optimize prediction of responsiveness to digital interventions for eating disorder symptoms. *Int. J. Eat. Disord.* **55**, 845–850 (2022).
33. Nacke, B. et al. everyBody–Tailored online health promotion and eating disorder prevention for women: study protocol of a dissemination trial. *Internet Inter.* **16**, 20–25 (2019).
34. Galmiche, M., Déchelotte, P., Lambert, G. & Tavolacci, M. P. Prevalence of eating disorders over the 2000–2018 period: a systematic literature review. *Am. J. Clin. Nutr.* **109**, 1402–1413 (2019).
35. American Psychiatric Association. Treatment of patients with eating disorders,third edition. *Am. J. Psychiatry* **163**, 4–54 (2006).
36. Linardon, J., Shatte, A., Messer, M., Firth, J. & Fuller-Tyszkiewicz, M. E-mental health interventions for the treatment and prevention of eating disorders: An updated systematic review and meta-analysis. *J. Consult. Clin. Psychol.* **88**, 994–1007 (2020).
37. Hilvert-Bruce, Z., Rossouw, P. J., Wong, N., Sunderland, M. & Andrews, G. Adherence as a determinant of effectiveness of internet cognitive behavioural therapy for anxiety and depressive disorders. *Behav. Res. Ther.* **50**, 463–468 (2012).
38. Pedersen, D. H., Mansourvar, M., Sortsø, C. & Schmidt, T. Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. *J. Med. Internet Res.* **21**, e13617 (2019).
39. Forsell, E. et al. Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: a single-blind randomized clinical trial with insomnia patients. *Am. J. Psychiatry* **176**, 315–323 (2019).
40. Andaur Navarro, C. L. et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* n2281 https://doi.org/10.1136/bmj.n2281 (2021).
41. Nettleton, D. F., Orriols-Puig, A. & Fornells, A. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artif. Intell. Rev.* **33**, 275–306 (2010).
42. Rodriguez-Galiano, V. F. & Chica-Rivas, M. Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models. *Int. J. Digit. Earth* **7**, 492–509 (2014).
43. Saseendran, A., Setia, L., Chhabria, V., Chakraborty, D. & Barman Roy, A. *Impact Noise Dataset Mach. Learn. Algorithms* https://doi.org/10.13140/RG.2.2.25669.91369 (2019).
44. Smink, W. A. C. et al. Analysis of the emails from the dutch web-based intervention "Alcohol de Baas": assessment of early indications of drop-out in an online alcohol abuse intervention. *Front. Psychiatry* **12**, 575931 (2021).
45. Hornstein, S., Forman-Hoffman, V., Nazander, A., Ranta, K. & Hilbert, K. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach. *Digit. Health* **7**, 205520762110606 (2021).
46. Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C. & Popp, J. Sample size planning for classification models. *Anal. Chim. Acta* **760**, 25–33 (2013).
47. Jacobi, C. et al. Indicated web-based prevention for women with anorexia nervosa symptoms: randomized controlled efficacy trial. *J. Med. Internet Res.* **24**, e35947 (2022).
48. Jacobi, C., Völker, U., Trockel, M. T. & Taylor, C. B. Effects of an Internet-based intervention for subthreshold eating disorders: a randomized controlled trial. *Behav. Res. Ther.* **50**, 93–99 (2012).
49. Jacobi, C. et al. Maintenance of internet-based prevention: a randomized controlled trial. *Int. J. Eat. Disord.* **40**, 114–119 (2007).
50. Beintner, I., Emmerich, O. L. M., Vollert, B., Taylor, C. B. & Jacobi, C. Promoting positive body image and intuitive eating in women with overweight and obesity via an online intervention: results from a pilot feasibility study. *Eat. Behav.* **34**, 101307 (2019).
51. Nacke, B. et al. Tailored online eating disorder prevention and health promotion for women: Results of a dissemination trial. (2024).

52. Donkin, L. et al. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. *J. Med. Internet Res.* **15**, e231 (2013).
53. Beintner, I. et al. Adherence reporting in randomized controlled trials examining manualized multisession online interventions: systematic review of practices and proposal for reporting standards. *J. Med. Internet Res.* **21**, e14181 (2019).
54. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
55. McKinney, W. Data Structures for Statistical Computing in Python. in 56–61 (Austin, Texas). https://doi.org/10.25080/Majora-92bf1922-00a (2010).
56. Roderick, J. A. & Rubin, D. *Statistical Analysis with Missing Data*. (John Wiley & Sons, Ltd). https://doi.org/10.1002/9781119013563.fmatter, (2002).
57. Günther, F., Yau, C., Elison-Davies, S. & Wong, D. On the Difficulty of Predicting Engagement with Digital Health for Substance Use. *Stud. Health Technol. Inform.* **302**, 967–971 (2023).
58. Killen, J. D. et al. Pursuit of thinness and onset of eating disorder symptoms in a community sample of adolescent girls: a three-year prospective analysis. *Int. J. Eat. Disord.* **16**, 227–238 (1994).
59. Zhang, H. The Optimality of Naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference* 2 (2004).
60. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
61. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*. (Springer US, New York, NY). https://doi.org/10.1007/978-1-0716-1418-1 (2021).
62. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
63. Kraemer, H. C. et al. Measures of clinical significance. *J. Am. Acad. Child Adolesc. Psychiatry* **42**, 1524–1529 (2003).
64. Sun, X. & Xu, W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* **21**, 1389–1393 (2014).
65. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
66. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a 'Kneedle' in a Haystack: Detecting Knee Points in System Behavior. in *2011 31st International Conference on Distributed Computing Systems Workshops* 166–171 (IEEE, Minneapolis, MN, USA). https://doi.org/10.1109/ICDCSW.2011.20 (2011).
67. Fairburn, C. G. & Beglin, S. J. Eating Disorder Examination Questionnaire. In *Cognitive Behavior Therapy and Eating Disorders*. (Guildford Press, New York, NY, USA, 2008).
68. Tylka, T. L. Development and psychometric evaluation of a measure of intuitive eating. *J. Couns. Psychol.* **53**, 226–240 (2006).
69. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9. *J. Gen. Intern. Med.* **16**, 606–613 (2001).
70. Spitzer, R. L., Kroenke, K., Williams, J. B. W. & Löwe, B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch. Intern. Med.* **166**, 1092–1097 (2006).
71. Rammstedt, B., Kemper, C., Klein, M., Beierlein, C. & Kovaleva, A. *Eine Kurze Skala Zur Messung Der Fünf Dimensionen Der Persönlichkeit: Big-Five-Inventory-10 (BFI-10)*. (2012).
72. Carey, K. B., Neal, D. J. & Collins, S. E. A psychometric analysis of the self-regulation questionnaire. *Addict. Behav.* **29**, 253–260 (2004).
73. Rosenberg, M. Society and the Adolescent Self-Image. in *Society and the Adolescent Self-Image* (Princeton University Press). https://doi.org/10.1515/9781400876136 (1979).
74. Bush, K. et al. The AUDIT Alcohol Consumption Questions (AUDIT-C): an effective brief screening test for problem drinking. *Arch. Intern. Med.* **158**, 1789–1795 (1998).

Acknowledgements

The original trial and data collection was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 634757. The funding agency was not involved in any decisions regarding the data collection or study methodology. The authors would like to thank Ina Beintner, Bianka Vollert, and Juliane Schmidt-Hantke for their extensive contributions during the design, preparation and conduction of the primary trial. This open-access publication was funded by the German Research Foundation (DFG).

Author contributions

B.N., D.G., and C.J. designed the trial that provided the data for the current study, B.N. and C.J. conducted the trial. D.G. was the trial statistician and responsible for the data management. K.Z., B.N., D.G., S.H., and B.F. contributed to the development of the analysis performed. K.Z. developed the idea, wrote the code, analyzed the data, and wrote the first draft of the paper. B.N., S.H., B.F., C.J., and D.G. reviewed and contributed to the final draft.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01360-w.

Correspondence and requests for materials should be addressed to Kirsten Zantvoort.

Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–11

DOI: 10.1177/20552076241248920
journals.sagepub.com/home/dhj

Sage

Dataset size versus homogeneity: A machine learning study on pooling intervention data in e-mental health dropout predictions

Kirsten Zantvoort[1], Nils Hentati Isacsson[2], Burkhardt Funk[1] and Viktor Kaldo[2,3]

Abstract

Objective: This study proposes a way of increasing dataset sizes for machine learning tasks in Internet-based Cognitive Behavioral Therapy through pooling interventions. To this end, it (1) examines similarities in user behavior and symptom data among online interventions for patients with depression, social anxiety, and panic disorder and (2) explores whether these similarities suffice to allow for pooling the data together, resulting in more training data when prediction intervention dropout.

Methods: A total of 6418 routine care patients from the Internet Psychiatry in Stockholm are analyzed using (1) clustering and (2) dropout prediction models. For the latter, prediction models trained on each individual intervention's data are compared to those trained on all three interventions pooled into one dataset. To investigate if results vary with dataset size, the prediction is repeated using small and medium dataset sizes.

Results: The clustering analysis identified three distinct groups that are almost equally spread across interventions and are instead characterized by different activity levels. In eight out of nine settings investigated, pooling the data improves prediction results compared to models trained on a single intervention dataset. It is further confirmed that models trained on small datasets are more likely to overestimate prediction results.

Conclusion: The study reveals similar patterns of patients with depression, social anxiety, and panic disorder regarding online activity and intervention dropout. As such, this work offers pooling different interventions' data as a possible approach to counter the problem of small dataset sizes in psychological research.

Keywords

e-mental health, dropout, machine learning, prediction, ICBT

Submission date: 3 August 2023; Acceptance date: 4 April 2024

Introduction

Modern societies struggle to provide adequate mental health care,[1–3] as traditional therapy alone is not meeting the increasing need.[4] Internet-based Cognitive Behavioral Therapy (ICBT) promises to improve care levels by achieving similar goals as face-to-face therapies through efficient digital means.[2,5] With the rise of ICBTs, a large variety of user online activity data becomes recordable. Applying advanced analytics to this data holds great promise to

[1]Institute of Information Systems, Leuphana University, Lueneburg, Germany
[2]Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, & Stockholm Health Care Services, Stockholm, Sweden
[3]Department of Psychology, Faculty of Health and Life Sciences, Linnaeus University, Växjö, Sweden

Corresponding author:
Kirsten Zantvoort, Institute of Information Systems, Leuphana University, Universitätsallee 1, 21335 Lüneburg, Germany.
Email: kirsten.zantvoort@leuphana.de

individualize and improve care.[6–8] One task that presents itself to be solved with Machine Learning (ML) is that of intervention dropout predictions.[4,6] A patient dropping out of an intervention is significantly less likely to have positive outcomes.[9] Yet, upon starting the intervention, costs occur, and scarce resources are occupied.[10] Measures such as guidance from therapists lower dropout rates,[11] but are often too costly to be provided to all patients.[12] Identifying patients at risk of dropout early on allows for the personalized allocation of measures. If more individuals' needs for support are met, resource allocation is optimized, and health benefits increase.[13] In contrast to the direct prediction of health outcomes, dropout predictions include patients that otherwise tend to be ignored due to missing symptom data.[14,15]

Initial studies using ML to predict dropout based on user behavior data show promise.[6,16–21] Nevertheless, there are still few ML applications in ICBTs, especially dropout predictions.[16,21–24] A recent systematic literature review found that, despite the widespread consensus about its value, only 3 out of 94 digital interventions use ML to personalize interventions for depression.[8] This scarcity of work is attributed to the small size of available datasets for training,[8,25] as it limits the accuracy and generalizability of predictions.[26–28] Collecting large health datasets is costly,[29] and the median dataset size across 59 studies using ML for outcome predictions in depression treatments was found to be only 115 patients.[30] Similarly, the median for related work predicting dropout in online interventions was 342 patients.[6,16–21] Thus, with ML approaches in ICBTs, the question arises of how to produce accurate predictions despite small datasets.[7,23,31]

Albeit the lack of large datasets, the number of smaller datasets available was already reported to be in the hundreds 5 years ago.[32] Data sharing between providers creates larger training datasets but poses significant challenges regarding data privacy, data interoperability, and conflicting interests.[33] However, many providers themselves offer similar interventions for different but related primary symptoms, such as depression and anxiety disorders.[20] As symptoms and behavior are interconnected,[34] common behavioral patterns between patients could be leveraged by pooling several interventions into one dataset. If successful, this not only improves prediction results but also lowers model maintenance efforts. However, pooling the data may be detrimental if contexts and user behaviors differ significantly. Most papers predicting dropout focus on a single intervention,[6,16,17,20,21] while one gathers two different interventions for the same symptoms,[19] and one gathers interventions for three different symptoms.[18] While this shows that different options are possible, no study comparatively evaluates the approaches. It, generally, remains unclear how patients with different but related target symptoms vary in their online intervention behavior. Clustering analyses have successfully identified user archetypes in mental health apps in general[35] and within specific interventions.[36] Applying such a clustering analysis to users of different but related ICBTs could offer valuable insights into the similarities in user behavior.

In this study, we use demographic, symptom, and online user behavior data of 6418 routine care patients from the Internet Psychiatry in Stockholm, Sweden. The main research question is if the value of pooling intervention data for social anxiety (SAD), major depression (MDD), and panic disorder (PD) outweighs the downside of losing data homogeneity when predicting intervention dropout. The goal is to identify patients who will end up leaving the intervention early without benefiting, already in intervention week 4 of 12. For this, four different supervised ML methods (i.e. Logistic Regression, Support Vector Machines, Random Forest, and AdaBoost classifiers) are compared. We further investigate the relationship between prediction performance and dataset size. To this end, the training on individual versus pooled datasets is repeated on samples of the median dataset sizes of related work. A clustering analysis is the intermediate step to understanding the differences and similarities between the interventions' data. Through these proposed steps, this study aims at (1) exploring the heterogeneity of online intervention data across three highly prevalent mental disorders and (2) providing insights into what pooling these different intervention datasets into one dataset yields for dropout prediction. Finally, this work adds to the limited body of research investigating dropout predictions across three large routine care interventions.

Methods

Interventions and participants

This study uses routine care outpatient data from the Internet Psychiatric clinic in Stockholm, Sweden from 2008 to 2020.[37] The data comprises all available patients undergoing treatment while the platform in question was used. The datasets consist of 1633 PD, 1907 SAD, and 2902 MDD patients. The treatments have previously been evaluated with positive results.[38–40] After a psychiatric assessment, each patient received 12 weeks of disorder-specific intervention. The assessment and treatment evaluation are based on established patient self-rating measures; for PD, the Panic Disorder Severity Scale-Self Report[41]; for SAD, the Liebowitz Social Anxiety Scale-Self Report,[42] and for MDD, the Montgomery–Åsberg Depression Rating Scale-Self Report[43,44] were used. Each of the three interventions was administered in the same clinical context where patients self-refer and then go through a web-based screening and the established semistructured M.I.N.I diagnostic assessment interview[45] with a psychiatrist. During the interview, the clinician provides information about (I)CBTs in general, and the expected effort required

from the patient to complete treatment. As such, all patients are ensured to have a relevant diagnosis as well as being informed about the treatment, and sufficiently motivated to engage in it.

Included patients receive the same therapist support routine across all three interventions. The interventions reside on the same technical platform and are very similar in structure, but clearly differ in therapeutic content. All interventions start with psychoeducation and consist of cognitive behavioral therapy techniques specific for each condition divided into 10 modules. For example, the SAD and PD interventions include symptom-specific exposure exercises, whereas the MDD has a focus on behavioral activation. Each intervention's modules consist mainly of exercises, including homework, messages from and to a therapist and weekly symptom assessments. More detailed information about the interventions is summarized in Supplemental Appendix 1 and has previously been published for SAD,[38] depression,[39] and PD.[40]

Features

The prediction is based on the first 4 weeks of data as a trade-off between gathering sufficient data versus maintaining sufficient time to intervene to prevent dropout.[6] A previous study has shown personalizing the support level in week 4 to have a positive effect for patients at risk.[13] For the features, thus, all data gathered after week 4 is disregarded to prevent target leak. First, we include the common sociodemographic variables, age, and gender.[21] Second, given their importance,[8,44] the symptom measures at screening, and at the beginning of weeks 1, 2, 3, and 4, respectively, are included. In addition to the actual scores, the time needed to fill out the questionnaires is included. Third, we use the basic information of the intervention set-up (i.e. year of start, start in winter or summer, and target disorder).[21] Fourth, the character length of the homework assignments and messages are each summed per week, which have previously been found to be predictive of both dropout and health outcomes.[46] To account for the therapist messages, the percentage of characters sent in the conversation produced by the therapist is included. Following the insights from Bremer et al.,[6] several features are generated from patients' log data. This includes the sum of time spent on the intervention, the number of pages clicked, the number of sessions, and number of days that patient logged in. In addition, the time patterns of the login behavior are gathered across all weeks by looking into the percentage of sessions per weekdays and daytimes. Further, we record how many days a patient needed to finish each module. Further information on the preprocessing and feature engineering steps can be found in Supplemental Appendix 2.

The operationalization of the dropout variable is aimed at identifying the patients who are most likely to leave the intervention too early to sufficiently benefit.[47] The intended symptom improvement is determined by either the final symptom score below the absolute cutoff for remittance (8 for PDSS-SR,[48] 35 for LSAS-SR,[49] and 11 for MADRS-S[50]) or a 50% improvement since the start of the treatment.[51] If no symptom score after week 8 is available, their symptom improvement is classified as "unknown." Module completion has been found to be the adherence measure with most consistently positive power toward explaining therapy outcome.[52] For this study, we use module 8 of 10 as it contains all unknown leavers, it includes all of the content introductions, as the last 2 modules are repetition and maintenance,[52,53] and produces considerably balanced classes. A more detailed explanation of the dropout variable can also be found in Supplemental Appendix 2. The averages, SD and units of the resulting dataset can be found in Supplemental Appendix 3. All steps, including the subsequent modeling, are implemented in Python, using the pandas,[54] Numpy,[55] Kneed algorithm,[56] and Scikit-learn[57] libraries.

Exploration of heterogeneity between interventions

This analysis addresses the first goal: the exploration of heterogeneity in patients across the three interventions using the demographic, intervention, symptom, online activity, and character counts variables. The general purposes of clustering are to gain insight into the data, identify natural groups, and be able to summarize them based on segment prototypes.[58] First proposed in 1967, k-means algorithms are among the most used clustering approaches due to their computational efficiency and easy implementation.[58,59] The k-means algorithm is an optimization algorithm that iteratively finds a set of k centroids, such that the total sum of distance between each point and its nearest centroid is minimized.[60] As such, it optimizes for groups that are as similar as possible within themselves but as different as possible from each other.[60] The number of clusters needs to be decided apriori and is inferred using the Elbow (or Knee) method.[61] In essence, this method uses the explained data variance to reveal where the marginal gain of a new cluster is outbalanced by the increasing number of clusters. As commonly done, a Principal Component Analysis is conducted prior to the clustering to lower the number of dimensions, reduce multicollinearity and facilitate the visualization of clusters.[62] The number of principal components is also automatically identified by using the Kneed algorithm.[56] The critical question is whether the algorithm will rely on the target disorder variable identifying each intervention as primary splitting criteria. If the different target symptoms result in different online behavior, the clustering algorithm would be expected to reproduce the three intervention groups. Only if patients behave sufficiently similar across interventions, mixed clusters can be expected. To ensure comparability, this process is

done only on the first 4 weeks of data also available to the prediction models. This excludes the dropout and health outcome variable for all patients, which will, however, be added after completing the clustering for the evaluation of clusters.

Dropout prediction

The second goal is investigating the effects of pooling patients from interventions for depression, SAD, and PD to one dataset when predicting dropout. To this end, models trained on each of the interventions' data individually are compared to models trained on all three interventions pooled. As the data has almost twice as many MDD as PD patient, we add a pooled run where we under sample the large interventions to have balanced ratios of one-third each. The training dataset sizes at hand are in the four digits, which is already unusually big.[30] To increase the usefulness of results for future studies, the training process is repeated on smaller samples—the median dataset sizes of related work for outcome prediction (115)[30] and dropout prediction (342).[6,16–21] Taking away an assumed 15% data for a holdout test set results in training data of 98 and 291 patients per single intervention.

While the training data differs in dataset size, all models are evaluated on the same 20% stratified test set to maintain comparability across runs. The final evaluation is done (1) relatively and (2) absolutely, both focusing on balanced accuracy (BACC). For the relative evaluation, the single versus pooled data results are compared. For the absolute comparison, the benchmark of (1) better than chance and (2) 67% balanced accuracy as minimally necessary to be valuable in an ICBT to adapt treatment as proposed by Forsell et al.[12] are used. Further, to adhere to the standards of medical ML studies, accuracy, balanced accuracy, specificity, recall, and area under the curve (AUC) are provided for the test set performances.[63]

In terms of algorithms, Logistic Regression (LR), Support Vector Machines (SVMs),[64] Random Forest (RF), and AdaBoost[65] classifiers are chosen as they cover a range from simple and robust to more sophisticated and flexible options.[31,62] As extensively argued,[66,67] choosing the algorithm to use in the same step as optimizing the hyperparameters comes with a significant risk of overfitting. Therefore, the model selection will be done through 5×10 nested-cross-validation (CV). The inner CV ptimizes hyperparameters via grid search, and the outer CV score determines the one algorithm to use. That algorithm is then retrained on the whole training data with a 10-fold CV, returning a single model to be evaluated on the test set. An intervention-based scaler is added to the pipeline, such that a standard scaler is fitted to the training data per intervention and then applied on the respective holdout fold.[63]

To choose the range of hyperparameters, initial values are run and added to if the outer points seem too low or high. For the algorithms that allow for balancing class weights, the class weights are balanced. For the LR, the choice of L1 and L2 feature selection is optimized as a hyperparameter for the liblinear solver.[68] The C value is searched across the range [0.001, 0.01, 0.05, 0.1, 0.20, 1]. The SVMs optimize over an RBF and a linear kernel with respective C values [0.001, 0.01, 0.1, 0.25, 0.5, 1]. The RF model searches across the number of estimators [5, 10, 25, 50, 500, 1200], the minimum samples [10, 25, 50, 100, 200], the maximum depths [5, 25, 50, 100, 500, 750], and a binary indicator for bootstrapping. Lastly, the AdaBoost Classifier trades off the number of estimators [1, 2, 5, 10, 25, 100, 1500] with their respective learning rate [0.001, 0.01, 0.1, 1, 2, 2.5].

Results

Data heterogeneity

To better understand the differences in online behaviors, user characteristics and symptom patterns, the 1631 PD, 1906 SAD, and 2881 MDD patients and their 57 input variables that result from preprocessing as described in Supplemental Appendix 2 are clustered. The kneed method suggests four principal components to represent the input data. Feeding these components into the k-means algorithm with k ranging from one to 11 suggests 3 most prevalent clusters. However, this optimal value only coincides with the number of interventions, as each intervention spreads comparatively evenly across clusters (Figure 1). The biggest intervention group, MMD, makes up 42–51% of each cluster, SAD accounts for 29–30%, and the smallest intervention, PD, spreads at 20–28% per cluster. To facilitate understanding, the clusters will be referred to as active, middle, and inactive clusters from now on for the reasons explained below. The middle cluster is by far the biggest as it contains 46% of all patients, with the inactive cluster following at 35% and the very active cluster tailing at 19%. All cluster means reported in this section can also be found in the overview table in Supplemental Appendix 3.

Inactive patients are more than 6 times as likely to have missing symptom scores (0.84/5) in the first 4 weeks as active patients (0.13), who are similar to the middle cluster (0.16). Further, the average lengths of messages and homework in the first weeks are 6 times as high for the active cluster (459/1331 characters) as for the inactive cluster (81/224 characters), with the middle clusters averaging at 157/1008 characters. Similarly, login data such as sessions, pages and duration per week are almost all 2–4 times as high for active as inactive patients, with the middle cluster somewhere in-between. The most extreme differences are in the durations where inactive patients, on average, spent one-third or one-fourth of the time (12,071, 6698, 6198, and 6878 s per weeks 1–4) that

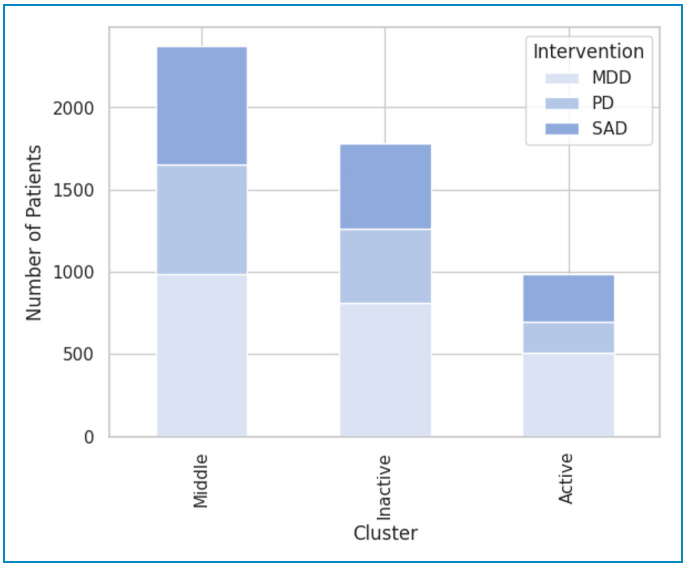


Figure 1. Distribution of patients per intervention and cluster. MDD: major depressive disorder; PD: panic disorder; SAD: social anxiety disorder.

active patients (32,393, 25,850, 24,738, and 24,171) spent. This is also reflected in the average number of modules completed in the first 4 weeks, with 2.7 for inactive, 4.2 for the middle, and 4.7 for the active patients.

While the starting symptom scores barely differ, inactive patients have a higher average symptom improvement between screening and treatment start (−12%) than middle (−9%) or active (−8%) patients. However, this changes in the following weeks. While inactive patients still see 12% improvement in week 2, they have next to no change (0%, −1%) in week 3 or 4. While the strength of change also lessens for middle and active patients, they still continuously improve (middle: −16%, −3%, −4% and active: −15%, −4%, −5%). The least differentiating variables are the start year, time variables (e.g. time and weekday of intervention use), symptom questionnaire duration, if they started the intervention in the winter, and the patient's age and sex. Retrospectively joining the dropout variable to the clustered data shows that patients from the active cluster are less than half as likely to drop out (25%) as patients from the inactive cluster (65%), with the middle cluster being closer to the active cluster (36%). Despite the dropout rates heavily differing across interventions (PD 28%, MDD 45%, and SAD 57%), the differences in dropout probability per cluster remain. Dropout ratios for the inactive, middle, and active groups are 68%, 36%, and 24% for MDD, 41%, 23%, and 16% for PD, and 80%, 49%, and 33% for SAD. Doing the same for the health outcomes shows that 53% of active and middle patients are treatment successes while only 34% of inactive patients are. For 15% of inactive patients, their health outcome is unknown, whereas the middle cluster has 5% and the active cluster only 2%. As a result, the percentage of not successful treatments is close together between active (45%), middle (43%), and inactive (50%) patients.

Prediction

The train-test split leads to a maximum of 5132 training data points and 1289 test data points, for which the averages of all variables can be found in Supplemental Appendix 3. Of these data points, 45% are MDD, 30% are SAD, and 25% are PD patients, resulting in the unbalanced pooled training datasets in Figure 2. The small and medium balanced pooled training data have the same total as the unbalanced run; however, they have balanced ratios of one-third per intervention. For the large data, balancing is dictated by the smallest intervention (PD), resulting in a sample size of 1304 each. The nine single intervention runs (three per disorder) with 98, 291, and 1304/1524/2303 are not separately shown in Figure 2.

