

# **Promises and Pitfalls of Machine Learning Modeling in Psychological Research**

Kumulative Dissertation zur Erlangung des akademischen Grades

Doktor der Philosophie (Dr. phil.)

Vorgelegt im Fachbereich 01 Humanwissenschaften der Universität Kassel

Von Kristin Jankowsky, M.Sc.

Eingereicht im Januar 2024

Tag der Disputation: 25.04.2024



Gutachter:

Prof. Dr. Ulrich Schroeders (Universität Kassel)

Prof. Dr. Johannes Zimmermann (Universität Kassel)

PD Dr. Timo Gnambs (Leibniz-Institut für Bildungsverläufe)



## Acknowledgements

Writing a dissertation has the potential to be a lonely endeavor, so I would like to thank the people who made sure that this was not the case for me.

First, I would like to thank my supervisor Ulrich Schroeders for the great company and guidance on the way down the machine learning rabbit hole. I learned a lot, always felt encouraged and supported, and it never got boring.

Second, I want to thank Johannes Zimmermann for the valuable feedback on manuscript drafts and research ideas and the willingness to review this thesis. For this, I also thank Timo Gnambs.

Additionally, I am grateful for the wonderful group of colleagues and friends I met at the office over the years. I especially want to thank Gabriel Olaru for being a partner in crime in fun side projects and Priscilla Achaa-Amankwaa, Geraldine Jung, Steffen Müller, Florian Scharf, Kim Speck, Diana Steger, and Leon Wendt for making the office a great place to be. I also want to thank all permanent or sporadic members of our journal club for the opportunity to discuss the current literature and research ideas with a group of highly motivated people.

Further, I would like to thank my family, starting with my sister Simone Wiedemann (for being the best big sister one could have) and my brother-in-law Sebastian Wiedemann (for the introduction to parallel programming that accelerated my analyses by years). Finally, I want to thank my parents Agnes and Günter Jankowsky who have always been my biggest supporters in everything I set my mind to.

## **Abstract**

Machine learning algorithms are becoming increasingly popular across psychology and its subdisciplines. They are often praised for their ability to efficiently deal with collinearity of predictors and complex relationships between predictors and outcomes. Despite their advantages, there are also critical voices pointing out the current limitations of machine learning predictions and questioning whether the algorithms live up to the expectations. In particular, there are increasing reports of incorrect model validation contributing to inflationary results. In this thesis, I investigate the usefulness of machine learning across four prediction use cases, namely the prediction of attrition in longitudinal studies (manuscript 1), of suicide attempts of adolescents (manuscript 2), of treatment response in psychotherapy (manuscript 3), and of psychotherapy dropout (manuscript 4). In each of these research areas, machine learning algorithms are increasingly being used with the aim to proactively prevent negative outcomes. In the prologue and in the four studies, I present and address typical pitfalls in machine learning modeling common for each of these areas. In the epilogue, I discuss different aspects (time, settings, cultures, measures, and methods) that can affect the generalizability of predictive models and that have not been sufficiently considered in the psychological research literature so far. In addition, I address several aspects where there is still room for improvement not only in machine learning modeling in psychological research in general, but also with respect to the prediction models included in this thesis (e.g., a more stringent feature selection or a more rigorous combination of machine learning modeling and open science practices).

## **Zusammenfassung**

Algorithmen des maschinellen Lernens werden in der Psychologie und ihren Teildisziplinen immer beliebter. Sie werden häufig dafür angepriesen, dass sie mit der Multikollinearität von Prädiktorvariablen und komplexen Beziehungen zwischen Prädiktoren und Kriterien effizient umgehen können. Trotz ihrer Vorteile werden auch immer mehr kritische Stimmen laut, die auf die derzeitigen Grenzen von Vorhersagen durch maschinelles Lernen hinweisen und die Frage aufwerfen, ob die Algorithmen den Erwartungen gerecht werden. Insbesondere gibt es immer mehr Berichte über fehlerhafte Modellvalidierungen, die wiederum inflationäre Ergebnisse bedingen. In dieser Dissertation untersuche ich die Nützlichkeit von Algorithmen des maschinellen Lernens für die Vorhersage von Abbrüchen in längsschnittlichen Studien (Manuskript 1), Suizidversuchen von Jugendlichen (Manuskript 2), dem Ansprechen auf eine Psychotherapie (Manuskript 3) und dem Abbruch einer Psychotherapie (Manuskript 4). In jedem dieser Gebiete werden Machine Learning Algorithmen immer häufiger mit dem Ziel genutzt, negative Ergebnisse proaktiv zu vermindern. Innerhalb des Prologs gehe ich jeweils auf typische Probleme bei der Machine Learning Modellierung innerhalb dieser Gebiete ein und adressiere diese innerhalb der vier Studien. Im Epilog diskutiere ich verschiedene Aspekte (Zeit, Umgebung, Kultur, Messinstrumente und -verfahren, und Methoden), die sich auf die Generalisierbarkeit von Vorhersagemodellen auswirken können und bisher bei dessen Betrachtung innerhalb der psychologischen Forschung zu kurz kommen. Darüber hinaus thematisiere ich verschiedene Aspekte, bei denen es innerhalb von Machine Learning Modellierung in der psychologischen Forschung allgemein, aber auch für meine Vorhersagemodelle noch Verbesserungspotential gibt (z.B. eine strengere Variablenvorabauswahl oder die stärkere Verknüpfung von Machine Learning Modellierung und Open Science Praktiken).