Figure 3 presents the BACC results for each intervention, dataset size, and type of training data. The box plots show the 10 outer CV scores of the training data, while the single bullet point shows the performance on the test set. An ideal result graph has a high y-axis value (balanced accuracy) with a narrow boxplot (low variance in training results). The boxplot's horizontal stripe (median) should be close to the test set point (neither overfitting nor unexpectedly high results). For pooled data results in the following, the unbalanced (UB) results are mentioned first, followed by the balanced (B) results. To answer the main research question, the single dataset runs are compared to their respective pooled counterparts (e.g. 98 data points single intervention vs $3 \times 98 = 294$ data points pooled run). All results, including the numbers discussed here,

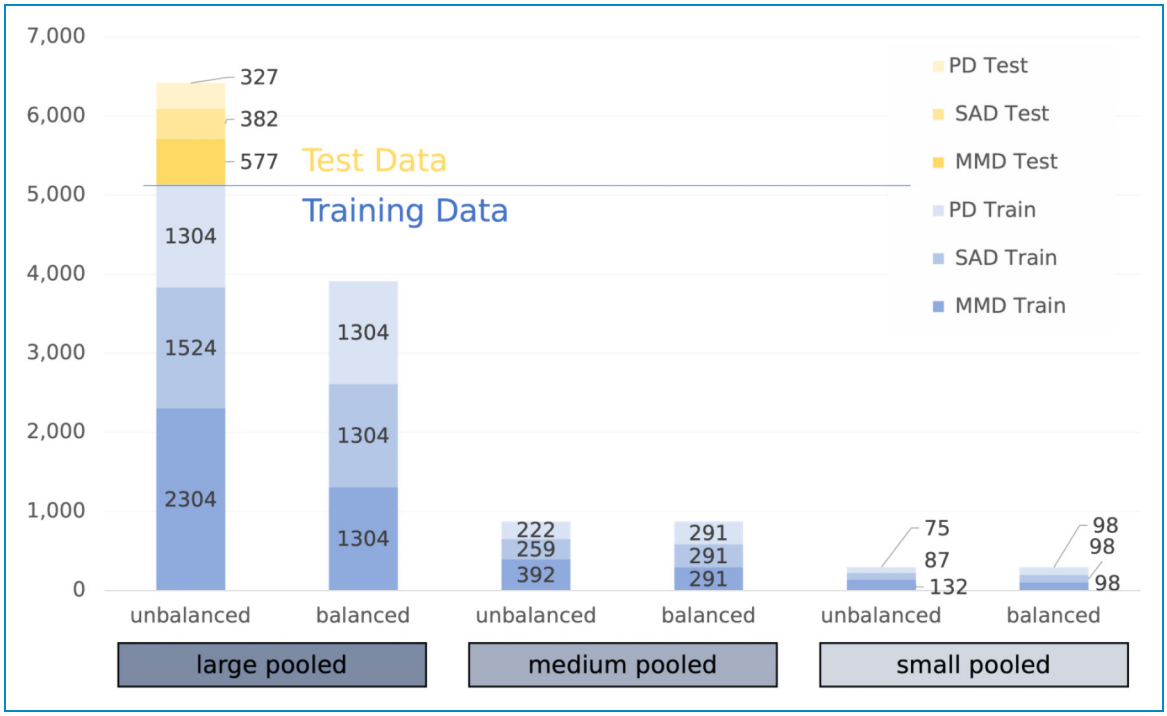


Figure 2. Training and test data sample per pooled run. MDD: major depressive disorder; PD: panic disorder; SAD: social anxiety disorder.

accuracy, recall, specificity, AUC score, and the model type chosen in the CV can be found in the result table in Supplemental Appendix 4.

Training data results. Only looking at training results, single runs outperform the pooled datasets in predicting dropout. The single intervention runs have a higher median outer CV score than both pooled runs in six out of nine cases with an average advantage of 0.037 BACC. Further, single runs are higher than at least one pooled run in two more cases. Hence, the only exception is the small (98 data points) PD dataset, where training results for the single run are lower than both pooled data BACCs. For MDD, the median CV score increases as the dataset size increases from small to medium to large, with a total difference of +0.04 BACC. The opposite pattern is visible for PD, such that the training scores decrease as more data is added with −0.05 BACC. Similarly, SAD has the highest score for the small data but then has the smallest score for the medium and an average score for the large data with a total range of −0.08. Considering the range between the first (Q1) and third (Q3) quartile of the CV scores, pooled runs have a lower average range than single interventions. The spread of the training results (Q3–Q1) is lowest for the large datasets at an average of 0.045 in BACC. For the small datasets, it is more than four times as much (average at 0.202), with the medium dataset size in-between (average 0.101).

Test data results. Which setting (single vs pooled) performs best is reversed when looking at the test instead of training results. Here, in seven out of nine cases, both pooled runs outperform the single intervention. Additionally, the unbalanced pooled data outperforms the single intervention for the smallest MDD run (0.64 vs UB 0.66/B 0.62). Hence, the single intervention is only superior in one case, the medium PD run. As such, PD patients have both one of the biggest gains and biggest loss (+0.063/−0.044 BACC) when using the pooled model instead of one trained on the PD patients alone. The runs with all available PD data barely differ (+0.007 for pooled). For SAD patients, the small data gains somewhat from pooling (UB +0.046/B +0.026), the medium dataset size has an average and the highest gain (+0.037/+0.072), and there is barely any difference in the large data (+0.000/0.002). On the contrary, MDD patients almost always benefit from pooling the data, and the gains grow with the dataset size (small data UB +0.021/B −0.016, medium data +0.032/+0.028, large data +0.045/+0.040). Keeping the natural ratio is superior for all three small data runs and the medium and large data MDD runs. Balancing the data is favorable for all medium and large datasets of SAD and PD. The largest impacts of balancing the data are −0.037 BACC for the small MDD and +0.035 for the medium SAD datasets. For MDD and SAD, the pooled data's higher BACC also means a higher recall than the single intervention datasets. However, in the case of PD, the pooled datasets have high specificity (average 0.80)

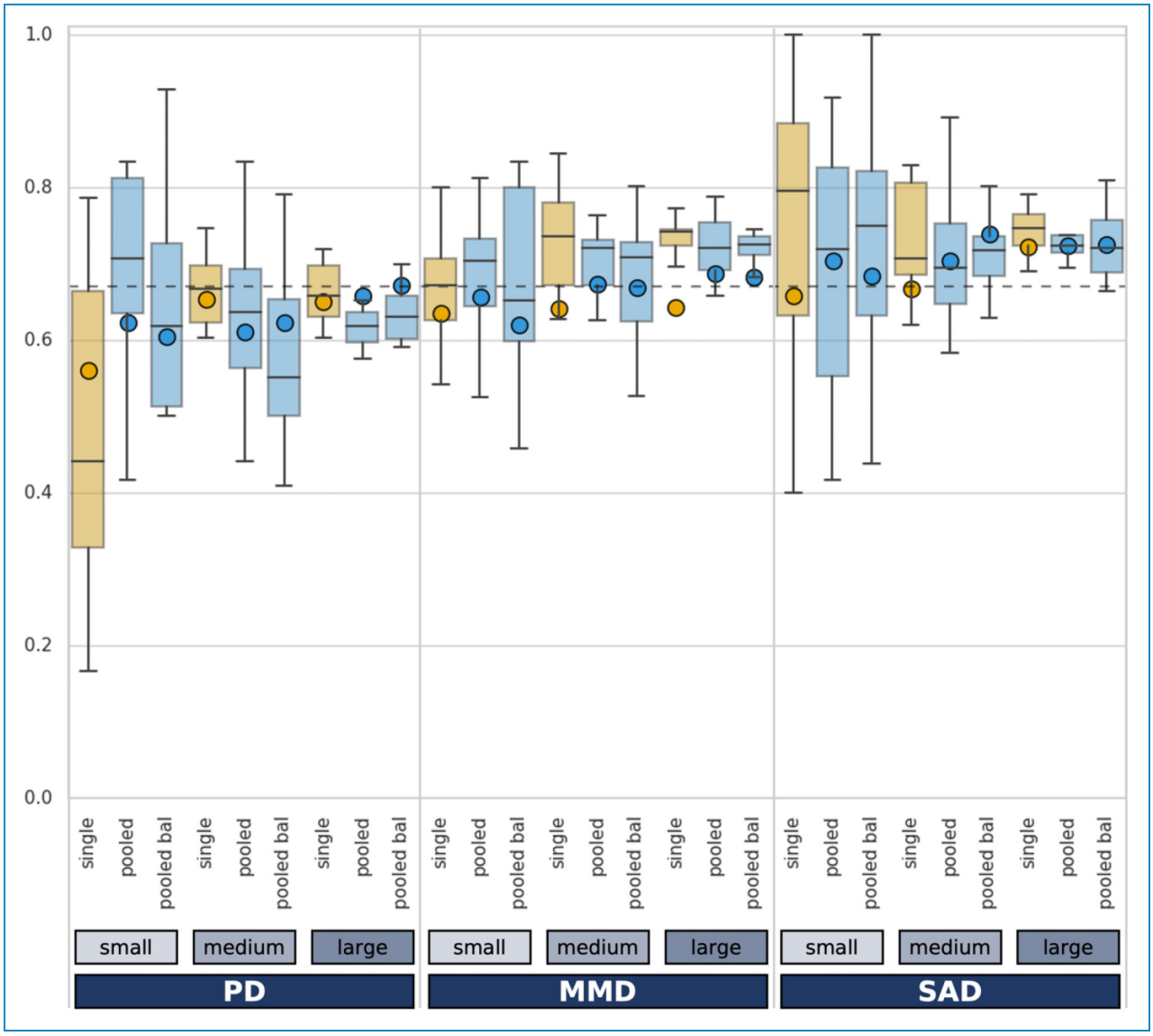


Figure 3. Balanced accuracy for training and test results per run with clinical threshold as dotted line. Bal: balanced intervention shares; MDD: major depressive disorder; PD: panic disorder; SAD: social anxiety disorder.

but lower recall (average 0.46) whereas the single intervention runs are more balanced (average specificity: 0.57, recall: 0.67).

Difference between training and test results. In terms of potential overfitting, the pooled runs have a smaller absolute difference between the test and training results in seven out of nine cases. The only exceptions are the medium and large PD datasets. This results in an average absolute train-test difference in BACC of 0.063 for single and much lower 0.034/0.037 (UB/B) for pooled data. The smallest dataset of SAD has the largest gap (−0.14) with the highest training results out of all runs but the lowest test results of all SAD models. Similarly, the training and test results of the MDD models are twice as much apart for the single data as for the pooled data, and the pooled model achieves better test results in all cases. The only exception to this rule is the medium PD dataset, where the single model achieves better and closer test and training results. However, if pooled runs outperform the single interventions in the training data, they consistently also outperform it in the test set.

Absolut evaluation. Regarding the absolute evaluation, all models achieve a balanced accuracy of more than 0.5 BACC, and are, therefore, better than chance. Further, 13 of the 27 models achieve a BACC on the test set of 0.67 or higher, with results differing across interventions and

settings. For SAD, all but the single intervention small patient models achieve the threshold, with a maximum BACC of 0.74. For PD and MDD, not one model trained on the single intervention achieves clinically relevant prediction results. However, for MDD, both pooled model for the medium (0.67/0.67) and large (0.69/0.68) data achieve the threshold. For PD, only the largest balanced pooled model achieves clinically relevant prediction results on the test set.

Discussion

Researching the heterogeneity in patient data for ICBTs for MDD, SAD, and PD, we found intervention-overarching patient groups in the first weeks of the interventions. Despite differing dropout rates per intervention (28–57%), the algorithm identified the respective most likely and least likely clusters to dropout. The active, middle and inactive clusters' correspondence to low, middle and high dropout is in line with previous findings.[47] Our first finding, that SAD, PD, and MDD patients have similar clusters of activity patterns may help the design and delivery of both individual and transdiagnostic interventions.

The answer to the first research question already hints toward the answer to the second; pooling the data was almost always favorable and doubled the likelihood of achieving clinically relevant test results. Most noticeably, having 873 mixed intervention training data points outperformed having 2304 individual intervention MDD or 1524 SAD patients. A possible hypothesis for this is that pooling different interventions forces the model to focus on general patterns rather than intervention-specific noise. Beyond better results, pooling data comes with the upside of less resources necessary for deploying and maintaining one versus three models. PD patients' overall low results might partly be explained by their high class imbalance regarding dropouts and completers.

Two further interrelated key findings are the importance of independent test sets and risk of overfitting on small datasets. If the decision about whether to pool the data was made on the training CV scores, single intervention runs would have been preferred. Further, even with pooled data, in two out of three interventions the small datasets seemingly outperform the much larger datasets in the training score. This aligns with Sajjadian et al.'s[30] findings that dataset size is significantly negatively correlated to the reported prediction accuracy. Our study's large test sets of 327–577 patients provides evidence that these good training results are biased as they fail to generalize. Sajjadian et al.[30] further find that many studies do not even use an adequate training set up, instead relying on a single train-test split. As can be seen in the box plots, this can result in extremely high or low results, neither of which represent the expectable prediction performance. Making a deployment decision on such ungeneralizable training results comes with a myriad of problems: risk of suboptimal care, wasted resources and ultimately the corrosion of trust in the use of ML in clinical care.[7,30,63] As this article shows, pooling different interventions enables providers to mitigate at least some of the risks when presented with the challenge of limited data availability.

The article, thus, contributes to e-mental health care by exploring the trade-off between data heterogeneity and dataset size and discussing the risk of overfitting.[30]

Limitations

At the same time, several limitations apply. For one, the routine care data in this study only includes self-referred patients, which leaves it unclear if the insights generalize to different patient selection methods. Further, it is yet to be investigated if the similarities between patients translate to the same clinical actions against dropout being effective. Third, using k-means for the clustering analysis is an industry standard,[58] but generative,[60] or density-based methods[69] may allow different insights. For the prediction task, the arguably biggest challenge in scaling the proposed approach is the availability of comparable interventions. While differing in content, the interventions at hand have a lot in common, the technical platform, the structure of treatments, the clinical routines for referral, assessment, therapist support, and the clinical staff. Therefore, our results do not warrant any prognosis about how the absence of these similarities would affect results. Lastly, this article neither compares the gains of pooled data to other options such as federate learning,[33] nor offers definitive insights on what minimal dataset size is necessary to produce generalizable results. In the end, pooling data in the proposed way is only one possible tool in the attempt to produce more generalizable and useful prediction models in psychological research.

Conclusion

Using ML to improve mental health care is a promising and growing research field. However, the lack of large datasets available hamper generalizability and cause biased results. This article addresses this issue by investigating the effects of pooling data from different interventions together to increase the training dataset size available.

A total of 6418 routine care patients' data from ICBTs for depression, SAD, and PD is used to (1) investigate heterogeneity in patient online behavior between interventions and (2) analyze the benefits of data pooling when predicting intervention dropout. Regarding the first question, the cluster analysis suggests three intervention-overarching groups that are defined more by their online behavior and other clinical characteristics than by which ICBT-program they are in. The finding that patients across the three interventions have similar behavioral patterns is further supported in the prediction results. Ultimately, data pooling

doubles the number of results that reach the threshold of clinical usefulness on the test set results. We, therefore, answer the second research question by concluding that data pooling is the superior approach based on our dataset.

Acknowledgement: We want to thank Silvan Hornstein for his valuable feedback on the research question and manuscript.

Contributorship: VK and NHI were involved in protocol development, gaining ethical approval and the organization of the data collection. KZ researched literature and derived the research question. NHI and KZ did the data preprocessing. KZ conducted the analysis and predictions; and wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Code availability: The underlying code for the model training can be made available upon request from the corresponding author.

Data availability: The datasets analyzed in the current study are not publicly available due to the sensitivity of health data and data privacy requirements. An artificial sample to give insights into the data structure can be made available from the corresponding author on reasonable request.

Declaration of conflicting interests: The authors declared no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval and consent: The Swedish Ethical Review Authority, the national governmental body responsible for all research-related ethical applications in Sweden, has approved the study Stockholm (DNR: 2011/2091-31/3, amendment 2016/21-32, 2017/2320-32 and 2018/2550-32), and the opt-out consent routine that is used for all patients at the routine healthcare service including the Internet Psychiatry Unit. This routine bases on the Swedish regulation that pseudonymized healthcare data may be used for research. The patients were provided with the written information how to opt-out in case they want to exercise this right.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Vetenskapsrådet, Avtal om Läkarutbildning och Forskning Agreement, Familjen Erling-Perssons Stiftelse, Fredrik och Ingrid Thurings Stiftelse. This open-access publication was funded by the German Research Foundation (DFG).

Guarantor: VK.

ORCID iD: Kirsten Zantvoort https://orcid.org/0000-0001-9876-054X

Supplemental material: Supplemental material for this article is available online.

References

1. Ebert DD, Harrer M, Apolinário-Hagen J, et al. Digital interventions for mental disorders: key features, efficacy, and potential for artificial intelligence applications. In: Kim YK (ed.) *Frontiers in psychiatry*. Singapore: Springer Singapore, 2019, pp.583–627. (Advances in Experimental Medicine and Biology; vol. 1192).
2. Becker D, van Breda W, Funk B, et al. Predictive modeling in e-mental health: a common language framework. *Internet Interv* 2018; 12: 57–67.
3. The Lancet Global Health. *Mental health matters [Internet]*. Amsterdam, Netherlands: Elsevier Ltd., 2020 [cited 2023 Feb 14]. Available from: https://doi.org/10.1016/S2214-109X(20)30432-0
4. Lamo Y, Mukhiya SK, Rabbi F, et al. Towards adaptive technology in routine mental health care. *Digital Health* 2022; 8.
5. Cuijpers P, Karyotaki E, Weitz E, et al. The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *J Affect Disord* 2014; 159: 118–126.
6. Bremer V, Chow PI, Funk B, et al. Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: machine learning approach. *J Med Internet Res* 2020; 22: e17738.
7. DeMasi O, Kording K and Recht B. Meaningless comparisons lead to false optimism in medical machine learning. Jan YK, editor. *PLoS ONE* 2017; 12: e0184604.
8. Hornstein S, Zantvoort K, Ulrike L, et al. Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms. *Front Digital Health* 2023; 5: 1170002.
9. Donkin L, Christensen H, Naismith SL, et al. A systematic review of the impact of adherence on the effectiveness of e-therapies. *J Med Internet Res* 2011; 13: e52.
10. Kaltenthaler E, Sutcliffe P, Parry G, et al. The acceptability to patients of computerized cognitive behaviour therapy for depression: a systematic review. *Psychol Med* 2008; 38: 1521–1530.
11. Baumeister H, Reichler L, Munzinger M, et al. The impact of guidance on Internet-based mental health interventions—a systematic review. *Internet Interv* 2014; 1: 205–215.
12. Forsell E, Jernelöv S, Blom K, et al. Clinically sufficient classification accuracy and key predictors of treatment failure in a randomized controlled trial of Internet-delivered Cognitive Behavior Therapy for insomnia. *Internet Interv* 2022; 29: 100554.
13. Forsell E, Jernelöv S, Blom K, et al. Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: a single-blind randomized clinical trial with insomnia patients. *Am J Psychiatry* 2019; 176: 315–323.
14. Barrett MS, Chua W-J, Crits-Christoph P, et al. Early withdrawal from mental health treatment: implications for psychotherapy practice. *Psychother Theory Res Pract Train* 2008; 45: 247–267.
15. Wu MS, Chen SY, Wickham RE, et al. Predicting non-initiation of care and dropout in a blended care CBT intervention: Impact of early digital engagement, sociodemographic, and clinical factors. *Digital Health* 2022; 8: 205520762211 33760.

16. Pedersen DH, Mansourvar M, Sortsø C, et al. Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. *J Med Internet Res* 2019; 21: e13617.
17. Wallert J, Gustafson E, Held C, et al. Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the U-CARE heart randomized controlled trial. *J Med Internet Res* 2018; 20: e10754.
18. Cote-Allard U, Pham MH, Schultz AK, et al. Adherence forecasting for guided internet-delivered cognitive behavioral therapy: a minimally data-sensitive approach. *IEEE J Biomed Health Inform* 2022; 27: 1–12.
19. Linardon J, Fuller-Tyszkiewicz M, Shatte A, et al. An exploratory application of machine learning methods to optimize prediction of responsiveness to digital interventions for eating disorder symptoms. *Int J Eat Disord* 2022; 55: 845–850.
20. Smink WAC, Sools AM, Postel MG, et al. Analysis of the emails from the Dutch web-based intervention "Alcohol de Baas": assessment of early indications of drop-out in an online alcohol abuse intervention. *Front Psychiatry* 2021; 12: 575931.
21. Moshe I, Terhorst Y, Paganini S, et al. Predictors of dropout in a digital intervention for the prevention and treatment of depression in patients with chronic back pain: secondary analysis of two randomized controlled trials. *J Med Internet Res* 2022; 24: e38261.
22. Bzdok D and Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2018; 3: 01643933.
23. Symons M, Feeney GFX, Gallagher MR, et al. Machine learning vs addiction therapists: a pilot study predicting alcohol dependence treatment outcome from patient data in behavior therapy with adjunctive medication. *J Subst Abuse Treat* 2019; 99: 156–162.
24. Lee Y, Ragguett R-M, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *J Affect Disord* 2018; 241: 519–532.
25. Bzdok D and Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*, 2018; 3: 223–230.
26. Lateh MA, Kamilah Muda A, Yusof ZIM, et al. Handling a small dataset problem in prediction model by employ artificial data generation approach: a review. *J Phys* 2017; 892: 012016.
27. van Smeden M, Moons KGM, de Groot JAH, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res* 2019; 28: 2455–2474.
28. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 1998; 10: 1895–1923.
29. Pasini A. Artificial neural networks for small dataset analysis. *J Thorac Dis* 2015; 7: 953–960.
30. Sajjadian M, Lam RW, Milev R, et al. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. *Psychol Med* 2021; 51: 2742–2751.
31. Aafjes-van Doorn K, Kamsteeg C, Bate J, et al. A scoping review of machine learning in psychotherapy research. *Psychother Res* 2021; 31: 92–116.
32. Carlbring P, Andersson G, Cuijpers P, et al. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cogn Behav Ther* 2018; 47: 1–18.
33. Loftus TJ, Ruppert MM, Shickel B, et al. Federated learning for preserving data privacy in collaborative healthcare research. *Digital Health* 2022; 8: 20552076221134455.
34. Beard C, Millner AJ, Forgeard MJC, et al. Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychol Med* 2016; 46: 3359–3369.
35. Aziz M, Erbad A, Belhaouari SB, et al. Who uses mHealth apps? Identifying user archetypes of mHealth apps. *Digital Health* 2023; 9: 20552076231152175.
36. Chien I, Enrique A, Palacios J, et al. A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions. *JAMA Netw Open* 2020; 3: e2010791.
37. Titov N, Dear B, Nielssen O, et al. ICBT in routine care: a descriptive analysis of successful clinics in five countries. *Internet Interv* 2018; 13: 108–115.
38. El Alaoui S, Hedman E, Kaldo V, et al. Effectiveness of Internet-based Cognitive-Behavior Therapy for social anxiety disorder in clinical psychiatry. *J Consult Clin Psychol* 2015; 83: 902–914.
39. Hedman E, Ljótsson B, Kaldo V, et al. Effectiveness of Internet-based cognitive behaviour therapy for depression in routine psychiatric care. *J Affect Disord* 2014; 155: 49–58.
40. Hedman E, Ljótsson B, Rück C, et al. Effectiveness of Internet-based Cognitive Behaviour therapy for panic disorder in routine psychiatric care. *Acta Psychiatr Scand* 2013; 128: 457–467.
41. Houck PR, Spiegel DA, Shear MK, et al. Reliability of the self-report version of the panic disorder severity scale. *Depress Anxiety* 2002; 15: 183–185.
42. Baker SL, Heinrichs N, Kim H-J, et al. The liebowitz social anxiety scale as a self-report instrument: a preliminary psychometric analysis. *Behav Res Ther* 2002; 40: 701–715.
43. Montgomery SA and Åsberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry J Ment Sci* 1979; 134: 382–389.
44. Svanborg P and Åsberg M. A new self-rating scale for depression and anxiety states based on the Comprehensive Psychopathological Rating Scale. *Acta Psychiatr Scand* 1994; 89: 21–28.
45. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The mini-International neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 1998; 59: 22–33. quiz 34-57.
46. Zantvoort K, Scharfenberger J, Boß L, et al. Finding the best match—a case study on the (text-) feature and model choice in digital mental health interventions. *J Healthc Inform Res* 2023; 7: 447–479.
47. Beintner I, Vollert B, Zarski A-C, et al. Adherence reporting in randomized controlled trials examining manualized multisession online interventions: systematic review of practices and proposal for reporting standards. *J Med Internet Res* 2019; 21: e14181.
48. Furukawa TA, Katherine Shear M, Barlow DH, et al. Evidence-based guidelines for interpretation of the Panic Disorder Severity Scale. *Depress Anxiety* 2009; 26: 922–929.

49. von Glischinski M, Willutzki U, Stangier U, et al. Liebowitz Social Anxiety Scale (LSAS): optimal cut points for remission and response in a German sample. *Clin Psychol Psychother* 2018; 25: 465–473.
50. Fantino B and Moore N. The self-reported Montgomery-Asberg Depression Rating Scale is a useful evaluative tool in major depressive disorder. *BMC Psychiatry* 2009; 9: 26.
51. Karin E, Dear BF, Heller GZ, et al. Measurement of symptom change following web-based psychotherapy: statistical characteristics and analytical methods for measuring and interpreting change. *JMIR Ment Health* 2018; 5: e10200.
52. Donkin L, Hickie IB, Christensen H, et al. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. *J Med Internet Res* 2013; 15: e231.
53. Karyotaki E, Kleiboer A, Smit F, et al. Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. *Psychol Med* 2015; 45: 2717–2726.
54. McKinney W. *Data structures for statistical computing in Python*. In Austin, Texas: Python in Science, 2010 [cited 2024 Jan 4]. pp.56–61. Available from: https://conference.scipy.org/proceedings/scipy2010/mckinney.html
55. Harris CR, Millman KJ, Van Der Walt SJ, et al. Array programming with NumPy. *Nature* 2020; 585: 357–362.
56. Satopaa V, Albrecht J, Irwin D, et al. Finding a "kneedle" in a haystack: detecting knee points in system behavior. In: 2011 31st International conference on distributed computing systems workshops. Minneapolis, MN, USA: IEEE; 2011. p.166–171.
57. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–2830.
58. Jain AK, Murty MN and Flynn PJ. Data clustering: a review. *ACM Comput Surv* 1999; 31: 264–323.
59. Sinaga KP and Yang MS. Unsupervised K-means clustering algorithm. *IEEE Access* 2020; 8: 80716–80727.
60. Hastie T, Tibshirani R and Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York, NY: Springer, 2017, 745 p.(Springer series in statistics).
61. Bholowalia P and Kumar A. EBK-means: a clustering technique based on elbow method and K-means in WSN. *Int J Comput Appl* 2014; 105: 17–24.
62. James G, Witten D, Hastie T, et al. *An introduction to statistical learning: with applications in R*. New York, NY: Springer US, 2021, (Springer Texts in Statistics).
63. Cabitza F and Campagner A. The need to separate the wheat from the chaff in medical informatics. *Int J Med Inf* 2021; 153: 104510.
64. Cortes C and Vapnik V. Support-vector networks. *Mach Learn* 1995; 20: 273–297.
65. Schapire RE. Explaining AdaBoost. In: Schölkopf B, Luo Z and Vovk V (eds) *Empirical inference*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp.37–52.
66. Cawley GC and Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2017; 11: 2079–2107.
67. Bates S, Hastie T and Tibshirani R. Cross-validation: what does it estimate and how well does it do it? [Internet]. arXiv; 2022 [cited 2023 Feb 14]. Available from: http://arxiv.org/abs/2104.00673
68. Fan RE, Chang KW, Hsieh CJ, et al. LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 2008; 9: 1871–1874.
69. Kotu V and Deshpande B. *Predictive analytics and data mining: concepts and practice with RapidMiner*. 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2014, 446 p.

Journal of Healthcare Informatics Research
https://doi.org/10.1007/s41666-023-00148-z

RESEARCH ARTICLE

Finding the Best Match — a Case Study on the (Text-) Feature and Model Choice in Digital Mental Health Interventions

Kirsten Zantvoort[1] · Jonas Scharfenberger[1] · Leif Boß[2] · Dirk Lehr[2] · Burkhardt Funk[1]

Received: 30 September 2022 / Accepted: 29 August 2023

Abstract

With the need for psychological help long exceeding the supply, finding ways of scaling, and better allocating mental health support is a necessity. This paper contributes by investigating how to best predict intervention dropout and failure to allow for a need-based adaptation of treatment. We systematically compare the predictive power of different text representation methods (metadata, TF-IDF, sentiment and topic analysis, and word embeddings) in combination with supplementary numerical inputs (socio-demographic, evaluation, and closed-question data). Additionally, we address the research gap of which ML model types — ranging from linear to sophisticated deep learning models — are best suited for different features and outcome variables. To this end, we analyze nearly 16.000 open-text answers from 849 German-speaking users in a Digital Mental Health Intervention (DMHI) for stress. Our research proves that — contrary to previous findings — there is great promise in using neural network approaches on DMHI text data. We propose a task-specific LSTM-based model architecture to tackle the challenge of long input sequences and thereby demonstrate the potential of word embeddings (AUC scores of up to 0.7) for predictions in DMHIs. Despite the relatively small data set, sequential deep learning models, on average, outperform simpler features such as metadata and bag-of-words approaches when predicting dropout. The conclusion is that user-generated text of the first two sessions carries predictive power regarding patients' dropout and intervention failure risk. Furthermore, the match between the sophistication of features and models needs to be closely considered to optimize results, and additional non-text features increase prediction results.

Keywords Precision psychiatry · Health care analytics · Machine learning · Natural language processing · E-mental health

Extended author information available on the last page of the article

Published online: 18 September 2023

1 Introduction

Estimates suggest that even before 2020, only a third of people affected by mental health problems received the help they needed [1, 2]. This unmet need is accelerated by the psychological aftermath of the COVID-19 crisis, with estimated growth rates in the prevalence of major depression and anxiety disorders of more than 25% [3]. Consequently, offering effective help on a larger scale is of paramount importance for individuals and, considering costs and devastating impacts, for societies as a whole [4].

Digital mental health interventions (DMHIs) help to provide psychological treatment as they are easily accessible, economical, and scalable [5]. DMHIs pursue related goals to face-to-face therapy but are conducted through the means of online education formats. They mainly consist of self-help texts, video or audio manuals, and exercises and can be accessed independently of time and location. DMHIs can be unguided self-help interventions or can include guidance by e-coaches, for example, via calls or messages [6]. Meta-analyses demonstrate their efficacy in treating various mental health problems like stress [7] and stress-related disorders such as depression [8, 9] and anxiety [10]. At the same time, to be effective, a participant must finish at least a certain amount of the intervention to show health benefits [11, 12]. However, it is estimated that in unguided DMHIs, three out of four participants drop out too early. At one in three participants, the odds are better, yet still problematic, in guided DMHIs [13]. Such dropout is a key factor identified for participants' variance in response rates, causing Gan et al. [12] to call for strategies to help those who struggle. Measures such as e-coaches' guidance, reminders, and personalization positively influence overall completion rates and health outcomes [12, 14, 15]. However, the extent of guidance necessary differs among individuals and many complete and benefit from interventions with little or none of the usually costly support. Hence, in order to optimally allocate the limited resources and effectively help as many as possible, participants in need of attention must be identified [16].

Machine learning (ML) models can make individual predictions and have previously been used to estimate intervention dropout and failure probabilities [17]. Most of these attempts focus on user journey data, including log-in data and other indicators of online behavior [18–20]. At the same time, human language is the primary tool in psychiatry and psychology [21, 22]. Accordingly, DMHIs often include asynchronous text-driven communication with participants, generally involving (1) open-text intervention exercises and (2) direct communication with e-coaches [23]. Such texts are known to hold valuable information regarding a user's mental state and intentions that e-coaches can use to best support their participants [24]. Extracting information from these texts is a promising but time-consuming task and thus poses a major challenge with respect to scalability. Natural language processing (NLP) is a field of computer science specifically designed to handle text data. Using NLP methods to automate or augment parts of the e-coaches' work is a largely unexplored field of research [17]. First advances train ML models on the users' text to predict binge eating behavior [25]

as well as intervention outcomes for social anxiety [26], and depression interventions [27]. NLP methods are ample and differ in both their complexity and their requirements. Obtaining descriptive numbers (e.g., length of the text) and simple counts of words (i.e., bag-of-words approaches) is straightforward from a technical point of view. However, the amount of human decision-making and manual pre-processing is high, and the contextual meaning captured is essentially non-existent. Word embeddings based on neural networks can account for the context of words [28] and have set various NLP prediction performance benchmarks outside of DMHI text data [29]. However, first applications to intervention text data are disenchanting when paired with simple classifiers [25, 27]. With some results worse than random chance, Gogoulou et al. [27] conclude that "the task of predicting treatment outcome based on patient text is very difficult" [27, p. 578]. These results notwithstanding, word embedding features are successfully combined with more complex ML models in the related field of mental health diagnostics [30, 31]. As these conflicting findings show, deciding on a suitable combination of text representation techniques and ML models remains a largely unexplored problem in DMHIs. In addition, the predictive power of newer deep learning models, such as bidirectional encoder representations from transformers (BERT) [29] is yet to be explored in the context of intervention text data. Beyond the issues discussed thus far, Funk et al. [25] point out that the isolated investigation of text data overlooks the likely interaction with non-text features such as the age of participants – a hypothesis supported by several other authors' findings [26, 32, 33]. Hence, the main motives driving this research are (1) the open question of how to best combine automated text analysis with non-text features to optimize resource allocation in DMHIs, (2) the hypothesis that previous performances of word embeddings in DMHIs are limited by the subsequent classification models used, not the word embeddings themselves, and (3) the proposition that a BERT model pre-trained on a general corpus will have predictive power in the intervention setting as well.

Joining the rising efforts of ML applications and automatization in the health sector [34, 35], we tackle the problem of machine-learning-aided decision-making in E-mental health research. Within this research area, our clinical application is optimizing resource allocation to relieve an overstrained system by identifying those that most need additional support. The findings of our case study on 849 participants allow for the derivation of more concrete hypotheses for the further investigation of empirical generalization [36]. More precisely, our contribution is threefold: First, we systematically compare the predictive power of different text representation methods (i.e., metadata, TF-IDF, topic analysis, sentiment analysis, and word embeddings) in combination with supplementary numerical inputs (socio-demographic, evaluation, and closed-question data) for intervention dropout and failure. We complement related work by investigating which ML model types — ranging from linear to sophisticated deep learning models — are best suited for different features and outcome variables. Second, we account for the relatively long and sequential input texts by designing a task-specific neural network architecture which (in many settings) outperforms existing word embedding approaches on intervention text. Third, we demonstrate the potential of BERT models [29] pre-trained on generic text corpora

in dropout prediction. To this end, this paper is structured as follows: we summarize related work (Chapter 2), describe our research approach (Chapter 3), present the text representation techniques (Chapter 4) and ML models (Chapter 5), and thoroughly evaluate different combinations of text representation methods and ML models (Chapter 6). Finally, we discuss the limitations of our study and highlight future research directions (Chapter 7).

2 Background

In medical research, the number of ML applications has greatly increased in recent years as they promise improved care, scalability, and cost efficiency [37]. Such improvements are particularly needed in mental health care, where patients often go undiagnosed [38], and require long-time monitoring and care [39]. While many data types (e.g., log-in or questionnaire data) are available [17], text data presents itself as a propitious option in a field that has always primarily relied on language for diagnosis and treatment [21, 25, 40]. Several research branches emerged to leverage text data's vast occurrence in the context of mental health [33]. As their nature, accessibility, and use significantly differ, Becker et al. [40] call for differentiation between research on *pre*-intervention and intervention data. This chapter briefly explains both and outlines related work to derive the research gaps addressed in this study.

2.1 Pre-Intervention Text Data

Pre-intervention data is gathered before and, thus, outside of a clinical intervention. Use cases focus on diagnosing mental health disorders and generating insights. For this purpose, much attention has been placed on social media data [41–44]. These datasets gather users' natural communication with each other on platforms like Twitter or Reddit. One example are Cohan et al. [31], who tackle a multi-class diagnosis problem on a dataset of 20.406 self-reportedly diagnosed and 335.952 control users' social media posts. They find that sequential neural network approaches outperform their non-sequential models trained on Term-frequency Inverse Document Frequency (TF-IDF) [45] features in eight out of nine conditions. Yeruva et al. [44] compare insights on obesity and healthy eating — topics related to eating disorders [46] — from 103.609 Tweets versus 6.602 academic abstracts from PubMed. They propose a pipeline to construct social and contextual word embeddings, which produce valuable insights. Wongkoblap, Vadillo & Curcin [47] predict depression diagnoses for 4.169 Twitter users. On the one hand, they compare the dictionary-based Linguistic Inquiry and Word Count (LIWC) tool [48], a language model, topic analysis, and Usr2Vec [49] features paired with logistic regression (LR) or support vector machines (SVMs). On the other hand, they pair word embeddings with a one-dimensional convolutional neural network (CNN), as well as two task-specific (attention-based) neural network architectures. At AUCs of 0.91–0.93, their sequential models outperform their non-sequential models with AUCs of 0.79–0.88. They

explain this gap with the information loss non-sequential models suffer when features are aggregated across words. More recent studies in Mental Health diagnostics go one step further by using a more novel pre-trained BERT model [29], which yields good results [35, 50] and thus shows promise for other areas of text data in E-Mental Health research. As much more work exists than can be discussed here, reviews such as [20, 33], or [41] can be referred to for a more detailed picture.

While pre-intervention text data is usually publicly and easily accessible on large scales (e.g., through crawlers), it lacks health labels such as a reliable clinical diagnosis and must depend on self-published information. Further, anonymity and limited information verification options can cause issues with data quality [33, 47]. In summary, pre-intervention text data was produced in a non-clinical setting and primarily generates diagnoses and epidemiological insights.

2.2 Intervention Text Data

In contrast, intervention data comes from a clinical setting designed to help an already diagnosed user. Here, text is produced by health staff (e.g., Electronic Health Records [41, 51]) or by the users themselves [40]. In DMHIs, users primarily produce answers to open-text questions or conversation data with health staff. Because of the controlled setting, high-quality socio-demographic, longitudinal symptom, and user behavior data is usually available. However, gathering intervention data requires resource-intensive steps such as screening, diagnosis, and the assurance of weeks-long (guided) interventions. Consequently, such data points tend to be costly, and data sets stay small [52]. Additionally, access to existing datasets is extremely limited due to privacy concerns [6, 12]. As a result, Shatte, Hutchinson and Teague [17] find that only 1% of studies investigating ML in a mental health setting investigate intervention data, and barely any consider NLP methods. In agreement with these findings, several authors conclude that NLP on intervention data is vastly understudied despite its substantial potential [17, 25, 33, 41].

In mental health interventions, lack of adherence and responsiveness to treatment are major concerns [6, 11, 53]. As shown by Forsell et al. [16], Pedersen et al. [19], and Pihlaja et al. [54], targeted measures such as human guidance can improve upon these problems but, in an already overstrained system, cannot be offered to all participants. Here, supervised ML models provide great value by identifying those users that require additional care and allowing for individually targeted measures [16]. To present a comprehensive picture of previous work of NLP for dropout and intervention failure prediction, we search PubMed with the query (“Natural Language Processing” OR “NLP”) AND (“Psychology” OR “Psychiatry” OR “DMHI*”) AND (“Predict*” OR “Machine Learning”) AND (“Outcome” OR “Dropout” OR “Adherence”). We include papers that used ML models to make individual dropout or outcome predictions based on user-generated open-text data in DMHIs. We then follow the citations in the related work section for more relevant papers. Furthermore, a PubMed search including a similar query with the term “BERT” did not lead to any studies including user-generated intervention data.

Howes et al. [32] predict intervention outcomes based on chat data between therapists and 167 English-speaking users of a depression and anxiety intervention. Simple LR, linear SVMs, and decision tree (DT) models are trained for classification. The authors conclude that a combination of demographic and metadata yields better results than the slightly more sophisticated sentiment and topic analysis. The best-reported f1 measure improves the baseline from 0.57 to 0.7. However, as they point out, they split several messages of one patient between test and training set in their 10-fold cross-validation. With limited patients available, the combination of age, gender, and therapist can already allow a model to identify an individual participant and infer the result from the training example.

Hoogendorn et al. [26] retrieve information about sentiment, topics, writing style, and word usage from German emails written by 69 social anxiety patients, together with meta and demographic data. They investigate (1) averages and (2) trends per person. They choose the 20 features most correlated with their outcome variable — symptom levels at week 12 — mainly covering single words (17), topics (2), and writing style (1). For classification, they train LR, DT, and random forest (RF) models, arguing that these model types give reasonably good and understandable results. While socio-demographic data alone has no predictive value, complementing it with text data up to week six significantly enhances the prediction performance of their RF model (AUC 0.83).

Smink et al. [55] use 770 participants' first four out of an average of 20 emails written in a DMHI for alcohol abuse to predict dropout. They retrieve word count and LIWC [48] features and combine them with socio-demographic data. The classifiers used are LR, a neural network, XGBoost, and a Mixed Effect RF model. First, they aggregate the features as means across all four emails for the non-sequential models. Second, they input the features per email into their sequential neural network and RF. Hence, while sequential models are included, they only consider the order of emails, not the sequentially of language itself. The winning XGBoost model performs worse than their baseline, leading to the conclusion that they could not associate their simple email text features with intervention dropout.

Funk et al. [25] use 372 participants' English messages and intervention text snippets to predict binge eating episodes in the next 24 h. A total of 100 of these participants also have the 6-month follow-up health outcome. The authors compare an array of different methods of text representation: metadata, bag-of-words models including topic and sentiment scores, word embeddings, and Part-of-Speech tagging. To predict short-term symptom severity, they train an LR and an RF model, resulting in a maximum AUC of 0.57 for new users. Additionally, they use LASSO regression to determine the best out of their 220 variables for the long-term outcome prediction. None of the 50 element-wise averaged embedding dimensions are among the most informative features.

Gogoulou et al. [27] compare TF-IDF, Word2Vec, FastText [56], and Doc2Vec [57] text representation on Swedish homework reports of 1.986 users of a depression intervention. The three word embeddings are trained in advance on an additional 4.835 users' texts from other interventions. In their approach, TF-IDF outperforms the word embeddings in almost all settings, and in some cases, the latter even perform worse than the naïve baseline. With a maximum f1 score of 0.69 (baseline

of 0.58), they conclude there is a signal in intervention text data regarding outcome prediction, but word embeddings do not serve to extract it. While their paper has the by far largest sample for intervention text data and considers three different methods of word embeddings, it only uses a simple linear classifier. As such, they do not put their focus on leveraging the sequential nature of word embeddings [58].

In conclusion, only one paper investigates the prediction of dropout based on intervention text data, with little success. However, as the authors propose, features other than the two simple ones included should be investigated [55]. For outcome prediction, several studies find that combining text features with non-text features such as socio-demographic data leads to the best results [26, 32, 33]. This results in the first research focus of this paper presented in the introduction; the question of how to best combine text analysis with non-text features to optimize resource allocation in DMHIs. The works so far suggest that simpler text representation features are superior in their predictive performance. However, datasets were almost always smaller than 250 users, and the focus has been on linear and simpler tree-based classifiers. Those papers including more sophisticated models, only used simple features. Thus, the performance of more sophisticated models, such as ensemble methods and deep neural network classifiers in combination with complex features, remains to be investigated in typical DMHI prediction tasks. This leads us to our second research proposition: Previous performance of word embeddings in DMHIs is limited by the subsequent classification models used, not the word embeddings themselves. Further, successful examples from research on pre-intervention data [35, 50] let us arrive at our third proposition for this paper: That a BERT model pre-trained on a general corpus will have predictive power in the intervention setting as well.

3 Study Set-Up

This study addresses the gap in existing research by systematically exploring the predictive power of text (i.e., different metadata, TF-IDF, sentiment and topic analysis, Word2Vec and FastText word embeddings) and non-text data types (i.e., socio-demographic and symptom data, evaluation data, and closed-question data) and their interplay with different model types (i.e., LR, SVMs, XGBoost, AdaBoost, LSTMs, and BERT). We investigate these results for intervention dropout and outcome to provide insights into the use of ML methods to optimize resource allocation. The final goal is better outcomes with equal or lower costs [16, 19]. A key focus of this paper is the investigation of the gap between the word embeddings' theoretical power and the lack of its manifestation when used on intervention text data. To this end, two different word embeddings are trained and then (1) averaged for non-sequential models and (2) used as they are with a sequential model. Furthermore, we employ BERT to make predictions based on the intervention text data, which — to the best of our knowledge — has not yet been investigated. At the same time, Occam's razor principle suggests that — ceteris paribus — the simplest model is preferable [59]. Because of this, feature extraction methods and models of different sophistication levels are pit against each other in this exploratory study of how

to best predict intervention failure and dropout. With 849 participants, the dataset at hand is larger than all but one of the previous works on intervention text.

3.1 Data Description

For our case study, we consider the data of 927 participants from six randomized controlled trials (Table 1) of an internet-based stress management intervention called GET.ON Stress [7]. The training program comprises seven sessions, planned to be held on a weekly schedule. Each session consists of general information, quizzes, audio and video files, downloadable worksheets, and interactive exercises. The interactive exercises are the most important element in each session. Users work through the exercises by reading or listening to short instructions and then writing their answers into text boxes. In subsequent sessions, many of the text inputs are picked up and displayed again to the user by the system. The core stress coping strategies included in the training program are problem-solving [60] and emotion regulation [61]. At the beginning of the program, participants write about their stressors, goals, and motivations. In each subsequent session, the participants are asked to choose pleasant activities, plan to implement them into their lives, and to reflect on how it went in the subsequent session. In the second and third sessions, participants learn a systematic six-step problem-solving method that can be applied to their problems, again reflecting on it in the subsequent sessions. In sessions four to six, participants learn and practice different emotion regulation techniques, such as muscle and breathing relaxation [61]. In the seventh session, participants reflect on their goals for the training and plan how to continue practicing stress coping in the future. Four weeks after completing session seven, an optional booster session eight is provided. Depending on the trial, participants went through the program as a self-help intervention, were able to ask for feedback, or automatically received written feedback by e-coaches after every session. For more detailed information on the set-up of the intervention and each of the studies, please refer to the primary publications cited in Table 1.

In this study, intervention dropout is defined as having finished less than the six core sessions out of eight total sessions. Sessions 7 and 8 are not considered core

Table 1 Overview of the intervention studies included in this analysis

Study	German clinical trials register no.	Publication	Level of human support
1	DRKS00004749	Heber et al. [62]	Intensive guidance[a]
2	DRKS00005112	Ebert et al. [63]	Guidance on demand[b]
3	DRKS00005384	Ebert et al. [64]	No guidance[c]
4	DRKS00005687	Nixon et al. [65]	Guidance on demand
5	DRKS00005990	Ebert et al. [66]	Guidance on demand
6	DRKS00005699	Nixon et al. [67]	No guidance

[a]participants receive written feedback (avg. 30 min) after each session; [b]participants receive feedback on demand; [c]participants receive technical support only

sessions as they do not convey new material but instead serve as a reflection and repetition session, respectively. As such, the dropout definition follows the consensus of operationalizing dropout reported by Donkin et al. [11] and is recommended to use by Gan et al. [12]. The second session is chosen as the point of prediction due to the trade-off between text gathered and time left to intervene [18]. Choosing this prediction point results in 849 German-speaking participants who completed exercises in the first two sessions — 25% of whom are considered dropouts. Intervention failure is defined as an improvement of fewer than 5.16 points on the Perceived Stress Scale (PSS) [68, 69], the primary health outcome metric. This threshold value of 5.16 is based on the reliable change index indicating a clinically meaningful change in symptomatology introduced by Jacobson and Truax [70]. The average baseline PSS score is 25 and, after finishing an average of 6.6 sessions, ends at 17. In total, 37% of users considered are intervention failures. A total of 40 participants did not fill out the PSS questionnaire after finishing the intervention and, therefore, cannot be considered for intervention failure prediction. Losing many data rows because participants did not fill out the final symptom questionnaire is a common problem when predicting intervention outcome. For example, Gogoulou et al. [27] disregard 38% of their participants because their low adherence prevents the calculation of the target features. Attempting to predict the 6-month follow-up, Funk et al. [25] even lose 73% of their data. In this dataset, from those with unknown outcomes, 85% dropped out. Excluding these participants runs the risk of ignoring those most in need of additional support. Therefore, we provide insights into both, dropout (keeping more participants) and intervention failure (the more exact outcome measure) predictions.

3.2 Non-Text Data

Related work suggests that a combination of text and non-text features is most promising when retrieving information about a user's mental state [23]. Thus, unsurprisingly, a myriad of the above-mentioned studies includes non-text variables in their analysis. We train benchmark models on each of the non-text and text feature types by themselves and then combine them to be able to differentiate between individual, and interaction effects.

Baseline variables such as socio-demographics or symptom data have been thoroughly investigated in terms of their predictive power for dropout and intervention failure, howbeit with limited consensus in results (e.g., [71, 72]). We include these variables in our analysis based on the assumption that ways of expressing oneself are dependent on users' characteristics such as age and gender [23]. Asking users to fill out a baseline questionnaire before starting the intervention is common, as seen in the related work section. Our eleven socio-demographic variables cover different information about the participants' age, gender, educational background (2 features), occupation (5), and family status (2). 77 participants did not indicate their income level, they are accounted for in an additional feature. The descriptive statistics and data types of all included socio-demographic features can be found in the supplementary material 1. The majority of participants identify as female (78%), hold a college degree (60%) and are on average 42 years old, where the age distribution

is bimodal with two peaks around 30 and 50 years. In addition to the socio-demographic variables, five symptom-related variables provide the baseline PSS subscores of *Helplessness* and *Self-Efficacy* [69], and carry information about previous experiences with training and therapy (3 features). The mean values of the PSS subscores before the intervention are 16 and 9, respectively. The aforementioned variables are supplemented by the intervention support level and an indicator of whether the user found out about the intervention via their health insurance company.

Evaluation data providing information on the user's attitude towards the intervention can easily be argued to be an evident factor for their intention to continue it. Therefore, this data proposes a promising alternative to the resource-intensive process of text-analysis. At the same time, it requires an additional questionnaire after each session, hence straining the limited user attention available. To investigate this trade-off, it will be included in the analysis. The users evaluate the (1) easiness and (2) usefulness of each session on a scale of 1 (very useful/very easy) to 5 (not useful at all/very difficult). Furthermore, the users were asked to estimate the time they needed to complete the respective session on a rating scale from 1 (less than 30 min) to 4 (more than 90 min). On average, users rate the easiness with 2.3, the usefulness at 1.8 and the time required between 30 and 90 min. Furthermore, users have the chance to articulate well-liked and improvable aspects of each session in an open-text format. For the text representation, we append this text to the rest of the user's generated text of the corresponding session. In total, 735 participants answered the evaluation questions for at least one of the first two sessions, and missing values are accounted for in an additional feature.

Closed question data is structured data in the form of questionnaire items that have a limited set of pre-defined answer options, which Cook et al. [73] found to have better performance than open-text questions when predicting suicidal intentions. Such closed questions are often inherent in the intervention design and are easier to handle than unstructured text data from a technical standpoint. Exemplary impressions of how the users saw such questions can be found in the supplementary material 2. In our dataset, three closed-form intervention exercises sum up to an additional 13.298 user entries. These questions address the perceived stress levels, the percentage of successfully implemented goals from the previous session, and the intended day of finishing the upcoming session. We extract the relevant numbers and — depending on the nature of the question — include them as they are or aggregate them (i.e., sums, averages, or counts). We fill missing values with 0 s and create additional features indicating missing values.

4 Text Representation

In total, the 849 users produced 61.290 open-text answers to intervention exercises and another 3.647 answers to the open-text evaluation questions. Given the point in time of the prediction, only the text from sessions one and two are used. This leaves 15.773 entries, 1.597 of which are open-text evaluation answers. As a first step, 1.064 entries that do not contain any relevant information (e.g., "xxx", "…", "-") are deleted, which are found via the investigation of the answers with less than

five characters. A feature counting the number of such entries is included in the simple metadata. Since text representation techniques typically cannot handle numbers well [25], digits are replaced by '#'. Second, we scrape a list of commonly used German abbreviations and manually adjust and supplement them to better fit the context of this intervention. The abbreviations are replaced with their long-form, and special characters, as well as smileys, are deleted. A spell check based on the Hunspell package is tried but does not increase cross-validation scores and, therefore, is not used in the final results. Third, we lemmatize the participants' text using the Python library SpaCy. As upper-case letters carry significant meaning in German [74], the texts are only lower-cased after lemmatization. Since bag-of-words methods usually benefit from lemmatized texts [74], while neural network approaches are not expected to [75], we keep both. Lastly, we aggregate the text per user and session, resulting in a concatenated string of all user text inputs that can be used as-is or be further aggregated across sessions 1 and 2.

4.1 Metadata

Especially when thinking about dropout as the binary manifestation of engagement, the effort invested in the exercises is a promising candidate for its prediction [5]. Assuming that a longer answer to a given task requires more effort than a short one, the arguably most straightforward measure is the length of the answer. Hence, we create a *simple metadata* representation of the participants' texts by measuring the word and character count. An average intervention text in sessions 1 and 2 together contain 617 words in 4.105 characters and an additional 48 words in 313 characters for the evaluation questions. To account for the participants' willingness to answer the intervention questions, a feature counting the number of *useless* (defined as above) entries is added. Additionally, the usage of upper cases, exclamation, question marks, and positively or negatively connoted smileys are counted before they are deleted in the text cleaning. The *advanced metadata* is based on Ewbank et al.'s [51] finding that different therapeutic intentions and topics have different impacts on outcomes in face-to-face intervention. In sessions 1 and 2, tasks aim to gather information about the user's motivation and build skills in problem-solving, stress analysis, behavior reflection, and behavioral planning. Thus, all text snippets are categorized, and text lengths per category are retrieved to investigate whether this additional information can improve predictions.

4.2 Bag-of-Words

Bag-of-words approaches count the occurrences of each word in a document (i.e., intervention answers) in an attempt to extract similarities or differences in texts. A popular bag-of-words method is *Term Frequency-Inverse Document Frequency* (TF-IDF) [45]. The word occurrence count is rescaled based on the relative occurrence of all documents. The scikit-learn TF-IDF vectorizer is used on the word level, considering uni- and bigrams to produce the vector per participant. This approach results in a very large and highly sparse matrix; both attributes that many ML

models cannot handle well. To reduce the size of the matrix, features used by more than 70% of documents are discarded, as they are assumed to be stop words. In order to keep fewer features than data points [25], the number of TF-IDF features kept is determined by the number of users minus the number of additional non-text features. In two additional steps, sentiment and topic analysis are used to reduce the matrix dimensionality by grouping similar words. *Sentiment properties* of polarity and subjectivity are retrieved per text snippet to extract variation in sentiments depending on the exercise (e.g., "What stressed you today?" vs "What makes you feel good"). The German version of the text blob package - a rule-based approach — is used on the lemmatized text, as per the recommendation of Fehle, Schmidt and Wolff [74]. Sentiment polarity is recorded on a scale from [-1,1], with the minimum indicating a negative and the maximum indicating a positive connotation. In addition, the subjectivity variable indicates the level of opinion, emotions, or judgments between 0 (objective) and 1 (subjective). As both the average sentiment and the range of sentiment are considered valuable information [25], the mean, max, and minimum scores across sessions are included as features. Another way of reducing the dimensions is *Latent Dirichlet Allocation (LDA),* which tries to identify latent topics in the documents. LDA assumes that a document touches upon different topics operationalized by a list of common words associated with each topic [76]. Considering the number of relatively small entries and the likely tendency that similar exercises produce similar answers, this step is done on the already aggregated text, and the number of topics considered is set to 10 [25]. The topic model is calculated on the training data corpus only and then applied to the test data text.

4.3 Embeddings

Based on the assumption that similar words appear in similar contexts, word embeddings attempt to analyze word co-occurrences and represent each word by *n*-dimensional vectors of real numbers. Thus, words used in akin contexts tend to be mapped to vectors with small distances. Word2Vec [58] and FastText [56] are frequently used word embedding techniques based on neural networks. Word2Vec offers two different network architectures to learn word representations by (1) predicting a current word based on its surrounding words (CBOW) or (2) predicting the surrounding words based on a current word (Skip-gram). While the learned representations of the words in the training corpus are mostly meaningful, unseen words cause difficulties. In order to find a vector representation of these words, a fraction of rare words is typically mapped to an out-of-vocabulary (OOV) token during training allowing unseen words to be mapped to this generic OOV vector. FastText [56] is an extension of Word2Vec, which tries to tackle the problem of unseen words by building embeddings for each word in the corpus as well as the n-grams each word consists of. Hence, word vectors for unseen words can be generated based on the n-grams in a more meaningful way. Both word embeddings can be trained from scratch on custom datasets or word embeddings pre-trained on large text corpora in languages (e.g., Wikipedia or News articles) can be used. Since the text produced by the study participants is different in its structure from generic corpora, we follow related work

[25, 27] and train the word embeddings on an extended dataset using the Gensim library. To enhance our small training dataset, we also use the texts generated by control group users and train the word embeddings at the sentence level. We treat the vector dimension n and the model architecture (i.e., CBOW or Skip-gram) as hyperparameters which are optimized during the training of our recurrent neural network (Section 5.2). To compare the sequential approach to results from related work [25, 27], we process the generated word embeddings by calculating the element-wise averages of every participant's text and use these averaged word embeddings as inputs for non-sequential models (Section 5.1).

5 Machine Learning Models

In the following, we present the different ML models that are trained to predict dropout and intervention failure. To match the complexity of the text representation methods, we consider three different model categories: (1) traditional ML models for non-sequential data, (2) deep learning models for sequential data, and (3) advanced pre-trained transformer-based models. While non-text features, meta-data, bag-of-words, and averaged word embeddings are combined with traditional ML models, we extend related work in this field by additionally maintaining the sequential nature of text by training recurrent neural networks as well as a BERT classification model. We set apart a hold-out test set (20% of the participants) beforehand to evaluate the models' out-of-sample performance (Chapter 6).

5.1 ML Models for Non-Sequential Data

We use four different classification models: LR, SVMs, AdaBoost, and XGBoost. The corresponding model hyperparameters are optimized in a fivefold cross-validation (CV), where each hyperparameter space is defined by initially choosing small intervals around the default values and incrementally considering adjustments if the boundaries perform best in the CV. For each data input (i.e., combinations of text representations and supplementary numerical inputs), one final model, chosen based on the CV scores, is trained on the entire training data, and evaluated on the hold-out test data. To account for the class imbalance in the dropout data, we create synthetic data of the minority classes by using SMOTE oversampling [77]. The sampling ratio is treated as a hyperparameter for all four models and is optimized during the CV.

Logistic regression as a linear model for binary classification is very popular due to its fast training times, good explainability, and reasonably good results. In light of the dataset size, the liblinear solver is chosen. Given the partially high number of predictors, L_1 or L_2 regularization are optimized as a hyperparameter in the CV, together with the respective penalization strength (0.01–10). *Support vector machines* classify by drawing decision boundaries between classes. SVMs can either use the feature space as is or use a non-linear kernel to map it into a higher dimensional space to make classes linearly separatable [78]. The use of a linear or a radial basis function kernel is optimized as a hyperparameter, each with their own

set of regularization parameters (C: 0.1–1000, gamma: 0.001–1) to balance over- and under-fitting. For both, LR and SVMs, a scaler is added to the ML pipeline. *XGBoost* is a fast and efficient implementation of a Gradient Boosting Tree that also allows for the regularization of features and thus avoids overfitting on smaller datasets [79]. As the XGBoost classifier has many non-trivial hyperparameters, Bayesian Search CV is used to allow a less computationally expensive grid search [80]. To constrain the architecture of the trees, the max. depth (3–5), and the minimum weight of a child (0.5–1) are optimized. Further measures against overfitting are the percentage of rows (0.5–1), and columns (0.5–1) used to build each tree, as well as the regularization parameters gamma (0/1) and lambda (1/2). The number of estimators (50–1000) is also investigated with the learning rate for each step (0.01–0.5). *AdaBoost* classifiers leverage the advantages of ensemble learning by combining a variety of weak learners to achieve better predictions [81]. The number of estimators (3–2000) used stands in a trade-off to the learning rate (0.001–2) — the weight given to each estimator — because of which these are optimized together. We implement our models in Python using the Scikit-learn and xgboost libraries. The non-sequential models can be trained on a standard laptop, and training times partially depend on the number of features. Including grid search, LR and SVMs usually need mere seconds while the AdaBoost model, on average, takes several minutes. Training times are the longest for the XGBoost models, where iterating through the entire hyperparameter space often takes longer than for the AdaBoost model, despite the use of Bayesian Search CV. Including the large number of TF-IDF features implies the longest training times at one or two hours each for the Ensemble models.

5.2 Recurrent Neural Network

Related work in this field demonstrates the inferior performance of word embeddings when element-wise averaged and used as inputs for models from the previous section [25]. Due to the relatively long input sequences in the second session (on average 370 words respectively 392 with evaluation texts), we assume that a carefully designed recurrent neural network can better leverage the potential of word embeddings than averaged word vectors and thus possibly achieve better results on our two classification tasks. To avoid enlarging the input sequence length further, we do not include text generated in the first session.

A naïve bidirectional LSTM-based [82, 83] model architecture, which consists of one input containing all text inputs of a given participant, barely achieves baseline performance on our validation set. This may be grounded in challenges arising from these long input sequences. Therefore, we decide to design a more sophisticated, task-specific model (Fig. 1) for our problem. The core model has four different blocks which aim to encode the participants' texts with respect to one of the four categories used in the second session — problem solving, reflection, stress analysis, and behavioral planning — and thus naturally reduces the input sequences' lengths. Each block consists of an input layer, an embedding layer (i.e., our pre-trained word embedding matrix), and two bidirectional LSTM layers. All outputs from the last bidirectional LSTM layers are concatenated and passed to a fully-connected neural

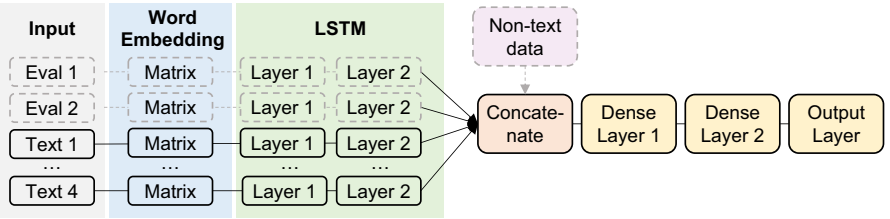


Fig. 1 Task-specific LSTM-based model architecture

network with dropout. Adding a further bidirectional LSTM layer after concatenation does not improve performance on our validation set. We consider the embedding dimension (FastText: 10, 25, 50, 100; Word2Vec: 25, 50, 100, 300), input sequence length (30, 50, 100, 200 words), number of units per LSTM layer (first layer: 0, 16, 32; second layer: 16, 32), number of neurons per dense layer (16, 32), and the dropout rates (0.1, 0.2) as hyperparameters which are optimized during training. If further text inputs are considered (i.e., evaluation texts), we extend our core model by two blocks processing the two different evaluation categories (i.e., feedback about liked contents and suggested improvements). If numerical inputs are considered (i.e., demographical data, numerical evaluation data, or extracted numbers from text), we extend our core model by another input layer which is normalized and directly passed to the concatenation layer. We try to account for the imbalanced class distribution by using a weighted binary cross-entropy loss function. The class weights are considered hyperparameters which are optimized during training. The Adam optimizer is used to train this network architecture where the learning rate (0.01, 0.001, 0.0005) yields the final hyperparameter. To tune all hyperparameters, we use 20% of the training data as a validation set and re-train our tuned models on the entire training set for 25 epochs with early stopping. Since the performance does not increase when fine-tuning the embedding layers, we freeze the embedding weights and only train the remaining weights of the network. The network is implemented in TensorFlow, and the hyperparameter tuning is executed on an Nvidia Tesla P100, which takes approximately six hours for each of the four different data inputs.

5.3 BERT

To represent the more complex recent transformer model architectures, we investigate the prominent "bidirectional encoder representations from transformers" (BERT) model [29] to predict dropout and intervention failure. In contrast to the previous approaches of separating the steps of text representation and training classification models, the BERT model combines these tasks. While training BERT from scratch requires a substantial amount of data, BERT models pre-trained on large datasets can be leveraged and are easily adaptable to new NLP tasks. On NLP benchmark tasks, pre-trained BERT models that are fine-tuned on custom datasets achieve better results than carefully crafted task-specific model architectures [19].

Therefore, we also follow this approach to both maintain the sequential structure of the texts and to reduce the manual effort in designing an appropriate architecture.

We build our classification model based on the BERT model pre-trained on three large German datasets ("bert-base-german-cased" from Huggingface's model repository) and fine-tune it on our dataset. To adapt this model to our two classification tasks, we slightly modify the model architecture: we use the 768-dimensional representation vector produced by the BERT model and feed it into a new classification head consisting of two hidden layers and a sigmoid output layer. When considering additional numerical inputs (i.e., baseline, evaluation, or closed question variables), we concatenate the 768-dimensional vector with the supplementary variables. The design of the classification head is optimized during training where the number of neurons per hidden layer (16, 32, 64) and the dropout rate (0.1, 0.2) are considered hyperparameters. Despite BERT's ability to handle input sequences up to 512 words, we only consider shorter lengths (64 and 128 words) due to the required computational resources. To compensate for the class imbalances, we make use of a weighted binary cross-entropy loss function and treat the class weights as additional hyperparameters. The aforementioned hyperparameters and the learning rate ($5 \cdot 10^{-4}$, 10^{-5}, $5 \cdot 10^{-5}$, 10^{-6}) of the Adam optimizer are optimized during training using a validation set of 20% of the training data. The final model is trained on the entire training data for 20 epochs with early stopping.

6 Results

We evaluate our final models on the test set of 170 (dropout) and 163 (intervention failure) participants. Although our test set is large when compared to most related work, this size still implies the risk of unrepresentative results. Since the area under the receiver operating characteristics (AUC) accounts for class imbalance [84] and thus eases the comparison of results of the two classification tasks, we choose this evaluation metric [85]. The two result tables for intervention failure (Table 2) and intervention dropout (Table 3) summarize the AUC scores on our test set, where columns represent different text representation methods and rows define supplementary non-text features. The benchmark model (BM) column provides a reference score trained exclusively on the corresponding numerical features. To identify the most predictive features, we calculate SHAP values [86] or use included feature importance measures for the non-sequential models.

6.1 Intervention Failure

Exclusively considering *text variables*, sentiment analysis (AUC of 0.65) outperforms the other text representation techniques on our test set. Other methods, such as Word2Vec combined with our LSTM architecture and advanced metadata with an LR model, achieve solid results (0.61 resp. 0.59) as well. While averaged word vectors combined with boosting classifiers perform very poorly (0.50–0.52), leveraging the sequential nature of the texts by using deep learning architectures

Table 2 Result table intervention failure prediction

AUC	BM	Sim. MD	Adv. MD	TF-IDF	Sentiment	LDA	W2V Avg	FT Avg	W2V NN	FT NN	BERT
Pure Text	0.500	0.546	0.588	0.545	0.649	0.550	0.500	0.522	0.605	0.553	0.550
Baseline	0.688	0.687	0.710	0.687	0.719	0.687	0.687	0.687	0.635	0.617	0.581
Eval	0.624	0.623	0.631	0.520	0.649	0.629	0.615	0.596	0.676	0.592	0.591
Closed-Q	0.529	0.524	0.578	0.554	0.577	0.534	0.508	0.530	0.572	0.623	0.564
Average	**0.614**	**0.595**	**0.627**	**0.577**	**0.649**	**0.600**	**0.578**	**0.584**	**0.622**	**0.596**	**0.572**

BM, baseline; *Sim./ Adv. MD*, simple or advanced metadata; *W2V*, Word2Vec; *FT*, FastText; *NN*, task-specific neural network; *Eval.*, evaluation data; *Closed-Q*, closed questions

Table 3 Result table intervention dropout prediction

AUC	BM	Sim. MD	Adv. MD	TF-IDF	Sentiment	LDA	W2V Avg	FT Avg	W2V NN	FT NN	BERT
Pure Text	0.500	0.651	0.541	0.626	0.559	0.547	0.531	0.610	0.644	0.630	0.618
Baseline	0.596	0.614	0.633	0.642	0.588	0.617	0.662	0.551	0.696	0.645	0.646
Eval	0.649	0.686	0.592	0.651	0.636	0.643	0.639	0.601	0.636	0.656	0.668
Closed-Q	0.584	0.554	0.563	0.652	0.599	0.606	0.575	0.632	0.639	0.663	0.668
Average	**0.610**	**0.626**	**0.582**	**0.643**	**0.596**	**0.603**	**0.610**	**0.607**	**0.654**	**0.649**	**0.650**

yields benefits (AUC 0.55–0.61). Thus, also performing equally or better than TF-IDF features with an AdaBoost model (0.55).

Our benchmark model (BM) trained on the numerical *baseline data* achieves an AUC score of 0.69 and, hence, is not outperformed by the vast majority of text representation techniques. This is most likely due to the initial PSS subscores included in the baseline data, which are expected to be important variables in predicting intervention failure [16]. While the additional baseline variables increase the performance of all text representation methods (compared to text-only models), only advanced meta-data and sentiment analysis (both combined with LR) achieve better AUC scores (0.71 and 0.72) than our baseline benchmark. In both cases, age is the most important feature, followed by the baseline PSS subscores and income category. While PSS subscores are among the five most important variables of our benchmark model as well, age and income are not, which possibly indicates a moderating function for text features. The deep learning approaches are the only approaches that do not benefit from adding baseline data and perform worse than the averaged word vectors combined with LR. Similar to baseline data, additional *evaluation data* (both textual and numerical) enhances the performance of nearly all representations. Besides the winning task-specific Word2Vec LSTM architecture (0.68), advanced metadata, sentiment analysis, and LDA (all using LR) achieve better results than the evaluation data benchmark by itself (0.62). TF-IDF, averaged word vector, and the remaining deep learning approaches cannot attain the benchmark scores, suggesting that more variables can have a harmful effect on the information-to-noise ratio. *Closed-question data* adds little value and, in some cases, even decreases the model performance when compared to text-only results. Only the task-specific FastText LSTM model leverages this additional information and achieves an AUC result of 0.62. While this clearly outperforms the benchmark (0.53) as well as the averaged word vectors on this task, various other approaches on different data inputs achieve better results.

To predict intervention failure, baseline data containing initial PSS subscores clearly benefits the models' performances. On our test set, sentiment analysis and advanced meta-data approaches yield solid results which perform better than benchmark models and other approaches considered. On average, advanced metadata (0.63) performs slightly better than simple metadata (0.60), offering evidence that the nature of the exercise done matters for the intervention outcome. Analyzing the model coefficients of our advanced metadata reveals that the largest coefficients are assigned to the text length of tasks concerning problem reflection, behavioral planning, and motivation. Since this model aims to predict rather than to explain, further research is necessary to investigate the causality. Among the two best-performing non-sequential approaches, LR and SVM are most frequently chosen (6 out of 8). BERT, TF-IDF (primarily used with boosting classifiers), and averaged word embeddings often perform below our benchmark. Although we demonstrate, on our test data, that word embeddings combined with our task-specific architecture often outperform the averaged word vector approaches, the deep learning models fail to achieve benchmark scores in many cases.

6.2 Dropout

Despite the theoretically assumed interrelation between dropout and intervention failure [11, 12, 87], well-performing text representation approaches and ML models differ significantly on our dataset. While TF-IDF and the deep learning approaches perform poorly in many settings when predicting intervention failure, these approaches, as well as the simple meta-data approach, dominate the results for dropout prediction. On pure *text data*, simple meta-data combined with a non-linear kernel SVM classifier yield the best AUC score (0.65), closely followed by TF-IDF combined with XGBoost (0.63) as well as the Word2Vec (0.64) and FastText (0.63) task-specific LSTM models. Word embeddings combined with our LSTM architecture increase performance in comparison to averaged word embeddings and an SVM classifier (0.53 and 0.61). Previously well-performing approaches such as advanced meta-data and sentiment analysis score mediocre results (0.54 resp. 0.56) on the task of dropout prediction. Akin to the intervention failure prediction, model performances' generally benefit from additional *baseline and evaluation variables*. Yet, for dropout prediction, evaluation data has a stronger impact, supporting the hypothesis that a participant's opinion on the intervention is a good predictor for discontinuation. The task-specific LSTM-based approach on the Word2Vec embeddings scores the best results (0.70) when using additional baseline variables, and other deep learning approaches also perform well (0.65) in this setting. TF-IDF features used with LR likewise achieve an AUC score (0.64) well above the benchmark (0.60) on this task. Simple meta-data combined with LR (0.69) and our fine-tuned BERT model (0.67) yield the best results when harnessing supplementary evaluation data. SHAP values, calculated for the evaluation benchmark and simple meta-data model, suggest that the number of useless entries, the session's perceived usefulness, and time adequacy are the most important features in this setting. Our FastText approach slightly surpasses the benchmark of 0.65 on the evaluation data. Using additional *closed-question data* mostly enhances the performance. BERT (0.67), FastText (0.66), Word2Vec (0.64), and SVM trained on TF-IDF features (0.65) clearly outperform the benchmark model (0.58). Averaged FastText (0.63) features combined with SVM achieve solid results, however, they cannot reach the results of our LSTM architecture.

In most cases adding non-text data increases the model performance, most evidently in the case of evaluation data. The most basic approach considered (simple meta-data) outperforms all other approaches when working on pure text data as well as in combination with evaluation data. Thus, a participant's' attitude in combination with how much they write is an easily attainable and well-performing prediction setup. Among the non-neural approaches, at nine times, SVMs with the non-linear kernel are the most commonly chosen classifiers, with an additional two wins for linear SVMs. At six or seven each, LR, AdaBoost, and XGBoost do not differ much in how often they were chosen. The more sophisticated approaches (TF-IDF, word embeddings in combination with task-specific LSTM architectures, and BERT) constantly achieve good results, and on average yield the best AUC scores on our test set. We notice a pattern in the embedding dimension and input sequence length hyperparameters: the most prominent embedding dimension among Word2Vec models is

25 with a maximum input sequence length of 100, whereas FastText models prefer shorter sequences of 50 words and embedding dimensions of 10 or 25. These findings also hold when predicting intervention failure, thus indicating the need to treat these numbers as hyperparameters instead of choosing default values. Furthermore, most models do not benefit from the second (optional) LSTM layer, which points towards an overwhelming model complexity considering our dataset size.

7 Discussion of Clinical Usefulness

As discussed by several authors such as Olczak et al. [85], Cabitza and Campagner [88], and Scott, Cater and Coiera [89], prediction performance metrics are only one subdimension when evaluating ML models in health care settings. Therefore, we use the ten questions proposed by Scott, Cater and Coiera [89] to summarize and evaluate the prospective clinical value of the proposed winning models.

(1) *What is the purpose and context of the algorithm?* The pain points the respective algorithms address are (1) high dropout rates and (2) low response rates in DMHIs in light of limited resources. The proposed models provide insights into who will likely drop out or not benefit after two out of eight sessions. As such, these predictions serve to adapt individual treatment plans (e.g., through additional guidance, sessions, or reminders) only if and where necessary.

(2) *How good were the data used to train the algorithm?* We use the five categories (i.e., completeness, correctness, concordance, plausibility, and currency) to assess data quality for clinical research proposed by the review of Weiskopf and Weng [90]. Regarding completeness, the data consists of all information else provided to the interventions' e-coach for decision-making. Furthermore, it spans a large variety of what previous work found relevant for intervention dropout and outcome. While additional outside information, such as previous health records or expert assessments, could possibly improve the predictions, the effort necessary to collect them requires extensive steps, deteriorating the cost-value ratio. Since the data stems from RCTs, research staff monitored the completeness of entries and missing data was very low, as seen in supplementary material 1. As for correctness, all non-text dimensions were manually investigated by the two first authors to find mistakes, and data quality was found to be high. The fact that spelling-mistake correction did not increase cross-validation scores indicates a good quality of the text data. Concordance of the data was, for example, internally validated by cross-checking modules completed with the submitted answers, running pivot tables for related variables (e.g., current employment status and leadership responsibility), and ensuring the correct time sequence of the entries. To check for plausibility, every feature's range and distribution were manually checked by two authors. Questions and findings, including averages and ranges, were discussed with the third author, who was involved in the data collection to check for plausibility, and no issues remained open. The currency of the data is high as the nature of the online setup allows the instant use of the

data as soon as the patient submits their answers. As such, a deployed model could inform clinical decisions immediately.

(3) *Were there sufficient data to train the algorithm?* The data set at hand is comparatively small for Data Science applications in general, thus presenting one of the major limitations of this study. Especially deep neural networks usually require large amounts of data to perform well. At the same time, this is a prerequisite that is rarely met in E-Mental Health research [52], and with almost 850 participants, the data set is large for DMHI standards. As seen in the related work section, only one other paper considering intervention data exceeds the dataset size presented in this work. A literature review found a dataset size of 100 to be minimally adequate for outcome predictions in DMHIs [91], but only 44% of the 56 studies investigated complied with this criterium. Further, they found that only 29% used a hold-out test set or adequate cross-validation method. At a test set size of 163/170 that was not used for training at any point, the results at hand can be considered among the more generalizable of the works currently available [91]. To address the small dataset size, we extend the pre-training corpus with texts generated by control group users and train the word embeddings at the sentence level. Further, our use of a pre-trained BERT model comes with the significant advantage that - as researchers from a field struggling with data collection - we can leverage large unrelated but available data sets [19]. The results for the deep learning models are stable and good within and across different settings. This suggests an at least minimally adequate data set size for them to compete with classical machine learning models.

(4) *How well does the algorithm perform?* With almost all average AUC scores well above 0.5, it can be concluded that the considered features have predictive power regarding intervention outcome and dropout. With the best scores reaching an AUC of 0.70 (dropout) and 0.72 (intervention outcome) after just two weeks, results are competitive with related work. For example, Bremer et al., [18] achieved an AUC of 0.6 when using the user journey data (e.g., time spent) of their first two out of seven sessions to predict dropout. The best prediction models proposed by us achieve a balanced accuracy of 0.66 and 0.67. Forsell et al. [16] did not reach similar balanced accuracy scores predicting outcome with only symptom data until week 3 or 4. The comparison to other related works is limited due to differences in baselines and time horizons. The performance in the sense of clinical usefulness will be discussed in question 8.

(5) *Is the algorithm transferable to new clinical settings?* The specific models with their respective (hyper)parameters and, in the case of the NNs, task-specific architecture, can likely not be deployed on a different intervention. However, the proposed process to train the two best-performing models can be replicated on any dataset including intervention text and socio-demographic data. As can be seen in the related work section, these are very common data types to be collected in a standard DMHI setting. The text pre-processing steps are generalizable for any German text and would only have to be slightly adapted for English text (i.e., different handling of capital cases). Transferring models from one language to another in the clinical context has been shown to be possible in other tasks, especially for languages from the same family [92]. The

fact that pre-trained neural networks for English text are more in number and more specific in problem-fit [93, 94] indicates that the prediction results of the neural networks could even improve for the English language. Once text features are produced, they can easily serve a variety of outcome measures. The related work section shows several options, ranging from 24 h symptom prediction to 6-month follow-ups. Other options could be to use it to personalize content or adapt the time of intervention.

(6) *Are the outputs of the algorithm clinically intelligible?* Considering the transparency of the decision process, neural networks' black-box nature is one of their major drawbacks. For the non-sequential models, SHAP values and built-in feature importance measures give first insights into the decision-making process. These efforts can easily be extended per the suggestions made by Yang [95] but are left for future research as interpretability is not the focus of this paper. However, the actual outputs of both models are binary and easily understandable as they represent dropout vs. completers and intervention successes vs. failures per the above-given definitions.

(7) *How will this algorithm fit into and complement current workflows?* As of now, e-coaches receive general guidelines on how much time to spend with their allocated participants. Within a given RCT, these suggestions did not differ across participants. Once implemented, the models' predictions could prompt individual suggestions. For example, a stop-light system could indicate green (no risk), yellow (moderate risk), or red (high risk of dropout) [19]. With this information, therapists or e-coaches can decide or be instructed as to which participant is most in need of their time. Pedersen et al. [19] report this approach to have been positively received by therapists in their study. Such risk profiles could also prompt automatic reminders, personalized feedback loops to identify the problem, or additional content (e.g., a module regarding motivation or goal setting).

(8) *Has use of the algorithm been shown to improve patient care and outcomes?* The next step to evaluating the practical value of the proposed model is the implementation within a live intervention. However, this exceeds the limits of this paper. At the same time, studies such as Forsell et al. [16] and Pedersen et al. [19] have empirically proven the superiority of adaptive care for both dropout and outcome predictions. In the baseline, the limited resources are currently being distributed at random. Empirical evidence shows that many patients benefit from unguided interventions [13] and Forsell et al. [16] show that at-risk patients — while significantly benefiting — even with enhanced care, barely reach the same health outcomes as not at-risk counterparts.

The best model predicting outcome recognizes 93% of intervention failures (recall) while avoiding overspending on 41% of the most likely completers (specificity). The same calculations for the slightly less balanced dropout predictions lead to 55% correctly identified dropouts while avoiding overspending on 80% of completers. These metrics can be off-traded through the threshold deciding between a dropout or failure, as exemplarily shown in Fig. 2. The histograms show the intervention failure probability as predicted by the winning model for each group – intervention failures and successes. As expected,

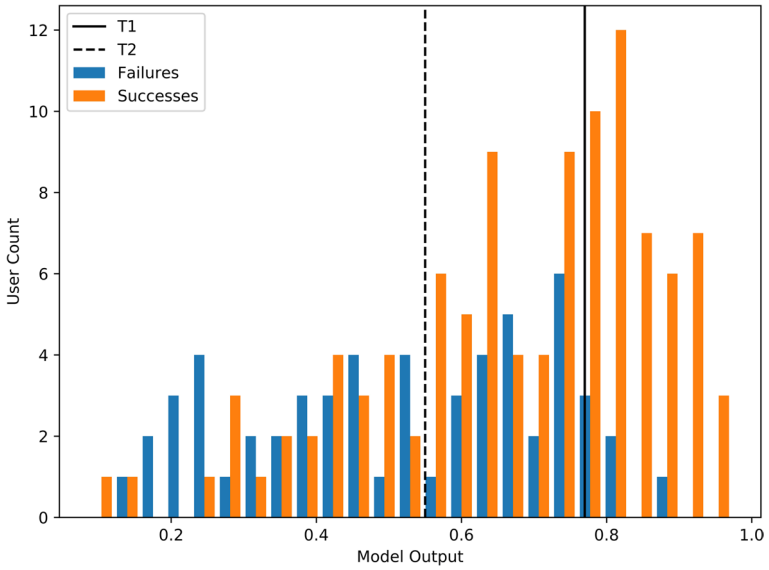


Fig. 2 Histogram failure prediction output

successes have a higher probability of being recognized as such (right side), failures are more prevalent in the low probabilities (left side), and there is a bulk of hard-to-identify participants in the middle. Changing the threshold from T1 (highest balance accuracy) to T2 decreases the recall to 53%; however, it avoids overspending on 75% of successes. Consequently, not much more than one-third of all participants receive enhanced care, lowering costs significantly while still addressing those most likely participants to benefit from support. The threshold can be adapted to fit the available resources and can even inform the number of participants accepted in the intervention. Considering the preventive nature of the stress intervention at hand, one application of the model could be to make the intervention available without guidance to reach as many participants as possible and only offer the available guidance to those who most need it. In the T2 scenario in Fig. 2, this increases the number of participants reached by threefold.

(9) *Could the algorithm cause patient harm?* The purpose is to optimize resource allocation while maintaining or improving the level of care over the entire population of participants. If it were used to reduce the average care level, it could harm those incorrectly classified as completers or successes through decreased levels of care. Such as prospect is especially worrisome when working with a population with severe symptoms. Depending on the importance of avoiding such false negatives, the recall can be increased at higher costs of resources. It, thus, must always be closely considered *how* to implement such a decision-support tool in *which* setting. At the same time, considering that right now, very limited resources are available, and many sick people are not being helped at all,

increasing the total number of participants treated is a factor to weigh in with individual effects.

(10) *Does use of the algorithm raise ethical, legal or social concerns?* Albeit the focus of early research being on establishing the overall feasibility, bias in the data must be considered early on. With primarily female participants that hold a university degree, the data at hand is — while typical for mental health interventions — not representative of the general population. Implementing such a model in routine care could disadvantage those groups with the already most extensive unmet needs and must be adjusted to ensure the best possible care for all [96]. In addition, ethical and legal aspects of an automated decision to change the level of care must be closely considered, especially in cases where the reason for the prediction is not transparent [95].

8 Conclusion

NLP methods can help make countless individual predictions based on text that would require impossible amounts of human resources to be analyzed. While the use of sophisticated NLP methods on non-clinical texts is continuously advancing in Mental Health diagnostics [30, 31, 35, 50], applications of NLP on E-mental health intervention text have been few and predominantly limited to simple models. In this case study, we train several ML models, considering various text representation methods and additional data inputs, to predict intervention failure and intervention dropout. For this, we use a dataset of 849 German-speaking participants of a stress intervention. By thoroughly evaluating combinations of the above-mentioned factors on our dataset, we contribute to the design choice of prediction models for intervention dropout and intervention outcome.

First, we demonstrate that *harnessing the sequential nature of text* by training deep learning models in combination with word embeddings outperforms the much simpler approach of using averaged word vectors on our test set. Thus, we complement existing research [25, 27] by proposing a task-specific LSTM architecture using word embeddings which successfully deals with the long input sequences and yields good results (average AUC score of 0.65) in dropout prediction. We further demonstrate the need to treat the embedding dimension as a hyperparameter rather than using the default values. Second, considering *supplementary baseline data* when predicting intervention outcome and *evaluation data* when predicting dropout yields the best-performing models. Thus, our findings support that the participants' background and attitude towards the intervention hold additional information in combination with text data. Third, we underline the *solid performance of easy-to-implement approaches* to predict dropout (simple meta-data and TF-IDF) and intervention outcome (advanced meta-data and sentiment analysis). By providing the insights from our case study, we seek to facilitate the development of ML-based tools which augment e-coaches' work in extracting valuable information from the participants' intervention texts

— hence, easing the task of identifying participants in need of human attention. With these predictions, necessary steps towards a more successful intervention in light of limited resources to face growing needs can be initiated.

Considering the still relatively small data set size and high specificity of our intervention set-up, this research is only a step towards better understanding, predicting, and ultimately influencing participants' behavior in DMHIs. Data sets such as this one can be considered the most promising approach to gathering knowledge in this research area. Yet, learning on few data points might not champion the same text representation methods and models, and more research is necessary to determine the generalizability of our findings. While we prove the potential of neural networks in this setting, they require large datasets, long training times, and have a black-box nature. However, the investigation of such complex methods is necessary to ensure the best possible results — especially considering the astonishing results deep learning models achieve on other NLP tasks. To truly understand human language, words must be considered beyond their lexical meaning, and the specific context needs to be understood — a task simple methods will never solve. One further way to address the problem of small datasets could be to use data augmentation methods as commonly used in computer vision, and more recently proposed for NLP tasks [97]. We suggest that employing attention-based [98] deep learning architectures can further enhance the model performance in prediction tasks such as ours. While designing task-specific network architectures like ours may be a complex and tedious task, large pre-trained text classification models can eliminate this work. To determine whether further research in applying pre-trained transformer models in this domain is auspicious, we examine the most prominent transformer model BERT and observe promising results in dropout prediction. Thus, we suggest investigating more sophisticated pre-trained transformer models (e.g., RoBERTa [99] or XLNet [100]) in such settings. In addition to an optimized pre-training strategy, XLNet tends to process long sentences better than BERT, which could be advantageous in cases like ours and further improve the model performance. Besides the particular transformer model, the text corpora used for pre-training, as well as the approaches to integrating the important non-text features into the model architecture, should be investigated in more detail (e.g., [101]). Furthermore, multi-task models (e.g., predicting intervention failure and dropout at the same time), which are frequently employed in other NLP tasks (e.g., [102]), can potentially improve results on both tasks. For the time being, simple feature representations such as metadata and classical statistical models should be considered an easy-to-implement yet competitive option for predicting intervention failure and dropout. In that regard, further research must be conducted to investigate how to improve these predictions, for example, more automized ways of finding the most important TF-IDF features [103].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41666-023-00148-z.

Author Contribution Kirsten Zantvoort and Jonas Scharfenberger conceptualized the paper under the supervision of Burkhardt Funk, conducted the analysis, and wrote the first draft of the paper. Dirk Lehr and Leif Boß were involved in the RCT study design and original data collection. Leif Boß contributed to the data selection and pre-processing decisions and supported data understanding. Leif Boß, Burkhardt

Funk, and Dirk Lehr provided substantial feedback on different versions of the manuscript. All authors contributed to the final draft. All authors contributed to the article and approved the submitted version.

Funding Open Access funding enabled and organized by Projekt DEAL. The present study has been funded by Leuphana University. The original RCTs were funded by the European Union (project EFRE: CCI 2007DE161PR001).

Data Availability Due to the sensitivity of text data in the context of DMHIs, the data cannot be made available. The source code can be made available upon request to the corresponding author.

Declarations

Ethics Approval Each RCT has been approved by the respective ethic commission, details can be found in the German Clinical Trials Register Numbers included in Table 1.

Consent to Participate The participants provided their written informed consent to participate in the respective RCTs.

Consent for Publication The participants provided their written informed consent for anonymized findings to be published.

Conflict of Interest Burkhardt Funk is a shareholder of GET.ON Institut für Online Gesundheitstrainings GmbH. The other authors declare they have no financial interests.

References

1. Wang PS, Lane M, Olfson M, Pincus HA, Wells KB, Kessler RC (2005) Twelve-month use of mental health services in the United States. JAMA Psychiatry 62(6):629–640. https://doi.org/10.1001/archpsyc.62.6.629
2. Rommel A, Bretschneider J, Kroll LE, Prütz F, Thom J (2017) Inanspruchnahme psychiatrischer und psychotherapeutischer Leistungen – Individuelle Determinanten und regionale Unterschiede. J Health Monit 68(08):e31
3. Santomauro DF, Herrera AMM, Shadid J, Zheng P, Ashbaugh C, Pigott DM, Abbafati C, Adolph C, Amlag JO, C.-1. M. D. Collaborators (2021) Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 398(10312):1700–1712. https://doi.org/10.1016/S0140-6736(21)02143-7
4. Ebert DD, Harrer M, Apolinário-Hagen J, Baumeister H (2019) Digital interventions for mental disorders: key features, efficacy, and potential for artificial intelligence applications, In Frontiers in Psychiatry, Singapore, Springer Natur, pp 584–627
5. Karyotaki E, Kleiboer A, Smit F, Turner D, Pastor A, Andersson G, Berger T, Botella C, Breton J, Carlbring P, Christensen H, de Graaf E, Griffiths K, Donker T, Farrer L, Huibers M, Lenndin J, Mackinnon A, Meyer B, Moritz S, Riper R (2015) Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. Psychol Med 45(13):2717–2726. https://doi.org/10.1017/S0033291715000665

6. Andersson G, e Carlbring, Rozental A (2019) Response and remission rates in internet-based cognitive behavior therapy: an individual patient data meta-analysis. Front Psychiatry 10. https://doi.org/10.3389/fpsyt.2019.00749
7. Heber E, Ebert DD, Lehr D, Cuijpers P, Berking M, Nobis S, Riper H (2017) The benefit of web- and computer-based interventions for stress: a systematic review and meta-analysis. J Med Internet Res 19(2):e32. https://doi.org/10.2196/jmir.5774
8. Reins JA, Buntrock C, Zimmermann J, Grund S, Harrer M, Lehr D, Baumeister H, Weisel K, Domhardt M, Imamura K, Kawakami N, Spek V, Nobis S, Snoek F, Cuijpers P, Klein JP, Moritz S (2021) Efficacy and moderators of internet-based interventions in adults with subthreshold depression: an individual participant data meta-analysis of randomized controlled trials. Psychother Psychosom 90(2):94–106.https://doi.org/10.1159/000507819
9. Karyotaki E, Ebert DD, Donkin L, Riper H, Twisk J, Burger S, Rozental A, Lange A, Williams AD, Zarski AC, Geraedts A, Straten Av, Kleiboer A, Meyer B, Ince BBÜ, Buntro C (2018) Do guided internet-based interventions result in clinically relevant changes for patients with depression? An individual participant data meta-analysis. Clin Psychol Rev 63:80–92. https://doi.org/10.1016/j.cpr.2018.06.007
10. Domhardt M, Letsch J, Kybelka J, Koenigbauer J, Doebler P, Baumeister H (2020) Are Internet- and mobile-based interventions effective in adults with diagnosed panic disorder and/or agoraphobia? A systematic review and meta-analysis. J Affected Disord 276:169–182. https://doi.org/10.1016/j.jad.2020.06.059
11. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N (2011) A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 13(3):e52. https://doi.org/10.2196/jmir.1772
12. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M (2021) Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 3. https://doi.org/10.3389/fdgth.2021.764079
13. Richards D, Richardson T (2012) Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev
14. Baumeister H, Reichler L, Munzinger M, Lin J (2014) The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interv 1(4):205–215. https://doi.org/10.1016/j.invent.2014.08.003
15. Hilvert-Bruce Z, Rossouw PJ, Wong N, Sunderland M, Andrews G (2012) Adherence as a determinant of effectiveness of internet cognitive behavioural therapy for anxiety and depressive disorders. Behav Res Ther 50(7-8):463–468. https://doi.org/10.1016/j.brat.2012.04.001
16. Forsell E, Jernelöv S, Blom K, Kraepelien M, Svanborg, Andersson G, Lindefors N, Kaldo V (2019) Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: a single-blind randomized clinical trial with insomnia patient. Am J Psychiatry 176(4):315–323. https://doi.org/10.1176/appi.ajp.2018.18060699
17. Shatte ABR, Hutchinson DM, Teague SJ (2019) Machine learning in mental health: a systematic scoping review of methods and applications. Psychol Med 49(9):1426–1448
18. Bremer V, Chow PI, Funk B, Thorndike FP, Ritterband LM (2020) Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: machine learning approach. J Med Internet Res 22(10)
19. Pedersen DH, Mansourvar M, Sortsø C, Schmidt T (2019) Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. J Med Internet Res 21(9). https://doi.org/10.2196/13617
20. Chekroud A, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, Cohen Z, Belgrave D, DeRubeis R, Iniesta R, Dwyer D, Choi K (2021) The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 20(2):154–170. https://doi.org/10.1002/wps.20882
21. Corcoran CM, Benavides C, Cecchi G (2019) Natural language processing: opportunities and challenges for patients, providers, and hospital systems. Psychiatr Annu 49(5):202–208. https://doi.org/10.3928/00485713-20190411-01
22. Abbe A, Grouin C, Zweigenbaum P, Falissard B (2015) Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res 25(2):86–100. https://doi.org/10.1002/mpr.1481

23. Calvo R, Milne D, Hussain M, Christensen H (2017) Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng 23(5):649–685. https://doi.org/10.1017/S1351324916000383
24. Bone D, Lee C-C, Chaspari T, Gibson J, Narayanan S (2017) Signal processing and machine learning for mental health research and clinical applications. IEEE Signal Process Magazin 34(5):196–195. https://doi.org/10.1109/MSP.2017.2718581
25. Funk B, Sadeh-Sharvit S, Fitzsimmons-Craft E, Trockel M, Monterubio G, Goel N, Balantekin K, Eichen D, Flatt R, Firebaugh M-L, Jacobi C, Graham A, Hoogendoorn M (2020) A framework for applying natural language processing in digital health interventions. J Med Internet Res 22(2):e13855. https://doi.org/10.2196/13855
26. Hoogendoorn M, Berger T, Schulz A, Stolz T, Szolovits P (2016) Predicting social anxiety treatment outcome based on therapeutic email conversations. IEEE J Biomed Health Inform 21(5):1449–1459. https://doi.org/10.1109/JBHI.2016.2601123
27. Gogoulou E, Boman M, Abdesslem FB, Isacsson N, Kaldo V, Sahlgren M (2021) Predicting treatment outcome from patient texts: the case of internet-based cognitive behavioural therapy. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics pp 575–580. https://doi.org/10.18653/v1/2021.eacl-main.46
28. Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155
29. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota 1:4171–4186. https://doi.org/10.18653/v1/N19-1423
30. Nobles AL, Glenn JJ, Kowsari K, Teachman eA, Barnes LE (2018) Identification of imminent suicide risk among young adults using text messages. In: Nobles AL et al (ed) Identification of Imminent Suicide Risk Among Young Adults using Text Messages. Proceedings of the SIGCHI conference on human factors in computing systems. CHI Conference, pp 1–11. https://doi.org/10.1145/3173574.3173987
31. Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N (2018) SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe pp 1485–1497
32. Howes C, Purver M, McCabe R (2014) Linguistic indicators of severity and progress in online text-based therapy for depression. Association for Computational Linguistics. In: Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Baltimore pp 7–16. https://doi.org/10.3115/v1/W14-3202
33. Calvo R, Milne DN, Hussain S, Christensen H (2017) Natural language processing in mental health applications using non-clinical texts 23(5):649–685
34. Oesterreich TD, Fitte C, Behne A, Teuteberg F (2020) Understanding the role of predictive and prescriptive analytics in healthcare: a multi-stakeholder approach. In: Proceedings of the 28th European Conference on Information Systems (ECIS) 28:1–19
35. Wołk A, Chlasta K, Holas P (2021) Hybrid approach to detecting symptoms of depression in social media entries, in Pacific Asia Conference on Information Systems Proceedings, Dubai, UAE
36. Tsang EW (2014) Case studies and generalization in information systems research: a critical realist perspective. J Strat Inf Syst 23:174–186
37. Eloranta S, Boman M (2022) Predictive models for clinical decision making: deep dives in practical machine learning. J Intern Med 292(2):278–295. https://doi.org/10.1111/joim.13483
38. Cepoiu M, McCusker J, Cole MG, Sewitch M, Belzile E, Ciampi A (2007) Recognition of depression by non-psychiatric physicians—a systematic literature review and meta-analysis. J Gen Intern Med 23(1):25–36. https://doi.org/10.1007/s11606-007-0428-5
39. DeMasi O, Kording K, Recht B (2017) Meaningless comparisons lead to false optimism in medical machine learning. PLoS One 12(9):e0184604. https://doi.org/10.1371/journal.pone.0184604
40. Becker D, Breda Wv, Funk B, Hoogendoorna M, Ruwaardc J, Riperc H (2018) Predictive modeling in e-mental health: a common language framework. Internet Interv 12:57–67. https://doi.org/10.1016/j.invent.2018.03.002
41. Le Glaz A, Haralambous Y, Kim-Dufor D-H, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C (2021) Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 23(5):e15708. https://doi.org/10.2196/15708

42. Paul A, Liao W-k, Alok Choudhary AA (2021) Harnessing psycho-lingual and crowd-sourced dictionaries for predicting taboos in written emotional disclosure in anonymous confession boards. J Healthc Inform Res 5:319–341
43. Masino AJ, Forsyth D, Fiks AG (2018) Detecting adverse drug reactions on twitter with convolutional neural networks and word embedding features. J Health Inform Res 2:25–43
44. Yeruva VK, Junaid S, Lee Y (2019) Contextual word embeddings and topic modeling in healthy dieting and obesity. J Healthc Inform Res 3:159–183
45. Spärck Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc 28(1):11–21. https://doi.org/10.1108/eb026526
46. Marcus MD, Wildes JE (2012) Obesity in DSM-5. Psychiatr Ann 42(11):431–435. https://doi.org/10.3928/00485713-20121105-10
47. Wongkoblap A, Vadillo M, Curcin V (2021) Depression detection of twitter posters using deep learning with anaphora resolution: algorithm development and validation. J Med Internet Res Ment Health 8(8). https://doi.org/10.3390/electronics11050676
48. Pennebaker J, Boyd R, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015. University of Texas at Austin, Austin
49. Coppersmith G, Carvalho P, Silva MJ, Wallace BC, Amir S (2017) Quantifying mental health from social media with neural user embeddings. In: Proceedings of the 2nd Machine Learning for Healthcare Conference, Boston 68:306–321
50. Bucur A-M, Cosma A, Dinu LP (2021) Early risk detection of pathological gambling, self-harm and depression using BERT. In: Proceedings of Conference and Labs of the Evaluation Forum, Bucharest, Romania
51. Ewbank MP, Cummins R, Tablan V, Bateup S, Catarino A, Martin AJ, Blackwell AD (2020) Quantifying the association between psychotherapy content and clinical outcomes using deep learning. JAMA Psychiatry 77(1):35–43. https://doi.org/10.1001/jamapsychiatry.2019.2664
52. Pasini A (2015) Artificial neural networks for small dataset analysis. J Thorac Dis 7(5). https://doi.org/10.3978/j.issn.2072-1439.2015.04.61
53. Eysenbach G (2005) The law of attrition. J Med Internet Res 7(1):1–9. https://doi.org/10.2196/jmir.7.1.e11
54. Pihlaja S, Lahti J, Lipsanen JO, Ritola V, Gummerus E-t, Stenberg J-H, Joffe G (2020) Scheduled telephone support for internet cognitive behavioral therapy for depression in patients at risk for dropout: pragmatic randomized controlled trial. J Med Internet Res 22(7):e15732. https://doi.org/10.2196/15732
55. Smink WAC, Sools AM, Postel MG, Sang ETK, Elfrink A, Libbertz-Mohr LB, Veldkamp BP, Westerhof GJ (2021) Analysis of the emails from the Dutch web-based intervention "Alcohol de Baas": assessment of early indications of drop-out in an online alcohol abuse intervention. Front Psychiatry 12:575931. https://doi.org/10.3389/fpsyt.2021.575931
56. Grave E, Joulin A, Mikolov T, Bojanowski P (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
57. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: In International conference on machine learning, Beijing 32(2):1188–1196
58. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A (2017) Advances in pre-training distributed word representations. arXiv:1712.09405
59. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK (1987) Occam's Razor. Inf Process Lett 24(6):377–380. https://doi.org/10.1016/0020-0190(87)90114-1
60. D'Zurilla TJ, Nezu AM (2010) Problem-solving therapies. In: Handbook of cognitive–behavioral therapies, vol 3. Guilford Press, pp 197–225
61. Berking M, Whitley B (2014) Affect regulation training - a practitioners' manual, New York. Springer, NY. https://doi.org/10.1007/978-1-4939-1022-9
62. Heber E, Lehr D, Ebert DD, Berking M, Riper H (2016) Web-based and mobile stress management intervention for employees: a randomized controlled trial. J Med Internet Res 18(1)
63. Ebert DD, Lehr D, Heber E, Riper H, Cuijpers P, Berking M (2016) Internet- and mobile-based stress management for employees with adherence-focused guidance: efficacy and mechanism of change. Scand J Work Environ Health 41(2):107–218. https://doi.org/10.5271/sjweh.3573
64. Ebert DD, Heber E, Berking M, Riper H, Cuijpers P, Funk B, Lehr D (2016) Self-guided internet-based and mobile-based stress management for employees: results of a randomised controlled trial. Occup Environ Med 73(5):315–323

65. Nixon P, Ebert DD, Boß L, Angerer P, Dragano N, Lehr D (n.d.) Web-based stress management intervention for employees experiencing effort-reward imbalance at work: a randomized controlled trial. Preprint
66. Ebert DD, Franke M, Zarski A-C, Berking M, Riper H, Cuijpers P, Funk B, Lehr D (2021) Effectiveness and moderators of an internet-based mobile-supported stress management intervention as a universal prevention approach: randomized controlled trial. J Med Internet Res 23(12):e22107. https://doi.org/10.2196/22107
67. Nixon P, Ebert DD, Boß L, Angerer P, Dragano N, Lehr D (2022) Efficacy of a web-based stress management intervention for employees experiencing adverse working conditions and occupational self-efficacy as mediator: a randomized controlled trial. J Med Internet Res 24(10). https://doi.org/10.2196/40488
68. Cohen S, Kamarck T, Mermelstein R (1983) A global measure of perceived stress. J Health Soc Behav 24(4):385–396. https://doi.org/10.2307/2136404
69. Schneider EE, Schönfelder S, Domke-Wolf M, Wessa M (2020) Measuring stress in clinical and nonclinical subjectsusing a German adaptation of the Perceived StressScale. Int J Clin Health Psychol
70. Jacobson NS, Truax P (1991) Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 59(1):12–19. https://doi.org/10.1037/0022-006X.59.1.12
71. Christensen H, Griffithi KM, Farrer L (2009) Adherence in internet interventions for anxiety and depression: systematic review. J Med Internet Res 11(2):e13. https://doi.org/10.2196/jmir.1194
72. Hedman E, Ljótsson B, Kaldo V, Hesser H, Alaoui SE, Kraepelien M, Andersson E, Rück C, Svanborg C, Andersson G, Lindefors N (2014) Effectiveness of Internet-based cognitive behaviour therapy for depression in routine psychiatric care. J Affect Disord 155:49–58. https://doi.org/10.1016/j.jad.2013.10.023
73. Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E (2016) Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. Comput Math Methods Med 2016:8708434. https://doi.org/10.1155/2016/8708434
74. Fehle J, Schmidt T, Wolff C (2021) Lexicon-based sentiment analysis in German: systematic evaluation of resources and preprocessing techniques. In: Proceedings of the 17th Conference on Natural Language Processing, Düsseldorf pp 86–103
75. Camacho-Collados J, Pilehvar MT (2018) On the role of text preprocessing in neural network architectures: an evaluation study on text categorization and sentiment analysis. In: Proceedings of the 2018 Conference of Empirical Methods in Natural Language Processing Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels
76. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
77. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority oversampling technique. J Artif Intell Res 16(1):321–357 https://doi.org/10.1613/jair.953
78. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn
79. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system, In Knowledge Discovery and Data Mining, San Francisco
80. Guyon I, Saffari A, Dror G, Cawley G (2011) Model selection: beyond the Bayesian/Frequentist divide. J Mach Learn Res 61–87
81. Schapire RE (2013) Explaining AdaBoost, In Empirical Inference, Heidelberg, Springer-Verlag Berlin Heidelberg
82. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Sig Process 2673–2681. https://doi.org/10.1109/78.650093
83. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
84. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30(7):1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2
85. Olczak J, Pavlopoulos J, Prijs J, Ijpma FFA, Doornberg JN, Lundström C, Hedlund J, Gordon M (2021) Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthop 92(5):513–525

86. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Conference on Neural Information Processing Systems, Long Beach pp 4768–4777
87. Barrett MS, Chua W-J, Crits-Christoph P, Gibbons MB, Casiano D, Thompson D (2008) Early withdrawal from mental health treatment: implications for psychotherapy practice. Psychotherapy 45(2):247–267. https://doi.org/10.1037/0033-3204.45.2.247
88. Cabitza F, Campagner A (2021) The need to separate the wheat from the chaff in medical informatics. Int J Med Inform 153:104510. https://doi.org/10.1016/j.ijmedinf.2021.104510
89. Scott I, Carter S, Coiera E (2021) Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inf 28:e100251. https://doi.org/10.1136/bmjhci-2020-100251
90. Weiskopf NG, Wenig C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 20(1):144–151. https://doi.org/10.1136/amiajnl-2011-000681
91. Sajjadian M, Lam RW, Milev R, Rotzinger S, Frey BN, Soares CN, Parikh SV, Foster JA, Turecki G, Müller DJ, Strother SC, Farzan F, Kennedy SH, Uher R (2021) Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. Psychol Med 51(16):2742–2751
92. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P (2018) clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semant 9(12):1–13
93. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E (2021) MentalBERT: publicly available pretrained language models for mental healthcare. Comput Lang. https://doi.org/10.48550/arXiv.2110.15621
94. Hugging Face, huggingface model overview, [Online]. Available: https://huggingface.co/models?language=de&sort=downloads. Accessed 23 09 2022
95. Yang CC (2022) Explainable artificial intelligence for predictive modeling in healthcare. J Healthc Inform Res 8:228–239
96. Gianfrancesco M, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178(11):1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763
97. Xiang R, Chersoni E, Lu Q, Huang CR, Li W, Long Y (2021) Lexical data augmentation for sentiment analysis. J Am Soc Inf Sci 72(11):1432–1447
98. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhin (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). NIPS, Long Beach, CA, USA, pp 6000–6010
99. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding, Proceedings of the 33rd International Conference on Neural Information Processing Systems., Curran Associates Inc., Red Hook, 517:5753–5763
100. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: a robustly optimized BERT pretraining approach. arxiv 1907.11692
101. Shen JX, Ma MD, Xiang R, Lu Q, Vallejos EP, Xu G, Huang CR, Long Y (2020) Dual memory network model for sentiment analysis of review text. Knowl-Based Syst 188:105004
102. Hashimoto K, Xiong C, Tsuruoka Y, Socher R (2017) A joint many-task model: growing a neural network for multiple NLP tasks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 1923–1933. https://doi.org/10.18653/v1/D17-1206
103. Zhang Y, Zhou Y, Yao J (2020) Feature extraction with TF-IDF and game-theoretic shadowed sets communications in computer and information science. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems, vol 1237. Springer, Cham, pp 722–733. https://doi.org/10.1007/978-3-030-50146-4_53

frontiers | Frontiers in Digital Health

TYPE Systematic Review
PUBLISHED 22 May 2023
DOI 10.3389/fdgth.2023.1170002

OPEN ACCESS

EDITED BY
Kwok Leung Tsui,
Virginia Tech, United States

REVIEWED BY
Rüdiger Christoph Pryss,
Julius Maximilian University of Würzburg, Germany
Sanne Booij,
University Medical Center Groningen, Netherlands

*CORRESPONDENCE
Silvan Hornstein
silvan.hornstein@hu-berlin.de

RECEIVED 20 February 2023
ACCEPTED 05 May 2023
PUBLISHED 22 May 2023

CITATION
Hornstein S, Zantvoort K, Lueken U, Funk B and Hilbert K (2023) Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms.
Front. Digit. Health 5:1170002.
doi: 10.3389/fdgth.2023.1170002

Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms

Silvan Hornstein[1]*, Kirsten Zantvoort[2], Ulrike Lueken[1], Burkhardt Funk[2] and Kevin Hilbert[1]

[1]Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany, [2]Institute of Information Systems, Leuphana University, Lueneburg, Germany

Introduction: Personalization is a much-discussed approach to improve adherence and outcomes for Digital Mental Health interventions (DMHIs). Yet, major questions remain open, such as (1) what personalization is, (2) how prevalent it is in practice, and (3) what benefits it truly has.
Methods: We address this gap by performing a systematic literature review identifying all empirical studies on DMHIs targeting depressive symptoms in adults from 2015 to September 2022. The search in Pubmed, SCOPUS and Psycinfo led to the inclusion of 138 articles, describing 94 distinct DMHIs provided to an overall sample of approximately 24,300 individuals.
Results: Our investigation results in the conceptualization of personalization as purposefully designed variation between individuals in an intervention's therapeutic elements or its structure. We propose to further differentiate personalization by what is personalized (i.e., intervention content, content order, level of guidance or communication) and the underlying mechanism [i.e., user choice, provider choice, decision rules, and machine-learning (ML) based approaches]. Applying this concept, we identified personalization in 66% of the interventions for depressive symptoms, with personalized intervention content (32% of interventions) and communication with the user (30%) being particularly popular. Personalization via decision rules (48%) and user choice (36%) were the most used mechanisms, while the utilization of ML was rare (3%). Two-thirds of personalized interventions only tailored one dimension of the intervention.
Discussion: We conclude that future interventions could provide even more personalized experiences and especially benefit from using ML models. Finally, empirical evidence for personalization was scarce and inconclusive, making further evidence for the benefits of personalization highly needed.
Systematic Review Registration: Identifier: CRD42022357408.

KEYWORDS
depression, digital mental health, personalization, precision care, iCBT, machine learning

1. Introduction

At an estimated lifetime prevalence of more than 10% (1, 2), major depressive disorder (MDD) is the second leading cause of years lived in disability (3). While this makes efficient treatments urgently needed, traditional approaches such as face-to-face psychotherapy are difficult to access for a significant part of patients (4–6). However, providing treatment through digital channels such as mobile applications and online formats (7) is effective in reducing depressive symptoms (8, 9) in a cost-effective way (10). Since most of the world

population has access to the internet (11) and/or a smartphone (12), digital mental health interventions (DMHIs) bypass barriers to traditional treatment.

Despite their potential, DMHIs inherit some of the general problems in depression treatment: Estimates for treatment dropout, as observed in RCTs, are up to 50% when considering publication bias (13). Moreover, response rates are unsatisfactory at less than 50% (14). Therefore, improving outcomes and reducing dropouts in DMHIs are expected to be highly impactful in facing the burden of depression.

Luckily, DMHIs' unique delivery channel provides new opportunities to improve the treatment of those suffering from depressive symptoms. Specifically, digital applications can efficiently be individualized to improve users' experience and outcomes, as observable across many other domains, ranging from e-commerce (15) over e-learning (16) towards social media (17). Simultaneously, the importance of accommodating patients' preferences for treatment outcomes in mental healthcare has been well established (18). Hence, the personalization of interventions to adapt treatment to individual needs is a promising approach to improving care, for depressive symptoms and beyond (19–22).

In line with that idea, a meta-analysis from 2013 showed that algorithm-based tailoring of DMHIs is associated with better outcomes (23). A review from 2022 found that none of the 26 reviewed apps for depression used just-in-time (JIT) adaptations, a mechanism for personalizing the timing of content delivery based on the individual or the situation (24). Another current systematic review investigated tailored interventions for workplace mental health (25), finding benefits on several outcomes when content or feedback was tailored towards the individual. Finally, a component network analysis examined the benefits of common internet-based cognitive behavioral therapy (iCBT) packages for depression, discovering small interactions between treatment components and patient characteristics (26).

While these publications are unified in their call for more personalization in DMHIs, they do not add up to a satisfactory empirical and theoretical ground for it. Firstly, the fragmented use of vocabulary fails to demarcate personalization from other distinct phenomena related to variability in DMHIs. For example, the term "tailoring" is used across various scopes and foci (23, 25, 27), while similar mechanisms are elsewhere called "individualized" (28) or "personalized" (26, 29). This diversity in vocabulary is shared with non-digital settings, as for traditional psychotherapy, 15 different terms for the same phenomena of varying treatment between individuals were reported (30). Secondly, in contrast to the breadth of used vocabulary, the focus of mechanisms within studies seems to be relatively narrow, focusing on specific mechanisms (23, 24) or areas (25, 28) of personalization. This potentially leads to an underestimation of variability already in place. Finally, while two of the mentioned reviews investigated the benefits of personalization through direct comparisons, they did so without a specific focus on depression and, related to the aforementioned narrow conceptualizations of personalization, with few studies being included. In conclusion, the concept, prevalence, and efficacy of personalization in DMHIs for depressive symptoms are not adequately delineated. Therefore, a disorder-specific review developing a conceptual framework for personalization and reviewing a wide span of interventions seems needed.

This systematic review aims to reduce the gap between the potential of personalization and its actual implementation by performing a comprehensive review of DMHIs for depressive symptoms with the following purposes:

1. Extract a conceptual framework that allows a clear and meaningful way of investigating, discussing, and classifying personalization.
2. Apply this framework to the available literature and report current use and mechanisms.
3. Evaluate the available evidence by identifying studies that directly compare interventions with different degrees of personalization.

2. Methods

This review was planned and reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (31). The protocol of this review was registered in the International Prospective Register of Systematic Reviews of the National Institute for Health Research (PROSPERO) under the ID CRD42022357408. The protocol was updated once after initial piloting to improve the alignment of the inclusion criteria and data extraction method with the scope of the review. Specifically, a new classification dimension for personalization was added that occurred in the literature and did not fit the pre-defined schema and the exclusion of e.g., prenatal depression was added to improve the comparability between included interventions. The final version of the protocol can be found in the **Supplementary Appendix S1**.

2.1. Search strategy

In the first step, a search was performed in three major databases (SCOPUS, PubMed, PsycInfo) to identify all published studies on DMHIs for depressive symptoms. The full search strings can be found in **Supplementary Appendix S2**. Additionally, three related reviews (13, 14, 32) were screened, and studies not yet included were added. Finally, papers brought to the author's awareness by being discussed in our included articles, not included yet but fulfilling our selection criteria, were added.

2.2. Selection criteria

We included empirical studies on DMHIs specifically targeting depressive symptoms, determining the interventions target by authors' self-report. This covered both, patients with diagnosed major depressive disorder (MDD), as well as with subclinical levels of symptoms. To be considered a DMHI, interventions

needed to be delivered through the internet and/or a smartphone. We included only empirical, peer-reviewed, English studies and conference articles with original data and patient cohort. To ensure a focus on the most relevant interventions for current use, we start our search from 2015 onwards.

To narrow down the focus of this review, studies on interventions targeting comorbid disorders such as anxiety were excluded. The same applied to those targeting a specific subtype of depression (e.g., prenatal depression), a single sub-symptom (e.g., rumination), or adolescent or elderly people (below 18 years or >64 years). Finally, those studies using digital technologies exclusively as a means of communication, such as one-on-one psychotherapy delivered via the web, were excluded as well.

2.3. Selection procedure

One of the researchers (S.H.) performed an initial screening based on the title and abstract of the studies identified through the search strategy. A second researcher (K.Z.) conducted the same procedure for a randomly chosen subset of 100 studies, resulting in excellent interrater reliability (0.94). The full description of the intervention was then read by both reviewers for all remaining papers to determine the final selection, extract interventions and code the variables of interest. Disagreements on any aspect of this process were solved by discussion between the reviewers until a consensus was reached. If full texts were unavailable, they were requested from the corresponding author. This occurred 12 times, with 8 of the articles made available on request.

2.4. Development of the conceptual framework

During the initial screening and before the update of the PROSPERO registration, we developed the proposed framework in an iterative process, considering usability, conceptual literature, and the observed interventions. Specifically, we discussed how we could classify personalization in a way that allows us to not just cover all mechanisms in the literature but

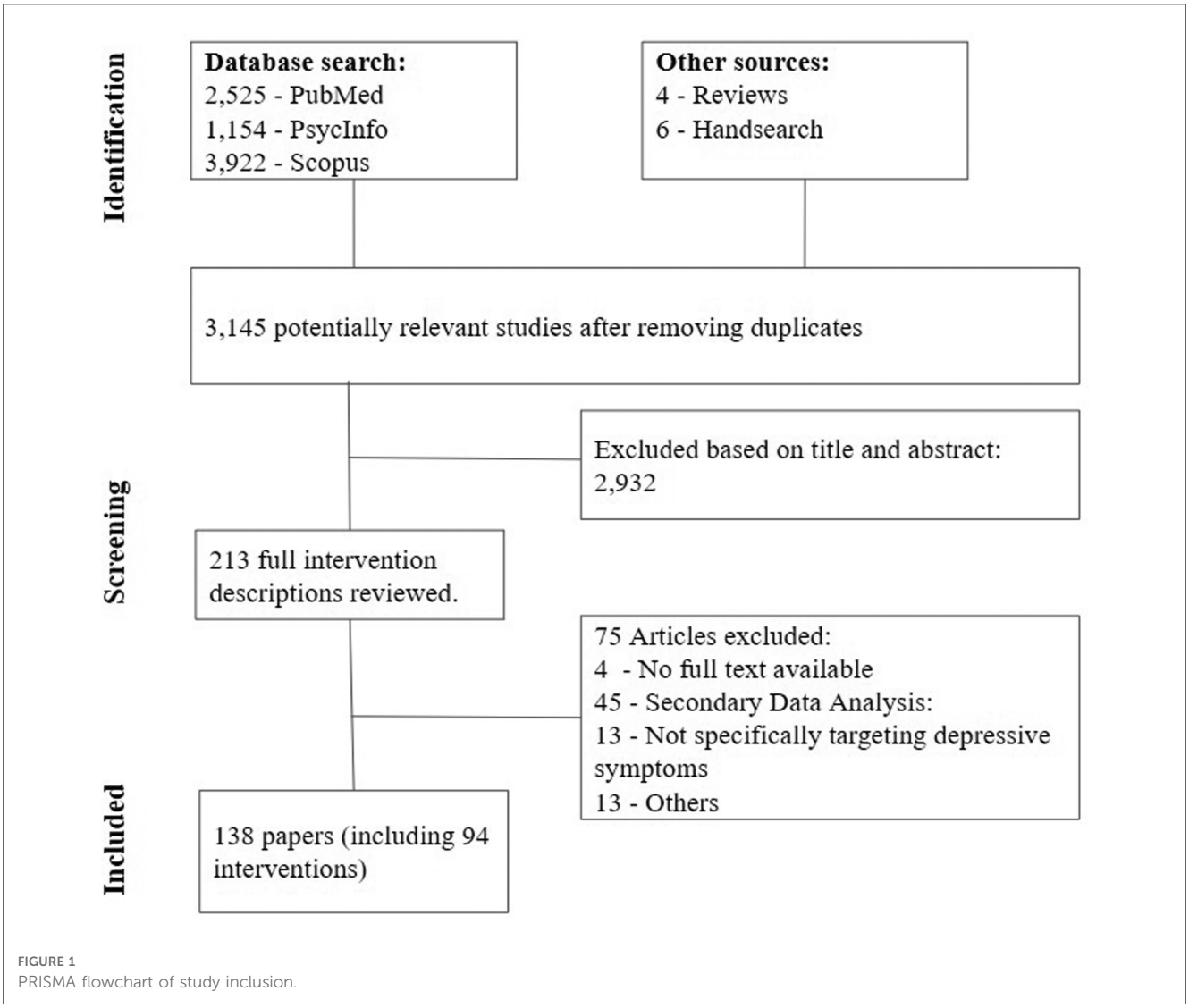


FIGURE 1
PRISMA flowchart of study inclusion.

also maximize usability by defining the dimensions as distinct as possible. We did this as we needed a method to classify personalization mechanisms during the systematic review and we could not find a satisfactory framework in the literature yet.

We departed from a common dictionary definition defining personalization as "the action of designing or producing something that meets someone's individual requirement" (33). Based on that, we intended to classify personalization in DMHIs in a broad enough way to cover the diversity of mechanisms present in related reviews and studies. At the same time, we intended to narrow down the concept to those mechanisms affecting the therapeutic content and structure, setting it apart from superficial sources of variability. Therefore, we excluded interactivity (34), the sole replay of user input as part of the app experience. For example, showing each patient their previously set goal might be a powerful tool, but it does not change the underlying therapeutic elements delivered. Additionally, we factored out customization (35), minor aesthetic adaptation such as users ability to change the color of an avatar. Finally, seeing personalization as referring to the level of the individual patient, we excluded group-based variability, such as cultural adaptation of the entire intervention (36).

Numerous screened interventions used a structured session-based approach to deliver their intervention—a common approach among manualized mental health interventions (37). Therefore, we identified (a) content (what is delivered during a session) and (b) order (how sessions are ordered) as potential areas of personalization. Since (c) guidance (level of human contact) is a highly relevant and variable aspect of DMHIs (38) we added it as another dimension. Finally, as we discovered prompts and mechanisms targeting the timing of interventions not being sufficiently represented in these three categories, we appended (d) communication as another dimension.

While, as mentioned above, we intended to exclude customization as minor user-choice-based adaptations of the intervention, we did not exclude user choice *per se* from our concept. This differs from the use in fields like marketing, where anything done by the user is defined as customization, not personalization (35). However, we saw the inclusion of actively designed user choice being justified for the following reasons: Firstly, user choice was a common mechanism described in the included interventions. Secondly, those mechanisms seem easily implementable and therefore highly relevant for practitioners interested in personalizing their intervention. Finally, user agency has been shown to be particularly relevant in mental healthcare (18). We also identified provider choice as another mechanism for guided and blended interventions. For data-driven personalization mechanisms, we saw rule-based and ML as distinct mechanisms applying static or learning criteria for personalization.

2.5. Data extraction

The framework developed above was applied to all identified interventions, coding the presence of personalization for each of the four (a–d) dimensions and classifying the underlying mechanism. For this, interventions had to be extracted from the included studies, and information from several studies on the same intervention had to be merged. If more than one distinct intervention was presented in a study, they were coded separately. Intervention versions in different languages were not coded separately if not reported to be clearly distinct in their content. If more than one study was available, a single observation of personalization resulted in a positive coding, but conflicting information was noted. Additionally, cited material such as older papers, weblinks, or appendices were consulted in the refrained from additional free-hand research on the reported interventions. In case information was indicative of personalization but insufficient for our coding, we contacted the corresponding author and asked for clarification. For this, we provided a four-week response window, including one reminder. Out of the seven authors contacted, six responded by providing

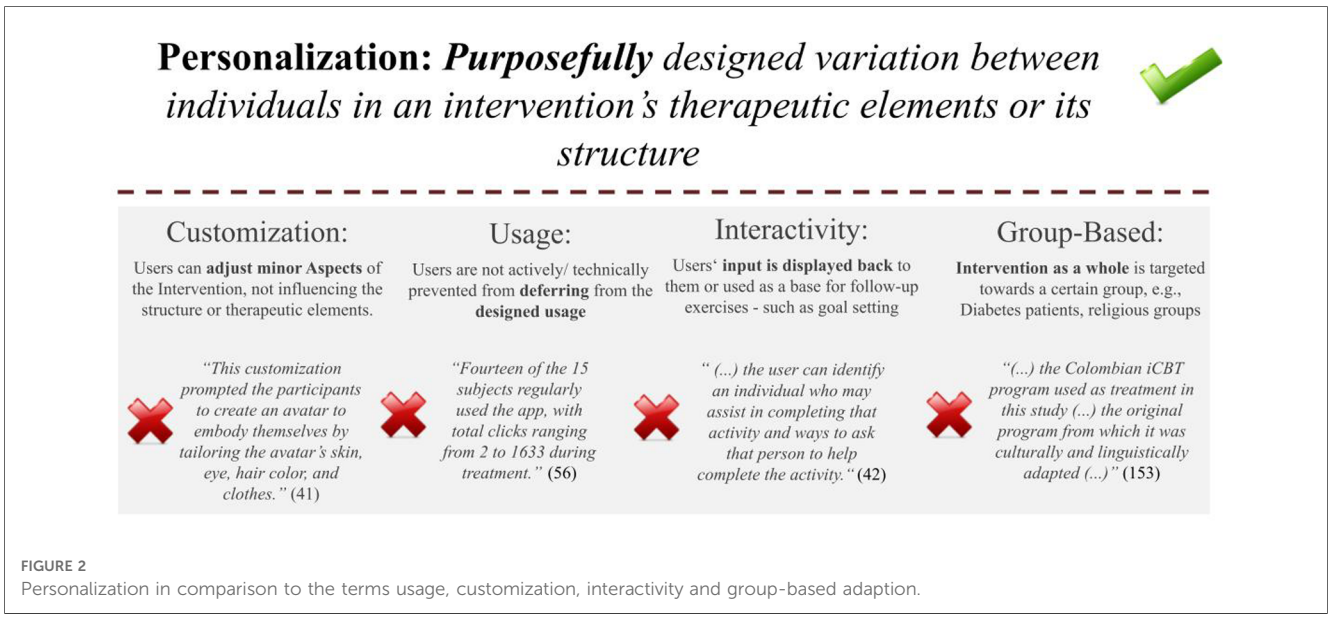


FIGURE 2
Personalization in comparison to the terms usage, customization, interactivity and group-based adaption.

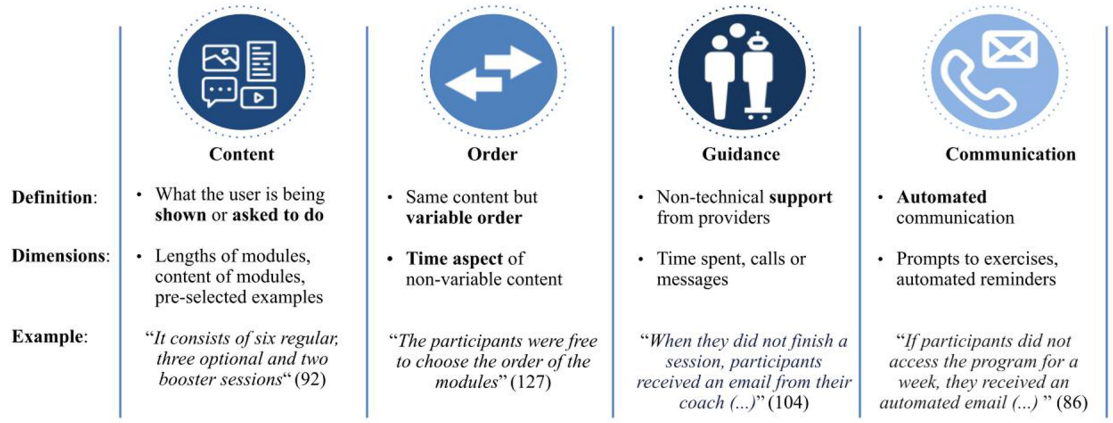


FIGURE 3
Dimensions of personalization.

additional information. In the single case where authors did not respond (39) we decided to code restrictively and assume the simpler of the potential mechanisms involved (rule-based instead of ML). Finally, for evaluating the evidence for personalization, we included every study that directly compared intervention versions that differed in their degree of personalization, according to our framework. We extracted effect sizes, dependent variables, and sample sizes for those.

3. Results

3.1. Study selection

Overall, we identified 3.143 potentially relevant publications and screened the title and abstract. For 213 of those, the full intervention description was reviewed, resulting in the final inclusion of $N = 138$ papers describing $k = 94$ distinct DMHIs for depressive symptoms (see **Figure 1**) (39–175).

3.2. Intervention and study characteristics

While mostly one study per intervention was included, for some up to seven publications on distinct trials were present and kept for further analysis. Across all studies, the reviewed interventions were deployed to approximately 24.300 participants, with an average sample size of 259 participants per intervention (range 1–2964). 75 of the interventions were used in a randomized controlled trial, with the remaining evidence coming from feasibility studies, naturalistic routine care data, and other study designs.

Most interventions had a duration between 6 and 12 weeks, and around 40 of the interventions report a structured module-/session-based design, delivering the content in pre-defined blocks. Finally, 38 interventions were unguided (no human contact within the intervention), 32 guided (including guidance from clinician or coach), and 14 blended (combining face-to-face and digital treatment), with the remaining 10 covering more than one of those categories. An overview of all characteristics can be found in **Supplementary Appendix S3**.

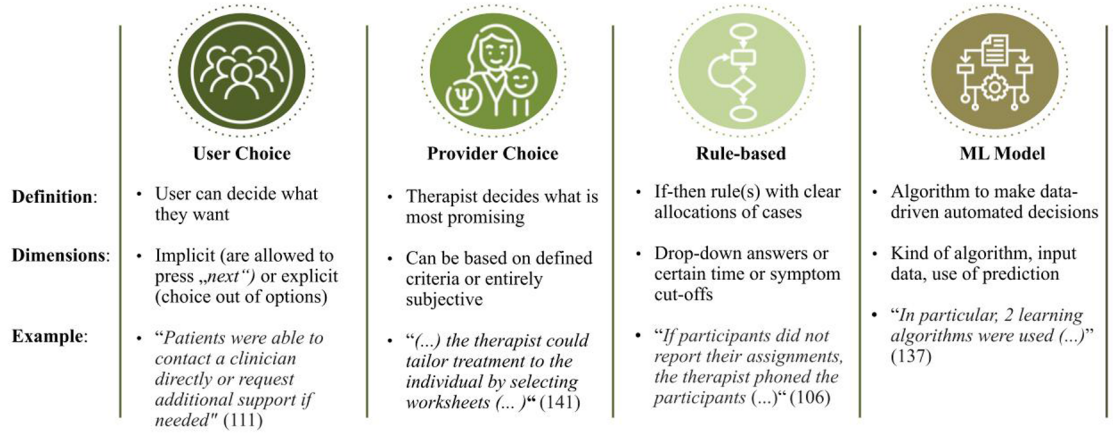


FIGURE 4
Mechanisms of personalization.

3.3. Conceptual framework

A conceptual framework of personalization in DMHIs was synthesized from the reviewed DMHIs and theoretical considerations. In summary, an understanding of personalization as *purposefully designed variation between individuals in an intervention's therapeutic elements or its structure* emerged. As such, personalization is differentiated from customization, usage, interactivity, and group-based adaptations. Customization describes minor adjustments, such as visual aspects, leaving the actual therapeutic ingredients unchanged. Usage refers to possible user-induced differences in app usage that were not actively or purposefully designed. For example, variability in the time spent on a module is usage, the offering of short and long versions of a module qualifies as personalization. Interactivity, the mere replay of user input as for example commonly used for goal-setting exercises, as this leaves the actual therapeutic elements and structure unchanged. Finally, as we understand personalization as referring to the level of the individual, we see it as being distinct from group-based variability, such as the adaptation for a particular cultural context (see **Figure 2**).

Within our definition of personalization, four personalizable intervention dimensions emerged, namely content, guidance level, order, and communication, as summarized in **Figure 3**. Content describes all variability in the delivered intervention material, such as exercises, psychoeducative material or topics presented. Order includes cases when patients receive the same content but in different order. Guidance refers to the extent of therapeutic support offered. Communication concerns the channel, timing, and content of actively offered information outside of the intervention's content. This primarily includes prompts or reminder messages. Mechanisms regarding the frequency and timing of the intervention, such as JIT mechanisms, also fall under communication.

Further, four different mechanisms beyond personalization emerged: user choice, provider choice, rule-based and ML-based personalization (see **Figure 4**). User choice covers intentionally designed personalization based on the direct choice of the participant. For provider choice, either the individual providing guidance, or the clinician involved in a blended setting makes the personalization decision. Among automated personalization mechanisms, rule-based (if-then-decision rules) from Machine Learning (decisions with "learned" decision criteria) personalization mechanisms are gathered.

3.4. Results on personalization

Applying the proposed framework for classifying variability in DMHIs, personalization was reported for 62 of the 94 interventions (66%). Most prominently, personalization mechanisms were used in the content for 30 of the interventions (32%). This was followed by personalized communication (30%), type (25%), and order (4%). 43 of the 62 (69%) interventions with a reported personalization mechanism did so for only a single dimension, while one DMHI reported a mechanism for all four subdomains of their intervention (60–66).

Across the 107 reported personalization mechanisms, rule-based was most prominent, being used in 51 cases (48%). User choice was observed in 39 cases (36%), and providers were involved in personalization 14 times (13%). The use of machine learning was reported three times (3%). Rule-based personalization was particularly prominent in the communication domain, accounting for 21 occurrences. Similarly, human guidance was personalized using decision rules 16 times. For content, user choice had a more prominent role, being reported 15 times. The use of personalization is summarized in **Figure 5**, with examples of the 3 most strategies being presented in **Table 1**. The share of interventions applying at least one personalization mechanism was the highest for guided interventions (72%), followed by unguided (63%) and tailed by blended (57%) interventions. Generally, the dimensions of personalization were equally spread across guidance levels. However, provider choice was nearly twice as common for blended than for guided interventions.

3.5. Use of automated decisions for personalization

Among the 55 automated mechanisms used, most were rule-based mechanisms of personalization. Here, activity data was heavily utilized, for example, for reminders in case of inactivity. Another common pattern was the use of symptom scores like the PHQ to step up care in the form of additional guidance (57) or the change from guided to blended care (169). While those approaches mostly used overall symptom severity, one exemption was the personalization based on suicide risk as e.g., in the form of additional prompts (146).

We identified three clear use cases of ML techniques for personalization. Firstly, EmoRecorder (70) used an activity recommendation system based on diverse data sources like app activity, sensor data and past recommendations. However, the intervention was at an early stage, being tested on a sample of only 15 healthy individuals. Secondly, the intervention MOSS (136) built on a JIT framework to assign intervention content depending on users' context and preferences. As such, it tested a recommender system with a sample of 126 adults. A third recommender system approach, so-called MUBS (137), applied a combination of ML and user choice by providing the 17 patients with a set of content recommendations.

3.6. Direct empirical comparison of more and less personalized interventions

Among the 138 papers in the final review, we identified two papers that included a direct comparison of a more and a less personalized version of an intervention. One study had participants fill out a questionnaire on motivational schemata and either matched them with an intervention arm to fit their

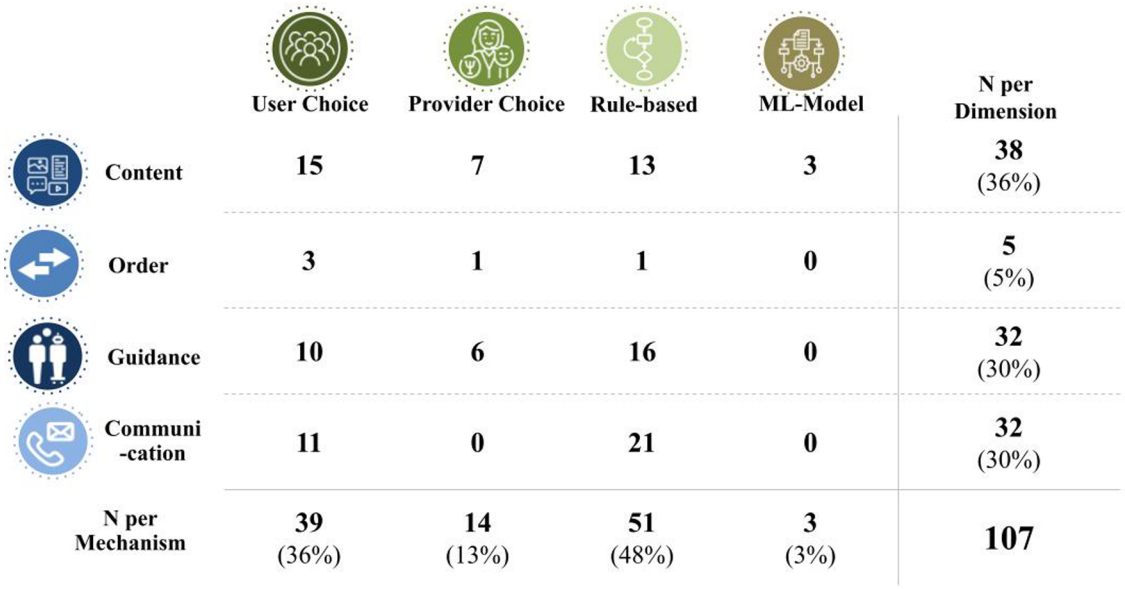


FIGURE 5
Personalization mechanisms per dimension of the intervention.

TABLE 1 Most prominent personalization strategies.

Number of interventions	Dimension	Mechanism	Description
21	Communication	Rule-based	e.g., Reminder for inactivity/non-completion
16	Guidance	Rule-based	e.g., Increased guidance/ clinician contact for symptom changes.
15	Context	User choice	e.g., Optional content selectable for patient.

motivational preference or a general one (40). Results showed effects for one of the two included motives ("being supported") on anticipated adherence, working alliance, and satisfaction; however, the overall sample size of this trial was just 55 participants. Secondly, a study compared a program version including JIT prompts with one without those prompts, therefore, differing the personalization in the communication domain between trial arms (93). While both versions showed significant effects compared to the waitlist, no effects were reported between the arms. Again, this should be interpreted with caution, considering the sample size of around 60 individuals per group.

4. Discussion

In recent years, personalization has been widely discussed as a promising avenue to improve DMHI adherence and outcomes. Nevertheless, it remains unclear what it entails and how it is used. In this review, we address this need for the case of depressive symptoms, by defining personalization as purposefully designed variation in intervention content, order, guidance, or communication. As possible mechanisms to operationalize personalization, we extract user choice, provider choice, decision rules, and ML. Applying this framework to 94 interventions for depressive symptoms reveals that two-thirds use at least one technique for personalization. Especially rule-based personalization of communication and guidance and user choice-based personalization of content is common. However, among interventions applying personalization, a majority does so just for one out of four dimensions of the intervention. Also, the use of ML models is scarce and limited to feasibility studies. Additionally, just two of the included studies investigated the benefits of personalization, both having small samples and just one finding supporting evidence.

Arguably, the biggest contrast between the proposed potentials in the personalization of DMHIs (19–22) and the existing literature is the lack of implemented ML mechanisms. Several of the implemented non-learning algorithms and decision rules were well designed. Yet, literature on ML in DMHIs reveals ample further promising and feasible use cases. Firstly, a notable body of research provides encouraging results in outcome (176, 177) and dropout (178, 179) predictions in DMHIs. Adapting the interventions for assumed non-responders is a low-hanging fruit and has already been successful for other disorders (180). Secondly, a prominent algorithmic approach to personalization in digital products is recommender systems (181–183). While all included ML approaches were such recommender systems, they were in early stages and deployed to very small sample sizes. Finally, all included ML approaches focused on the content of the intervention. However, ML also is a promising approach to personalize guidance, communication and order.

Contrasting theory and observations in another dimension, the data used for personalization just samples a fraction of the technically possible. While app usage patterns are an obvious data option, smartphones can also measure sleep patterns (184), physical activity (185), social interactions (186), and many other data points known to be relevant for depressive symptoms. Readily available toolkits like Apple's health kit (187) reduce the effort for implementation significantly. However, particularly passive sensing was rarely utilized in the reviewed interventions. Notably, the potential of ML-based personalization is heavily intertwined with the quality of the data available to them. Beyond that, aspects such as ethical responsibility in health care and privacy rights must be strongly considered, especially when investigating automated decisions (188).

Several interventions used self-reported symptoms for the personalization of the intervention. Noticeably, these mechanisms mostly used overall symptom severity. This approach disregards that symptom profiles can vary massively between patients with the same overall score (189). Some evidence points toward distinct symptom patterns being associated with different optimal treatment procedures (190). Therefore, while overall severity seems reasonable for varying guidance or communication, the sub-symptoms might be a promising ground for personalizing content and order.

The two included trials that manipulated personalization did so with small sample sizes and inconclusive results. Subsequently, one barrier to implementing personalization might be the lack of clear evidence for its benefits. However, RCTs investigating personalization are likely costly and require large sample sizes when assuming smaller effect sizes than for waitlist-controlled studies. Luckily, meta-analytic approaches allow summarizing evidence across studies, even when personalization is rarely directly manipulated. While we mentioned one such approach investigating interactions between individuals and benefits of iCBT packages (26), we consider similar approaches for other personalization mechanisms as very promising. However the identification and comparison of relevant studies in meta-analyses requires shared vocabulary and a common framework. We believe that such future work will benefit from the shared conceptual framework proposed in this article.

There are some limitations of this review that should be considered. Firstly, published studies are just one marker of what interventions are in use. While several included interventions originated in a commercial setting, those from academic settings will likely still be overrepresented in this review. Secondly, we focused on personalization within an intervention, excluding the personalization of interventions themselves. For example, past approaches investigated the data-driven personalization of therapy school (191) or the decision between medication and CBT (192). Thirdly, identifying interventions for depressive symptoms while excluding those addressing comorbid disorders, particularly anxiety, has proven challenging. One example is when anxiety was mentioned as intervention target in a cited study, but not in the original paper. While this seems understandable in light of the well-established comorbidity of depression and anxiety (193), this resulted in several edge cases of inclusion. Fourthly, we took interactivity, customization, and group-based adaptions out of the scope of this review due to their difference in nature to personalization. This should not be misunderstood as an assumed inferiority, and we call for the further investigation of these approaches to complement or even substitute personalization. Fifthly, we did not evaluate our framework by any methods besides the literature review. Approaches like expert interviews could help to determine and improve the usability of the proposed conceptualization. Sixthly, to provide a wide and less biased picture of the state of personalization, a broad search strategy was used. However, studies using more specific terminologies might be underrepresented. For example, a study on ecological momentary interventions (EMI) was not identified by our search strategy (194) as EMI was not used as a search term. Also, as pointed out by one of the reviewers, the mesh term "Telemedicine" was not used. Future approaches could therefore benefit from the application of additional techniques for iterating on the search strategy, such as the wider use of sentinel articles. Finally, as we developed our framework exclusively with studies on depressive symptoms, it remains unclear whether there are more aspects to consider with other disorders. However, we expect this framework to provide value beyond the use case of depressive symptoms and encourage future studies to investigate personalization strategies in other domains.

In conclusion, our conceptual development and empirical evaluation holistically characterizes the current use of personalization for DMHIs for depressive symptoms. A broad conceptualization of personalization reveals that most interventions incorporate personalization mechanisms. However, we conclude that we are barely scratching the surface of what is technically possible and already gold standard in other research and business areas. At the same time, we see the thin empirical ground as a barrier to implementation and call for more direct and meta-analytic evidence to delineate the benefits personalization has over an "one size fits all"-approach. Finally, as we see this question as equally pressing for other disorders, we hope for similar-minded approaches for those in the future.

Data Availability Statement

A file with all included studies as well as the related coding regarding the variables of interest can be found in Appendix 3. A file with the full search strings used can be found in Appendix 2. A file containing all studies screened during the selection process is available on request.

Author contributions

All authors designed the review and contributed to the reviews protocol. SH and KZ performed the literature search and data extraction. All authors contributed to the analysis of results and synthesis of insights. SH and KZ wrote the initial draft of the paper. All authors reviewed the draft and contributed to the final

paper. All authors contributed to the article and approved the submitted version.

Funding

The article processing charge was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 491192747 and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

Conflict of interest

Two of the authors declare no Competing Non-Financial Interests but the following Competing Financial Interests. SH is currently employed as Data Scientist by Elona Health, a digital mental health start-up building blended mental healthcare solutions for the German market. SH worked for Meru Health, a digital mental health company developing interventions, in the past. BF is a shareholder at HelloBetter, a digital mental health company developing digital interventions, and PersonalAIze, an AI consulting company.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2023.1170002/full#supplementary-material.

References

1. Lim GY, Tam WW, Lu Y, Ho CS, Zhang MW, Ho RC. Prevalence of depression in the community from 30 countries between 1994 and 2014. *Sci Rep*. (2018) 8:2861. doi: 10.1038/s41598-018-21243-x

2. Hasin DS, Sarvet AL, Meyers JL, Saha TD, Ruan WJ, Stohl M, et al. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry*. (2018) 75:336. doi: 10.1001/jamapsychiatry.2017.4602

3. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJL, et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLOS Med*. (2013) 10:e1001547. doi: 10.1371/journal.pmed.1001547

4. Wang PS, Demler O, Kessler RC. Adequacy of treatment for serious mental illness in the United States. *Am J Public Health*. (2002) 92:92–8. doi: 10.2105/ajph.92.1.92

5. Singer S, Engesser D, Wirp B, Lang K, Paserat A, Kobes J, et al. Effects of a statutory reform on waiting times for outpatient psychotherapy: a multicentre cohort study. *Couns Psychother Res*. (2022) 22:982–97. doi: 10.1002/capr.12581

6. Moroz N, Moroz I, D'Angelo MS, editors. Mental health services in Canada: barriers and cost-effective solutions to increase access. *Healthc Manag Forum*. (2020) 33:282–7. doi: 10.1177/0840470420933911

7. Tal A, Torous J. The digital mental health revolution: opportunities and risks. *Psychiatr Rehabil J*. (2017) 40:263–5. doi: 10.1037/prj0000285

8. Josephine K, Josefine L, Philipp D, David E, Harald B. Internet- and mobile-based depression interventions for people with diagnosed depression: a systematic review and meta-analysis. *J Affect Disord*. (2017) 223:28–40. doi: 10.1016/j.jad.2017.07.021

9. Moshe I, Terhorst Y, Philippe P, Domhardt M, Cuijpers P, Cristea I, et al. Digital interventions for the treatment of depression: a meta-analytic review. *Psychol Bull*. (2021) 147:749–86. doi: 10.1037/bul0000334

10. Donker T, Blankers M, Hedman E, Ljotsson B, Petrie K, Christensen H. Economic evaluations of internet interventions for mental health: a systematic review. *Psychol Med*. (2015) 45:3357–76. doi: 10.1017/S0033291715001427

11. International Telecommunication Union. Measuring digital development Facts and figures. *ITU*. (2019). https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ (Accessed December 15, 2022).

12. O'Dea S. Number of smartphone users worldwide from 2016 to 2021. *Statista*. (2021). https://www.statista.com/statistics/330695/number-of-smartphone-usersworld wide/ (Accessed December 15, 2022).

13. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. *J Affect Disord*. (2020) 263:413–9. doi: 10.1016/j.jad.2019.11.167

14. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijper P, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. *JAMA Psychiatry*. (2021) 78:361–71. doi: 10.1001/jamapsychiatry.2020.4364

15. Kaptein M, Parvinen P. Advancing e-commerce personalization: process framework and case study. *Int J Electron Commer*. (2015) 19:7–33. doi: 10.1080/10864415.2015.1000216

16. Zheng L, Long M, Zhong L, Gyasi JF. The effectiveness of technology-facilitated personalized learning on learning achievements and learning perceptions: a meta-analysis. *Edu Inf Tech*. (2022) 27:11807–30. doi: 10.1007/s10639-022-11092-7

17. Shanahan T, Tran TP, Taylor EC. Getting to know you: social media personalization as a means of enhancing brand loyalty and perceived quality. *J Retail Consum Serv*. (2019) 47:57–65. doi: 10.1016/j.jretconser.2018.10.007

18. Swift JK, Callahan JL, Cooper M, Parkin SR. The impact of accommodating client preference in psychotherapy: a meta-analysis. *J Clin Psychol*. (2018) 74:1924–37. doi: 10.1002/jclp.22680

19. Aung MH, Matthews M, Choudhury T. Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies. *Depress Anxiety*. (2017) 34:603–9. doi: 10.1002/da.22646

20. D'Alfonso S. AI In mental health. *Curr Opin Psychol*. (2020) 36:112–7. doi: 10.1016/j.copsyc.2020.04.005

21. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. *J Gen Intern Med*. (2013) 28:660–5. doi: 10.1007/s11606-013-2455-8

22. Andrews G, Williams AD. Internet psychotherapy and the future of personalized treatment. *Depress Anxiety*. (2014) 31:912–5. doi: 10.1002/da.22302

23. Lustria ML, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. *J Health Commun*. (2013) 18(9):1039–69. doi: 10.1080/10810730.2013.768727

24. Teepe GW, Da Fonseca A, Kleim B, Jacobson NC, Salamanca Sanabria A, Tudor Car L, et al. Just-in-time adaptive mechanisms of popular mobile apps for individuals with depression: systematic app search and literature review. *J Med Internet Res*. (2021) 23:e2941. doi: 10.2196/29412:9

25. Moe-Byrne T, Shepherd J, Merecz-Kot D, Sinokki M, Naumanen P, Hakkaart-van Roijen L, et al. Effectiveness of tailored digital health interventions for mental health at the workplace: a systematic review of randomised controlled trials. *PLOS Dig Heal*. (2022) 1:e0000123. doi: 10.1371/journal.pdig.0000123

26. Furukawa TA, Suganuma A, Ostinelli EG, Andersson G, Beevers CG, Shumake J, et al. Dismantling, optimising, and personalising internet cognitive behavioural

therapy for depression: a systematic review and component network meta-analysis using individual participant data. *Lancet Psychiat*. (2021) 8:500–11. doi: 10.1016/S2215-0366(21)00077-8

27. Ta Park VM, Ton V, Yeo G, Tiet QQ, Vuong Q, Gallagher-Thompson D. Vietnamese American dementia caregivers' perceptions and experiences of a culturally tailored, evidence-based program to reduce stress and depression. *J Gerontol Nurs*. (2019) 45:39–50. doi: 10.3928/00989134-20190813-05

28. Zagorscak P, Heinrich M, Bohn J, Stein J, Knaevelsrud C. How individuals change during internet-based interventions for depression: a randomized controlled trial comparing standardized and individualized feedback. *Brain Behav*. (2020) 10: e01484. doi: 10.1002/brb3.1484

29. Lau Y, Chee DGH, Chow XP, Cheng LJ, Wong SN. Personalised eHealth interventions in adults with overweight and obesity: a systematic review and meta-analysis of randomised controlled trials. *Prev Med*. (2020) 132:106001. doi: 10.1016/j.ypmed.2020.106001

30. Captari LE, Hook JN, Hoyt W, Davis DE, McElroy-Heltzel SE, Worthington EL Jr. Integrating clients' religion and spirituality within psychotherapy: a comprehensive meta-analysis. *J Clin Psychol*. (2018) 74:1938–51. doi: 10.1002/jclp.22681

31. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. (2009) 6(7):e1000097. doi: 10.1371/journal.pmed.1000097

32. Himle JA, Weaver A, Zhang A, Xiang X. Digital mental health interventions for depression. *Cogn Behav Pract*. (2022) 29(1):50–9. doi: 10.1016/j.cbpra.2020.12.009

33. Suprenant CF, Solomon MR. Predictability and personalization in the service encounter. *J Mark*. (1987) 51(2):86–96. doi: 10.1177/002224298705100207

34. Deighton J, Sorrell M. The future of interactive marketing. *Harv Bus Rev*. (1996) 74(6):151–60. https://hbr.org/1996/11/the-future-of-interactive-marketing

35. Sundar SS, Marathe SS. Personalization versus customization: the importance of agency, privacy, and power usage. *Hum Commun Res*. (2010) 36(3):298–322. doi: 10.1111/j.1468-2958.2010.01377.x

36. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet-and mobile-based interventions for mental disorders: a systematic review. *NPJ Digit Med*. (2021) 4(1):1–18. doi: 10.1038/s41746-021-00498-1

37. Luborsky L, DeRubeis RJ. The use of psychotherapy treatment manuals: a small revolution in psychotherapy research style. *Clin Psychol Rev*. (1984) 4(1):5–14. doi: 10.1016/0272-7358(84)90034-5

38. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. *JAMA Psychiatry*. (2021) 78 (4):361–71. doi: 10.1001/jamapsychiatry.2020.4364

39. Burton C, Szentagotai Tatar A, McKinstry B, Matheson C, Matu S, Moldovan R, et al. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. *J Telemed Telecare*. (2016) 22 (6):348–55. doi: 10.1177/1357633X15609793

40. Bücker L, Berger T, Bruhns A, Westermann S. Motive-oriented, personalized, internet-based interventions for depression: nonclinical experimental study. *JMIR Form Res*. (2022) 6(9):e37287. doi: 10.2196/37287

41. Schuster R, Leitner I, Carlbring P, Laireiter AR. Exploring blended group interventions for depression: randomised controlled feasibility study of a blended computer-and multimedia-supported psychoeducational group intervention for adults with depressive symptoms. *Internet Interv*. (2017) 8:63–71. doi: 10.1016/j.invent.2017.04.001

42. Six SG, Byrne KA, Aly H, Harris MW. The effect of mental health app customization on depressive symptoms in college students: randomized controlled trial. *JMIR Ment Health*. (2022) 9(8):e39516. doi: 10.2196/39516

43. Dahne J, Collado A, Lejuez CW, Risco CM, Diaz VA, Coles L, et al. Pilot randomized controlled trial of a Spanish-language Behavioral Activation mobile app (¡ Aptívate!) for the treatment of depressive symptoms among United States Latinx adults with limited English proficiency. *J Affect Disord*. (2019) 250:210–7. doi: 10.1016/j.jad.2019.03.009

44. Dahne J, Lejuez CW, Diaz VA, Player MS, Kustanowitz J, Felton JW, et al. Pilot randomized trial of a self-help behavioral activation mobile app for utilization in primary care. *Behav Ther*. (2019) 50(4):817–27. doi: 10.1016/j.beth.2018.12.003

45. Pérez JC, Fernández O, Cáceres C, Carrasco ÁE, Moessner M, Bauer S, et al. An adjunctive internet-based intervention to enhance treatment for depression in adults: randomized controlled trial. *JMIR Ment Health*. (2021) 8(12):e26814. doi: 10.2196/26814

46. Lüdtke T, Pult LK, Schröder J, Moritz S, Bücker L. A randomized controlled trial on a smartphone self-help application (be good to yourself) to reduce depressive symptoms. *Psychiatry Res*. (2018) 269:753–62. doi: 10.1016/j.psychres.2018.08.113

47. Forand NR, Barnett JG, Strunk DR, Hindiyeh MU, Feinberg JE, Keefe JR. Efficacy of guided iCBT for depression and mediation of change by cognitive skill acquisition. *Behav Ther*. (2018) 49(2):295–307. doi: 10.1016/j.beth.2017.04.004

48. Pfeiffer PN, Pope B, Houck M, Benn-Burton W, Zivin K, Ganoczy D, et al. Effectiveness of peer-supported computer-based CBT for depression among veterans in primary care. *Psychiatr Serv*. (2020) 71(3):256–62. doi: 10.1176/appi.ps.201900283

49. Gupta SK, Slaven JE, Liu Z, Polanka BM, Freiberg MS, Stewart JC. Effects of internet cognitive-behavioral therapy on depressive symptoms and surrogates of cardiovascular risk in human immunodeficiency virus: a pilot, randomized, controlled trial. *Open Forum Infect Dis*. (2020) 7(7):ofaa280. doi: 10.1093/ofid/ofaa280

50. Xiang X, Kayser J, Sun Y, Himle J. Internet-based psychotherapy intervention for depression among older adults receiving home care: qualitative study of participants' experiences. *JMIR Aging*. (2021) 4(4):e27630. doi: 10.2196/27630

51. Littlewood E, Duarte A, Hewitt C, Knowles S, Palmer S, Walker S, et al. A randomised controlled trial of computerised cognitive behaviour therapy for the treatment of depression in primary care: the randomised evaluation of the effectiveness and acceptability of computerised therapy (REEACT) trial. *Health Technol Assess*. (2015) 19(101):viii–171. doi: 10.3310/hta191010

52. Fuller-Tyszkiewicz M, Richardson B, Klein B, Skouteris H, Christensen H, Austin D. A mobile app-based intervention for depression: end-user and expert usability testing study. *JMIR Ment Health*. (2018) 5(3):e54. doi: 10.2196/mental.9445

53. Stiles-Shields C, Montague E, Kwasny MJ, Mohr DC. Behavioral and cognitive intervention strategies delivered via coached apps for depression: pilot trial. *Psychol Serv*. (2019) 16(2):233–38. doi: 10.1037/ser0000261

54. Blackwell SE, Browning M, Mathews A, Pictet A, Welch J, Davies J, et al. Positive imagery-based cognitive bias modification as a web-based treatment tool for depressed adults: a randomized controlled trial. *Clin Psychol Sci*. (2015) 3(1):91–111. doi: 10.1177/2167702614560746

55. Williams AD, O'Moore K, Blackwell SE, Smith J, Holmes EA, Andrews G. Positive imagery cognitive bias modification (CBM) and internet-based cognitive behavioral therapy (iCBT): a randomized controlled trial. *J Affect Disord*. (2015) 178:131–41. doi: 10.1016/j.jad.2015.02.026

56. Pictet A, Jermann F, Ceschi G. When less could be more: investigating the effects of a brief internet-based imagery cognitive bias modification intervention in depression. *Behav Res Ther*. (2016) 84:45–51. doi: 10.1016/j.brat.2016.07.008

57. Callan JA, Dunbar Jacob J, Siegle G, Dey A, Thase M, DeVito Dabbs A, et al. CBT Mobilework©: user-centered development and testing of a mobile mental health application for depression. *Cogn Ther Res*. (2021) 45:287–302. doi: 10.1007/s10608-020-10159-4

58. Ritvo P, Knyahnytska Y, Pirbaglou M, Wang W, Tomlinson G, Zhao H, et al. Online mindfulness-based cognitive behavioral therapy intervention for youth with major depressive disorders: randomized controlled trial. *J Med Internet Res*. (2021) 23(3):e24380. doi: 10.2196/24380

59. Eriksson MCM, Kivi M, Hange D, Petersson EL, Ariai N, Häggblad P, et al. Long-term effects of internet-delivered cognitive behavioral therapy for depression in primary care–the PRIM-NET controlled trial. *Scand J Prim Health Care*. (2017) 35(2):126–36. doi: 10.1080/02813432.2017.1333299

60. Beevers CG, Pearson R, Hoffman JS, Foulser AA, Shumake J, Meyer B. Effectiveness of an internet intervention (Deprexis) for depression in a United States adult sample: a parallel-group pragmatic randomized controlled trial. *J Consult Clin Psychol*. (2017) 85(4):367–80. doi: 10.1037/ccp0000171

61. Zwerenz R, Becker J, Knickenberg RJ, Siepmann M, Hagen K, Beutel ME. Online self-help as an add-on to inpatient psychotherapy: efficacy of a new blended treatment approach. *Psychother Psychosom*. (2017) 86(6):341–50. doi: 10.1159/000481177

62. Richter LE, Machleit-Ebner A, Scherbaum N, Bonnet U. How effective is a web-based mental health intervention (Deprexis) in the treatment of moderate and major depressive disorders when started during routine psychiatric inpatient treatment as an adjunct therapy? A pragmatic parallel-group randomized controlled trial. *Fortschr Neurol Psychiatr*. (2022) 04:10.1055/a-1826-2888. doi: 10.1055/a-1826-2888

63. Klein JP, Berger T, Schröder J, Späth C, Meyer B, Caspar F, et al. Effects of a psychological internet intervention in the treatment of mild to moderate depressive symptoms: results of the EVIDENT study, a randomized controlled trial. *Psychother Psychosom*. (2016) 85(4):218–28. doi: 10.1159/000445355

64. Gräfe V, Moritz S, Greiner W. Health economic evaluation of an internet intervention for depression (deprexis), a randomized controlled trial. *Health Econ Rev*. (2020) 10(1):19. doi: 10.1186/s13561-020-00273-0

65. Meyer B, Bierbrodt J, Schröder J, Berger T, Beevers CG, Weiss M, et al. Effects of an internet intervention (Deprexis) on severe depression symptoms: randomized controlled trial. *Internet Interv*. (2015) 2(1):48–59. doi: 10.1016/j.invent.2014.12.003

66. Fischer A, Schröder J, Vettorazzi E, Wolf OT, Pöttgen J, Lau S, et al. An online programme to reduce depression in patients with multiple sclerosis: a randomised controlled trial. *Lancet Psychiatry*. (2015) 2(3):217–23. doi: 10.1016/S2215-0366(14)00049-2

67. Crisp DA, Griffiths KM. Reducing depression through an online intervention: benefits from a user perspective. *JMIR Ment Health*. (2016) 3(1):e4. doi: 10.2196/mental.4356

68. Iacoviello BM, Murrough JW, Hoch MM, Huryk KM, Collins KA, Cutter GR, et al. A randomized, controlled pilot trial of the emotional faces memory task: a digital therapeutic for depression. *NPJ Digit Med*. (2018) 1:21. doi: 10.1038/s41746-018-0025-5

69. Hoch MM, Doucet GE, Moser DA, Hee Lee W, Collins KA, Huryk KM, et al. Initial evidence for brain plasticity following a digital therapeutic intervention for depression. *Chronic Stress*. (2019) 3:2470547019877880. doi: 10.1177/2470547019877880

70. Hung GC, Yang PC, Wang CY, Chiang JH. *A smartphone-based personalized activity recommender system for patients with depression*. In: *Proceedings of the 5th EAI international conference on wireless mobile communication and healthcare*; 2015 Oct 14–16; London, Great Britain. Belgium: Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (2015). p. 253–7. doi: 10.4108/eai.14-10-2015.2261655

71. Pinto MD, Greenblatt AM, Hickman RL, Rice HM, Thomas TL, Clochesy JM. Assessing the critical parameters of eSMART-MH: a promising avatar-based digital therapeutic intervention to reduce depressive symptoms. *Perspect Psychiatr Care*. (2016) 52(3):157–68. doi: 10.1111/ppc.12112

72. Arjadi R, Nauta MH, Scholte WF, Hollon SD, Chowdhary N, Suryani AO, et al. Internet-based behavioural activation with lay counsellor support versus online minimal psychoeducation without support for treatment of depression: a randomised controlled trial in Indonesia. *Lancet Psychiatry*. (2018) 5(9):707–16. doi: 10.1016/S2215-0366(18)30223-2

73. Buntrock C, Ebert D, Lehr D, Riper H, Smit F, Cuijpers P, et al. Effectiveness of a web-based cognitive behavioural intervention for subthreshold depression: pragmatic randomised controlled trial. *Psychother Psychosom*. (2015) 84(6):348–58. doi: 10.1159/000438673

74. Reins JA, Boß L, Lehr D, Berking M, Ebert DD. The more I got, the less I need? Efficacy of internet-based guided self-help compared to online psychoeducation for major depressive disorder. *J Affect Disord*. (2019) 246:695–705. doi: 10.1016/j.jad.2018.12.065

75. Ebert DD, Buntrock C, Lehr D, Smit F, Riper H, Baumeister H, et al. Effectiveness of web-and mobile-based treatment of subthreshold depression with adherence-focused guidance: a single-blind randomized controlled trial. *Behav Ther*. (2018) 49(1):71–83. doi: 10.1016/j.beth.2017.05.004

76. Braun L, Titzler I, Terhorst Y, Freund J, Thielecke J, Ebert DD, et al. Are guided internet-based interventions for the indicated prevention of depression in green professions effective in the long run? Longitudinal analysis of the 6-and 12-month follow-up of a pragmatic randomized controlled trial (PROD-A). *Internet Interv*. (2021) 26:100455. doi: 10.1016/j.invent.2021.100455

77. Ofoegbu TO, Asogwa U, Otu MS, Ibenegbu C, Muhammed A, Eze B. Efficacy of guided internet-assisted intervention on depression reduction among educational technology students of Nigerian universities. *Medicine (Baltimore)*. (2020) 99(6):e18774. doi: 10.1097/MD.0000000000018774

78. Thase ME, Wright JH, Eells TD, Barrett MS, Wisniewski SR, Balasubramani GK, et al. Improving the efficiency of psychotherapy for depression: computer-assisted versus standard CBT. *Am J Psychiatry*. (2018) 175(3):242–50. doi: 10.1176/appi.ajp.2017.17010089

79. Wright JH, Owen J, Eells TD, Antle B, Bishop LB, Girdler R, et al. Effect of computer-assisted cognitive behavior therapy vs usual care on depression among adults in primary care: a randomized clinical trial. *JAMA Netw Open*. (2022) 5(2):e2146716. doi: 10.1001/jamanetworkopen.2021.46716

80. Deady M, Johnston D, Milne D, Glozier N, Peters D, Calvo R, et al. Preliminary effectiveness of a smartphone app to reduce depressive symptoms in the workplace: feasibility and acceptability study. *JMIR Mhealth Uhealth*. (2018) 6(12):e11661. doi: 10.2196/11661

81. Strauss C, Dunkeld C, Cavanagh K. Is clinician-supported use of a mindfulness smartphone app a feasible treatment for depression? A mixed-methods feasibility study. *Internet Interv*. (2021) 25:100413. doi: 10.1016/j.invent.2021.100413

82. Fish MT, Saul AD. The gamification of meditation: a randomized-controlled study of a prescribed mobile mindfulness meditation application in reducing college students' depression. *Simul Gaming*. (2019) 50(4):419–35. doi: 10.1177/1046878119851821

83. Ying Y, Ji Y, Kong F, Wang M, Chen Q, Wang L, et al. Efficacy of an internet-based cognitive behavioral therapy for subthreshold depression among Chinese adults: a randomized controlled trial. *Psychol Med*. (2022):1–11. doi: 10.1017/S0033291722000599

84. Beiwinkel T, Eißing T, Telle NT, Siegmund-Schultze E, Rössler W. Effectiveness of a web-based intervention in reducing depression and sickness absence: randomized controlled trial. *J Med Internet Res*. (2017) 19(6):e6546. doi: 10.2196/jmir.6546

85. Bur OT, Krieger T, Moritz S, Klein JP, Berger T. Optimizing the context of support of web-based self-help in individuals with mild to moderate depressive symptoms: a randomized full factorial trial. *Behav Res Ther*. (2022) 152:104070. doi: 10.1016/j.brat.2022.104070

86. Gili M, Castro A, García-Palacios A, Garcia-Campayo J, Mayoral-Cleries F, Botella C, et al. Efficacy of three low-intensity, internet-based psychological interventions for the treatment of depression in primary care: randomized controlled trial. *J Med Internet Res*. (2020) 22(6):e15845. doi: 10.2196/15845

87. Zhao C, Wampold BE, Ren Z, Zhang L, Jiang G. The efficacy and optimal matching of an internet-based acceptance and commitment therapy intervention for depressive symptoms among university students: a randomized controlled trial in China. *J Clin Psychol*. (2022) 78(7):1354–375. doi: 10.1002/jclp.23329

88. Oehler C, Scholze K, Driessen P, Rummel-Kluge C, Görges F, Hegerl U. How are guide profession and routine care setting related to adherence and symptom change in iCBT for depression?—an explorative log-data analysis. *Internet Interv*. (2021) 26:100476. doi: 10.1016/j.invent.2021.100476

89. Varga A, Czeglédi E, Tóth MD, Purebl G. Effectiveness of iFightDepression® online guided self-help tool in depression–a pilot study. *J Telemed Telecare*. (2022) 0(0):1357633X221084584. doi: 10.1177/1357633X221084584

90. Schuster R, Fischer E, Jansen C, Napravnik N, Rockinger S, Steger N, et al. Blending internet-based and tele group treatment: acceptability, effects, and mechanisms of change of cognitive behavioral treatment for depression. *Internet Interv*. (2022) 29:100551. doi: 10.1016/j.invent.2022.100551

91. Oehler C, Scholze K, Reich H, Sander C, Hegerl U. Intervention use and symptom change with unguided internet-based cognitive behavioral therapy for depression during the COVID-19 pandemic: log data analysis of a convenience sample. *JMIR Ment Health*. (2021) 8(7):e28321. doi: 10.2196/28321

92. Paganini S, Terhorst Y, Sander LB, Lin J, Schlicker S, Ebert DD, et al. Internet- and mobile-based intervention for depression in adults with chronic back pain: a health economic evaluation. *J Affect Disord*. (2022) 308:607–15. doi: 10.1016/j.jad.2022.04.004

93. Everitt N, Broadbent J, Richardson B, Smyth JM, Heron K, Teague S, et al. Exploring the features of an app-based just-in-time intervention for depression. *J Affect Disord*. (2021) 291:279–87. doi: 10.1016/j.jad.2021.05.021

94. El Alaoui S, Ljótsson B, Hedman E, Svanborg C, Kaldo V, Lindefors N. Predicting outcome in internet-based cognitive behaviour therapy for major depression: a large cohort study of adult patients in routine psychiatric care. *PLoS One*. (2016) 11(9):e0161191. doi: 10.1371/journal.pone.0161191

95. Anguera JA, Jordan JT, Castaneda D, Gazzaley A, Areán PA. Conducting a fully mobile and randomised clinical trial for depression: access, engagement and expense. *BMJ Innov*. (2016) 2(1):14–21. doi: 10.1136/bmjinnov-2015-000098

96. Pratap A, Renn BN, Volponi J, Mooney SD, Gazzaley A, Arean PA, et al. Using mobile apps to assess and treat depression in Hispanic and Latino populations: fully remote randomized clinical trial. *J Med Internet Res*. (2018) 20(8):e10130. doi: 10.2196/10130

97. Mol M, Dozeman E, Provoost S, van Schaik A, Riper H, Smit JH. Behind the scenes of online therapeutic feedback in blended therapy for depression: mixed-methods observational study. *J Med Internet Res*. (2018) 20(5):e174. doi: 10.2196/jmir.9890

98. Noguchi R, Sekizawa Y, So M, Yamaguchi S, Shimizu E. Effects of five-minute internet-based cognitive behavioral therapy and simplified emotion-focused mindfulness on depressive symptoms: a randomized controlled trial. *BMC Psychiatry*. (2017) 17(1):1–14. doi: 10.1186/s12888-017-1248-8

99. Sweet AM, Pearlstein SL, Paulus MP, Stein MB, Taylor CT. Computer-delivered behavioural activation and approach-avoidance training in major depression: proof of concept and initial outcomes. *Br J Clin Psychol*. (2021) 60(3):357–74. doi: 10.1111/bjc.12287

100. Murillo LA, Follo E, Smith A, Balestrier J, Bevvino DL. Evaluating the effectiveness of online educational modules and interactive workshops in alleviating symptoms of mild to moderate depression: a pilot trial. *J Prim Care Community Health*. (2020) 11:2150132720971158. doi: 10.1177/2150132720971158

101. Kingston J, Becker L, Woeginger J, Ellett LA. Randomised trial comparing a brief online delivery of mindfulness-plus-values versus values only for symptoms of depression: does baseline severity matter? *J Affect Disord*. (2020) 276:936–44. doi: 10.1016/j.jad.2020.07.087

102. Lüdtke T, Westermann S, Pult LK, Schneider BC, Pfuhl G, Moritz S. Evaluation of a brief unguided psychological online intervention for depression: a controlled trial including exploratory moderator analyses. *Internet Interv*. (2018) 13:73–81. doi: 10.1016/j.invent.2018.06.004

103. Ly KH, Topooco N, Cederlund H, Wallin A, Bergström J, Molander O, et al. Smartphone-supported versus full behavioural activation for depression: a randomised controlled trial. *PLoS One*. (2015) 10(5):e0126559. doi: 10.1371/journal.pone.0126559

104. Kenter RM, Cuijpers P, Beekman A, van Straten A. Effectiveness of a web-based guided self-help intervention for outpatients with a depressive disorder: short-term results from a randomized controlled trial. *J Med Internet Res*. (2016) 18(3):e4861. doi: 10.2196/jmir.4861

105. Nyström MB, Stenling A, Sjöström E, Neely G, Lindner P, Hassmén P, et al. Behavioral activation versus physical activity via the internet: a randomized controlled trial. *J Affect Disord*. (2017) 215:85–93. doi: 10.1016/j.jad.2017.03.018

106. Johansson O, Bjärehed J, Andersson G, Carlbring P, Lundh LG. Effectiveness of guided internet-delivered cognitive behavior therapy for depression in routine psychiatry: a randomized controlled trial. *Internet Interv*. (2019) 17:100247. doi: 10.1016/j.invent.2019.100247

107. Nordgreen T, Blom K, Andersson G, Carlbring P, Havik OE. Effectiveness of guided internet-delivered treatment for major depression in routine mental healthcare—an open study. *Internet Interv*. (2019) 18:100274. doi: 10.1016/j.invent.2019.100274

108. Jakobsen H, Andersson G, Havik OE, Nordgreen T. Guided internet-based cognitive behavioral therapy for mild and moderate depression: a benchmarking study. *Internet Interv*. (2017) 7:1–8. doi: 10.1016/j.invent.2016.11.002

109. Nygren T, Brohede D, Koshnaw K, Osman SS, Johansson R, Andersson G. Internet-based treatment of depressive symptoms in a Kurdish population: a randomized controlled trial. *J Clin Psychol*. (2019) 75(6):985–98. doi: 10.1002/jclp.22753

110. Tulbure BT, Andersson G, Sălăgean N, Pearce M, Koenig HG. Religious versus conventional internet-based cognitive behavioral therapy for depression. *J Relig Health*. (2018) 57(5):1634–648. doi: 10.1007/s10943-017-0503-0

111. Hallgren M, Kraepelien M, Öjehagen A, Lindefors N, Zeebari Z, Kaldo V, et al. Physical exercise and internet-based cognitive–behavioural therapy in the treatment of depression: randomised controlled trial. *Br J Psychiatry*. (2015) 207(3):227–34. doi: 10.1192/bjp.bp.114.160101

112. Wong VWH, Ho FY, Shi NK, Tong JT, Chung KF, Yeung WF, et al. Smartphone-delivered multicomponent lifestyle medicine intervention for depressive symptoms: a randomized controlled trial. *J Consult Clin Psychol*. (2021) 89 (12):970–84. doi: 10.1037/ccp0000695

113. Salisbury C, O'Cathain A, Edwards L, Thomas C, Gaunt D, Hollinghurst S, et al. Effectiveness of an integrated telehealth service for patients with depression: a pragmatic randomised controlled trial of a complex intervention. *Lancet Psychiatry*. (2016) 3(6):515–25. doi: 10.1016/S2215-0366(16)00083-3

114. Pots WT, Fledderus M, Meulenbeek PA, ten Klooster PM, Schreurs KM, Bohlmeijer ET. Acceptance and commitment therapy as a web-based intervention for depressive symptoms: randomised controlled trial. *Br J Psychiatry*. (2016) 208 (1):69–77. doi: 10.1192/bjp.bp.114.146068

115. Kelders SM, Bohlmeijer ET, Pots WT, van Gemert-Pijnen JE. Comparing human and automated support for depression: fractional factorial randomized controlled trial. *Behav Res Ther*. (2015) 72:72–80. doi: 10.1016/j.brat.2015.06.014

116. Rauen K, Vetter S, Eisele A, Biskup E, Delsignore A, Rufer M, et al. Internet cognitive behavioral therapy with or without face-to-face psychotherapy: a 12-weeks clinical trial of patients with depression. *Front Digit Health*. (2020) 2:4. doi: 10.3389/fdgth.2020.00004

117. Addington EL, Cheung EO, Bassett SM, Kwok I, Schuette SA, Shiu E, et al. The MARIGOLD study: feasibility and enhancement of an online intervention to improve emotion regulation in people with elevated depressive symptoms. *J Affect Disord*. (2019) 257:352–64. doi: 10.1016/j.jad.2019.07.049

118. Moskowitz JT, Addington EL, Shiu E, Bassett SM, Schuette S, Kwok I, et al. Facilitator contact, discussion boards, and virtual badges as adherence enhancements to a web-based, self-guided, positive psychological intervention for depression: randomized controlled trial. *J Med Internet Res*. (2021) 23(9):e25922. doi: 10.2196/25922

119. Visser DA, Tendolkar I, Schene AH, van de Kraats L, Ruhe HG, Vrijsen JN, et al. A pilot study of smartphone-based memory bias modification and its effect on memory bias and depressive symptoms in an unselected population. *Cognit Ther Res*. (2020) 44(1):61–72. doi: 10.1007/s10608-019-10042-x

120. Bruhns A, Lüdtke T, Moritz S, Bücker L. A mobile-based intervention to increase self-esteem in students with depressive symptoms: randomized controlled trial. *JMIR Mhealth Uhealth*. (2021) 9(7):e26498. doi: 10.2196/26498

121. Lukas CA, Eskofier B, Berking M. A gamified smartphone-based intervention for depression: randomized controlled pilot trial. *JMIR Ment Health*. (2021) 8(7): e16643. doi: 10.2196/16643

122. Lukas CA, Berking M. Blending group-based psychoeducation with a smartphone intervention for the reduction of depressive symptoms: results of a randomized controlled pilot study. *Pilot Feasibility Stud*. (2021) 7(1):1–8. doi: 10.1186/s40814-021-00799-y

123. Raevuori A, Vahlberg T, Korhonen T, Hilgert O, Aittakumpu-Hyden R, Forman-Hoffman V. A therapist-guided smartphone app for major depression in young adults: a randomized clinical trial. *J Affect Disord*. (2021) 286:228–38. doi: 10.1016/j.jad.2021.02.007

124. Economides M, Lehrer P, Ranta K, Nazander A, Hilgert O, Raevuori A, et al. Feasibility and efficacy of the addition of heart rate variability biofeedback to a remote digital health intervention for depression. *Appl Psychophysiol Biofeedback*. (2020) 45 (2):75–86. doi: 10.1007/s10484-020-09458-z

125. Forman-Hoffman VL, Nelson BW, Ranta K, Nazander A, Hilgert O, de Quevedo J. Significant reduction in depressive symptoms among patients with moderately-severe to severe depressive symptoms after participation in a therapist-supported, evidence-based mobile health program delivered via a smartphone app. *Internet Interv*. (2021) 25:100408. doi: 10.1016/j.invent.2021.100408

126. Bücker L, Schnakenberg P, Karyotaki E, Moritz S, Westermann S. Diminishing effects after recurrent use of self-guided internet-based interventions in depression: randomized controlled trial. *J Med Internet Res*. (2019) 21(10):e14240. doi: 10.2196/14240

127. Titzler I, Egle V, Berking M, Gumbmann C, Ebert DD. Blended psychotherapy: treatment concept and case report for the integration of internet-and mobile-based interventions into brief psychotherapy of depressive disorders. *Verhaltenstherapie*. (2019) 32(1):1–15. doi: 10.1159/000503408

128. Wang F, Feng F, Zhang J, Cooper A, Hong L, Wang W, et al. Outcomes of an online computerized cognitive behavioral treatment program for treating Chinese patients with depression: a pilot study. *Asian J Psychiatr*. (2018) 38:102–7. doi: 10.1016/j.ajp.2017.11.007

129. Brabyn S, Araya R, Barkham M, Bower P, Cooper C, Duarte A, et al. The second randomised evaluation of the effectiveness, cost-effectiveness and acceptability of computerised therapy (REEACT-2) trial: does the provision of telephone support enhance the effectiveness of computer-delivered cognitive behaviour therapy? A randomised controlled trial. *Health Technol Assess*. (2016) 20 (89):1–64. doi: 10.3310/hta20890

130. Silverstone PH, Rittenbach K, Suen VY, Moretzsohn A, Cribben I, Bercov M, et al. Depression outcomes in adults attending family practice were not improved by screening, stepped-care, or online CBT during a 12-week study when compared to controls in a randomized trial. *Front Psychiatry*. (2017) 8:32. doi: 10.3389/fpsyt.2017.00032

131. McDermott R, Dozois DJ. A randomized controlled trial of internet-delivered CBT and attention bias modification for early intervention of depression. *J Exp Psychopathol*. (2019) 10(2):2043808719842502. doi: 10.1177/2043808719842502

132. Gilbody S, Littlewood E, Hewitt C, Brierley G, Tharmanathan P, Araya R, et al. Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. *Br Med J*. (2015) 351:i195. doi: 10.1136/bmj.h5627

133. Löbner M, Pabst A, Stein J, Dorow M, Matschinger H, Luppa M, et al. Computerized cognitive behavior therapy for patients with mild to moderately severe depression in primary care: a pragmatic cluster randomized controlled trial (@ktiv). *J Affect Disord*. (2018) 238:317–26. doi: 10.1016/j.jad.2018.06.008

134. Birney AJ, Gunn R, Russell JK, Ary DV. Moodhacker mobile web app with email for adults to self-manage mild-to-moderate depression: randomized controlled trial. *JMIR Mhealth Uhealth*. (2016) 4(1):e8. doi: 10.2196/mhealth.4231

135. Rohani DA, Tuxen N, Lopategui AQ, Faurholt-Jepsen M, Kessing LV, Bardram JE. *Personalizing mental health: a feasibility study of a mobile behavioral activation tool for depressed patients. Proceedings of the 13th EAI international conference on pervasive computing technologies for healthcare (PervasiveHealth'19)*. p. 282–91. doi: 10.1145/3329189.3329214

136. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR Mhealth Uhealth*. (2016) 4 (3):e111. doi: 10.2196/mhealth.5960

137. Rohani DA, Quemada Lopategui A, Tuxen N, Faurholt-Jepsen M, Kessing LV, Bardram JE. *MUBS: a personalized recommender system for behavioral activation in mental health. Proceedings of the 2020 CHI conference on human factors in computing systems (CHI '20)*. p. 1–13. doi: 10.1145/3313831.3376879

138. Young CL, Mohebbi M, Staudacher HM, Kay-Lambkin F, Berk M, Jacka FN, et al. Optimizing engagement in an online dietary intervention for depression (my food & mood version 3.0): cohort study. *JMIR Ment Health*. (2021) 8(3):e24871. doi: 10.2196/24871

139. Hirsch A, Luellen J, Holder JM, Steinberg G, Dubiel T, Blazejowskyj A, et al. Managing depressive symptoms in the workplace using a web-based self-care tool: a pilot randomized controlled trial. *JMIR Res Protoc*. (2017) 6(4):e51. doi: 10.2196/resprot.7203

140. Stawarz K, Preist C, Tallon D, Thomas L, Turner K, Wiles N, et al.. *Integrating the digital and the traditional to deliver therapy for depression: lessons from a pragmatic study*. In *Proceedings of the 2020 CHI conference on human factors in computing systems (CHI '20)*. pp. 1–14. doi: 10.1145/3313831.3376510

141. Schlosser DA, Campellone TR, Truong B, Anguera JA, Vergani S, Vinogradov S, et al. The feasibility, acceptability, and outcomes of PRIME-D: a novel mobile intervention treatment for depression. *Depress Anxiety*. (2017) 34(6):546–54. doi: 10.1002/da.22624

142. Mehrotra S, Sudhir P, Rao G, Thirthalli J, Srikanth TK. Development and pilot testing of an internet-based self-help intervention for depression for Indian users. *Behav Sci*. (2018) 8(4):36. doi: 10.3390/bs8040036

143. Sakata M, Toyomoto R, Yoshida K, Luo Y, Nakagami Y, Uwatoko T, et al. Components of smartphone cognitive-behavioural therapy for subthreshold depression among 1093 university students: a factorial trial. *Evid Based Ment Health*. (2022) 25(e1):e18–e25. doi: 10.1136/ebmental-2022-300455

144. Jelinek L, Arlt S, Moritz S, Schröder J, Westermann S, Cludius B. Brief web-based intervention for depression: randomized controlled trial on behavioral activation. *J Med Internet Res*. (2020) 22(3):e15312. doi: 10.2196/15312

145. Haller N, Lorenz S, Pfirrmann D, Koch C, Lieb K, Dettweiler U, et al. Individualized web-based exercise for the treatment of depression: randomized controlled trial. *JMIR Ment Health*. (2018) 5(4):e10698. doi: 10.2196/10698

146. Zagorscak P, Heinrich M, Sommer D, Wagner B, Knaevelsrud C. Benefits of individualized feedback in internet-based interventions for depression: a

randomized controlled trial. *Psychother Psychosom*. (2018) 87(1):32–45. doi: 10.1159/000481515

147. Mayer G, Hummel S, Oetjen N, Gronewold N, Bubolz S, Blankenhagel K, et al. User experience and acceptance of patients and healthy adults testing a personalized self-management app for depression: a non-randomized mixed-methods feasibility study. *Digit Health*. (2022) 8:20552076221091353. doi: 10.1177/20552076221091353

148. Montero-Marín J, Araya R, Pérez-Yus MC, Mayoral F, Gili M, Botella C, et al. An internet-based intervention for depression in primary care in Spain: a randomized controlled trial. *J Med Internet Res*. (2016) 18(8):e5695. doi: 10.2196/jmir.5695

149. Mira A, Bretón-López J, García-Palacios A, Quero S, Baños RM, Botella C. An internet-based program for depressive symptoms using human and automated support: a randomized controlled trial. *Neuropsychiatr Dis Treat*. (2017) 13:987–1006. doi: 10.2147/NDT.S130994

150. Richards D, Timulak L, O'Brien E, Hayes C, Vigano N, Sharry J, et al. A randomized controlled trial of an internet-delivered treatment: its potential as a low-intensity community intervention for adults with symptoms of depression. *Behav Res Ther*. (2015) 75:20–31. doi: 10.1016/j.brat.2015.10.005

151. Richards D, Murphy T, Viganó N, Timulak L, Doherty G, Sharry J, et al. Acceptability, satisfaction and perceived efficacy of "space from depression" an internet-delivered treatment for depression. *Internet Interv*. (2016) 5:12–22. doi: 10.1016/j.invent.2016.06.007

152. Salamanca-Sanabria A, Richards D, Timulak L, Connell S, Mojica Perilla M, Parra-Villa Y, et al. A culturally adapted cognitive behavioral internet-delivered intervention for depressive symptoms: randomized controlled trial. *JMIR Ment Health*. (2020) 7(1):e13392. doi: 10.2196/13392

153. Salamanca-Sanabria A, Richards D, Timulak L. Adapting an internet-delivered intervention for depression for a Colombian college student population: an illustration of an integrative empirical approach. *Internet Interv*. (2019) 15:76–86. doi: 10.1016/j.invent.2018.11.005

154. Kageyama K, Kato Y, Mesaki T, Uchida H, Takahashi K, Marume R, et al. Effects of video viewing smartphone application intervention involving positive word stimulation in people with subthreshold depression: a pilot randomized controlled trial. *J Affect Disord*. (2021) 282:74–81. doi: 10.1016/j.jad.2020.12.104

155. Takahashi K, Takada K, Hirao K. Feasibility and preliminary efficacy of a smartphone application intervention for subthreshold depression. *Early Interv Psychiatry*. (2019) 13(1):133–6. doi: 10.1111/eip.12540

156. Cuijpers P, Heim E, Abi Ramia J, Burchert S, Carswell K, Cornelisz I, et al. Effects of a WHO-guided digital health intervention for depression in Syrian refugees in Lebanon: a randomized controlled trial. *PLoS Med*. (2022) 19(6):e1004025. doi: 10.1371/journal.pmed.1004025

157. Cuijpers P, Heim E, Ramia JA, Burchert S, Carswell K, Cornelisz I, et al. Guided digital health intervention for depression in Lebanon: randomised trial. *Evid Based Ment Health*. (2022) 25(e1):e34–e40. doi: 10.1136/ebmental-2021-300416

158. Carswell K, Harper-Shehadeh M, Watts S, Van't Hof E, Abi Ramia J, Heim E, et al. Step-by-Step: a new WHO digital mental health intervention for depression. *Mhealth*. (2018) 4:34. doi: 10.21037/mhealth.2018.08.01

159. Heim E, Ramia JA, Hana RA, Burchert S, Carswell K, Cornelisz I, et al. Step-by-step: feasibility randomised controlled trial of a mobile-based intervention for depression among populations affected by adversity in Lebanon. *Internet Interv*. (2021) 24:100380. doi: 10.1016/j.invent.2021.100380

160. Piera-Jiménez J, Etzelmueller A, Kolovos S, Folkvord F, Lupiáñez-Villanueva F. Guided internet-based cognitive behavioral therapy for depression: implementation cost-effectiveness study. *J Med Internet Res*. (2021) 23(5):e27410. doi: 10.2196/27410

161. Roepke AM, Jaffee SR, Riffle OM, McGonigal J, Broome R, Maxwell B. Randomized controlled trial of SuperBetter, a smartphone-based/internet-based self-help tool to reduce depressive symptoms. *Games Health J*. (2015) 4(3):235–46. doi: 10.1089/g4h.2014.0046

162. Hatcher S, Whittaker R, Patton M, Miles WS, Ralph N, Kercher K, et al. Web-based therapy plus support by a coach in depressed patients referred to secondary mental health care: randomized controlled trial. *JMIR Ment Health*. (2018) 5(1):e5. doi: 10.2196/mental.8510

163. Titov N, Dear BF, Staples LG, Terides MD, Karin E, Sheehan J, et al. Disorder-specific versus transdiagnostic and clinician-guided versus self-guided treatment for major depressive disorder and comorbid anxiety disorders: a randomized controlled trial. *J Anxiety Disord*. (2015) 35:88–102. doi: 10.1016/j.janxdis.2015.08.002

164. Rosso I, Killgore WD, Olson EA, Webb CA, Fukunaga R, Auerbach RPM, et al. Internet-based cognitive behavior therapy for major depressive disorder: a randomized controlled trial. *Depress Anxiety*. (2017) 34(3):236–45. doi: 10.1002/da.22590

165. Smith J, Newby JM, Burston N, Murphy MJ, Michael S, Mackenzie A, et al. Help from home for depression: a randomised controlled trial comparing internet-delivered cognitive behaviour therapy with bibliotherapy for depression. *Internet Interv*. (2017) 9:25–37. doi: 10.1016/j.invent.2017.05.001

166. O'moore KA, Newby JM, Andrews G, Hunter DJ, Bennell K, Smith J, et al. Internet cognitive–behavioral therapy for depression in older adults with knee osteoarthritis: a randomized controlled trial. *Arthritis Care Res (Hoboken)*. (2018) 70(1):61–70. doi: 10.1002/acr.23257

167. Mewton L, Andrews G. Cognitive behaviour therapy via the internet for depression: a useful strategy to reduce suicidal ideation. *J Affect Disord*. (2015) 170:78–84. doi: 10.1016/j.jad.2014.08.038

168. Mohr DC, Lattie EG, Tomasino KN, Kwasny MJ, Kaiser SM, Gray EL, et al. A randomized noninferiority trial evaluating remotely-delivered stepped care for depression using internet cognitive behavioral therapy (CBT) and telephone CBT .*Behav Res Ther*. (2019) 123:103485. doi: 10.1016/j.brat.2019.103485

169. Schueller SM, Mohr DC. *Initial field trial of a coach-supported web-based depression treatment. Proceedings of the 9th international conference on pervasive computing technologies for healthcare (PervasiveHealth '15)* (2015). p. 25–8

170. Stiles-Shields C, Montague E, Lattie EG, Schueller SM, Kwasny MJ, Mohr DC. Exploring user learnability and learning performance in an app for depression: usability study. *JMIR Hum Factors*. (2017) 4(3):e7951. doi: 10.2196/humanfactors.7951

171. Schure MB, Lindow JC, Greist JH, Nakonezny PA, Bailey SJ, Bryan WL, et al. Use of a fully automated internet-based cognitive behavior therapy intervention in a community population of adults with depression symptoms: randomized controlled trial. *J Med Internet Res*. (2019) 21(11):e14754. doi: 10.2196/14754

172. Stuart R, Fischer H, Leitzke AS, Becker D, Saheba N, Coleman KJ. The effectiveness of internet-based cognitive behavioral therapy for the treatment of depression in a large real-world primary care practice: a randomized trial. *Perm J*. (2022) 26(3):53–60. doi: 10.7812/TPP/21.183

173. Lu SHX, Assudani HA, Kwek TRR, Ng SWH, Teoh TEL, Tan GCY. A randomised controlled trial of clinician-guided internet-based cognitive behavioural therapy for depressed patients in Singapore. *Front Psychol*. (2021) 12:668384. doi: 10.3389/fpsyg.2021.668384

174. Imamura K, Kawakami N, Tsuno K, Tsuchiya M, Shimada K, Namba K. Effects of web-based stress and depression literacy intervention on improving symptoms and knowledge of depression among workers: a randomized controlled trial. *J Affect Disord*. (2016) 203:30–7. doi: 10.1016/j.jad.2016.05.045

175. Chee W, Kim S, Ji X, Park S, Chee E, Tsai H, et al. The effect of a culturally tailored web-based physical activity promotion program on Asian American midlife women's depressive symptoms. *Asian Pac Isl Nurs J*. (2016) 1(4):162–73. doi: 10.9741/23736658.1044

176. Vieira S, Liang X, Guiomar R, Mechelli A. Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clin Psychol Rev*. (2022) 97:102193. doi: 10.1016/j.cpr.2022.102193

177. Hornstein S, Forman-Hoffman V, Nazander A, Ranta K, Hilbert K. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: a machine learning approach. *Digit Health*. (2021) 7:1–11. doi: 10.1177/20552076211060659

178. Bremer V, Chow PI, Funk B, Thorndike FP, Ritterband LM. Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: machine learning approach. *J Med Internet Res*. (2020) 22(10):e17738. doi: 10.2196/17738

179. Bennemann B, Schwartz B, Giesemann J, Lutz W. Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *Br J Psychiatry*. (2022) 220(4):192–201. doi: 10.1192/bjp.2022.17

180. Forsell E, Jernelöv S, Blom K, Kraepelien M, Svanborg C, Andersson G, et al. Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: a single-blind randomized clinical trial with insomnia patients. *Am J Psychiatry*. (2019) 176(4):315–23. doi: 10.1176/appi.ajp.2018.18060699

181. Alamdari PM, Navimipour NJ, Hosseinzadeh M, Safaei AA, Darwesh AA. Systematic study on the recommender systems in the e-commerce. *IEEE Access*. (2020) 8:115694–716. doi: 10.1109/ACCESS.2020.3002803

182. Liu Z, Zou L, Zou X, Wang C, Zhang B, Tang D, et al. Monolith: real time recommendation system with collisionless embedding table. *arXiv*. (2022):2209.07663. doi: 48550/arXiv.2209.07663

183. Kulkarni PV, Rai S, Kale R. *Recommender system in elearning: a survey. Proceeding of international conference on computational science and applications* (2020). p. 119–26. doi: 10.1007/978-981-15-0790-8_13

184. Ong AA, Gillespie MB. Overview of smartphone applications for sleep analysis. *World J Otorhinolaryngol Head Neck Surg*. (2016) 2(1):45–9. doi: 10.1016/j.wjorl.2016.02.001

185. Bort-Roig J, Gilson ND, Puig-Ribera A, Contreras RS, Trost SG. Measuring and influencing physical activity with smartphone technology: a systematic review. *Sports Med*. (2014) 44(5):671–86. doi: 10.1007/s40279-014-0142-5

186. Boukhechba M, Daros AR, Fua K, Chow PI, Teachman BA, Barnes LE. Demonicsalmon: monitoring mental health and social interactions of college

students using smartphones. *Smart Health*. (2018) 9-10:192–203. doi: 10.1016/j.smhl.2018.07.005

187. North F, Chaudhry R. Apple healthkit and health app: patient uptake and barriers in primary care. *Telemed J E Health*. (2016) 22(7):608–13. doi: 10.1089/tmj.2015.0106

188. Carr S. "AI gone mental": engagement and ethics in data-driven technology for mental health. *J Ment Health*. (2020) 29(2):125–30. doi: 10.1080/09638237.2020.1714011

189. Fried EI, Nesse RM. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med*. (2015) 13:1–11. doi: 10.1186/s12916-015-0325-4

190. Boschloo L, Bekhuis E, Weitz ES, Reijnders M, DeRubeis RJ, Dimidjian S, et al. The symptom-specific efficacy of antidepressant medication vs. Cognitive behavioral therapy in the treatment of depression: results from an individual patient data meta-analysis. *World Psychiatry*. (2019) 18(2):183–91. doi: 10.1002/wps.20630

191. Delgadillo J, Duhne PGS. Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach. *J Consult Clin Psychol*. (2020) 88(1):14. doi: 10.1037/ccp0000476

192. Gunlicks-Stoessel M, Klimes-Dougan B, VanZomeren A, Ma S. Developing a data-driven algorithm for guiding selection between cognitive behavioral therapy, fluoxetine, and combination treatment for adolescent depression. *Transl Psychiatry*. (2020) 10:321. doi: 10.1038/s41398-020-01005-y

193. Lamers F, van Oppen P, Comijs HC, Smit JH, Spinhoven P, van Balkom AJ, et al. Comorbidity patterns of anxiety and depressive disorders in a large cohort study: the Netherlands study of depression and anxiety (NESDA). *J Clin Psychiatry*. (2011) 72(3):3397. doi: 10.4088/JCP.10m06176blu

194. Bastiaansen JA, Ornée DA, Meurs M, Oldehinkel AJ. An evaluation of the efficacy of two add-on ecological momentary intervention modules for depression in a pragmatic randomized controlled trial (ZELF-i). *Psychol Med*. (2022) 52 (13):2731–40. doi: 10.1017/S0033291720004845