## Table of Contents

<b>Prologue</b>	<b>1</b>
Explanation vs. Prediction: A Tale as Old as Time?	3
Model Training and (Cross-)Validation	4
Model Validation	7
Predicting Attrition in Longitudinal Studies	9
Predicting Suicidal Behavior	13
Predicting Psychotherapy Dropout and Treatment Response	16
References	20
<b>Validation and generalizability of machine learning prediction models on attrition in longitudinal studies</b>	<b>30</b>
<b>Predicting lifetime suicide attempts in a community sample of adolescents using machine learning algorithms</b>	<b>39</b>
<b>First impressions count: Therapists' impression on patients' motivation and helping alliance predicts psychotherapy dropout</b>	<b>57</b>
<b>Predicting treatment response using Machine Learning: A registered report</b>	<b>84</b>
<b>Epilogue</b>	<b>104</b>
Manuscript 1: Validation and generalizability of machine learning prediction models on attrition in longitudinal studies.	105
Manuscript 2: Predicting lifetime suicide attempts in a community sample of adolescents using machine learning algorithms.	106
Manuscript 3: First impressions count: Therapists' impression on patients' motivation and helping alliance predicts psychotherapy dropout.	108
Manuscript 4: Predicting treatment response using machine learning: A Registered Report.	109
There is no Such Thing as a Validated Prediction Model	110



Time	111
Settings	111
Cultures	112
Measurement	113
Methods	114
Testing for Generalizability Across Cultures and Methods: An Empirical Example	114
Asking Better and Fewer Questions...	117
... at the Right Time?	120
Scientific Utopia for Machine Learning in Psychological Research	122
References	127

# Prologue

A decade ago, Kosinski et al. (2013) showed that it was possible to predict a person's demographics, personality, intelligence, and happiness based on their somewhat inattentively generated digital traces (i.e., Facebook Likes). In a follow-up study, Youyou et al. (2015) demonstrated that machine learning (ML) models using Facebook Likes were even more accurate at judging a target's personality than their friends or colleagues. Despite potential privacy issues, these and similar findings were notable in that they reactivated the debate about statistical modeling cultures (Breiman, 2001; Shmueli, 2010; Yarkoni & Westfall, 2017) and likely spurred the rise of ML modeling in psychological science. Currently, ML-based predictive models are gaining popularity across psychology and its subdisciplines: For example, they have been used to improve the assessment and the prediction of personality traits (Stachl et al., 2020), to exploit the nuances of questionnaire items to improve the prediction of life outcomes (Möttus et al., 2017; Seeboth & Möttus, 2018; Stewart et al., 2022), for personalized models of psychotherapy outcomes (Lutz et al., 2018; Schwartz et al., 2020) or to flag who is at high risk of dropping out of university (Behr et al., 2020).

ML algorithms are often praised for their ability to efficiently deal with large numbers of heterogeneous predictors and complex relationships (i.e., non-linear, and interactive) between predictors and outcomes (e.g., Zou & Hastie, 2005) without having to specify them a priori. In psychological research, the underlying belief is that ML algorithms can help to prepare and process new types of data for psychologically relevant research questions (e.g., Adjerid & Kelley, 2018), but also enable novel insights and improved predictive performances in re-analyses of existing data. So far, the findings on whether the algorithms are living up to these expectations are mixed: There is anecdotal evidence that ML algorithms are superior to more traditional modeling approaches in some cases (e.g., Ali & Ang, 2022), but these positive reports are accompanied by just as many criticisms of questionable or even flawed implementation of the methods (Kapoor & Narayanan, 2023).

The following prologue to this thesis is divided into three parts: First, I will discuss the (philosophical) differences between explanatory and predictive analyses in psychological research. Second, I will give an overview of the typical ML modeling process and some of the most common algorithms used in the literature I discuss and in the four manuscripts included in this thesis. Third, I will present the four manuscripts of this thesis, which span three different prediction use cases, a) the prediction of longitudinal attrition, b) the prediction of suicidal behavior, and c) the prediction of therapy outcomes. In each of these areas, predictive modeling has been used to proactively classify individuals at risk for undesirable outcomes, with the ultimate goal of prevention. By addressing field-specific pitfalls of ML modeling, I aim to further explore the value of ML in psychological research using rigorous model validation approaches in a computationally reproducible and accessible manner.

### **Explanation vs. Prediction: A Tale as Old as Time?**

Although the topic of explanation vs. prediction in the behavioral sciences seems to have been reactivated by the increasing use and power of complex ML models, the underlying debate has been going on for much longer. Shmueli (2010) dates the “conflation of explanation and prediction” (p. 292) in the philosophy of science back to the 1940s. While Breiman (2001) distinguishes between data models (which assume that data are generated by some stochastic model) and algorithmic models (which treat the data generation process as a “black box” and focus on predictive accuracy) and clearly argues for the latter to solve relevant application problems, Shmueli (2010) takes on a more diplomatic role: The author emphasizes the need for both explanatory and predictive modeling, criticizes a lack of understanding of the differences between the two, and outlines the extent to which predictive models can even be useful for theory building and testing. In contrast to Breiman (2001), Shmueli (2010) does not rely on any black box analogy, but rather defines explanatory modeling as the use of statistical models to test causal hypotheses about theoretical constructs

and predictive modeling as the use of statistical models to predict new or future observations. Somewhat ironically, a current research topic within the ML community is *causal machine learning*, which, as the name suggests, attempts to combine predictive modeling with causal inference (e.g., Kaddour et al., 2022; Knaus et al., 2021). Which statistical culture prevailed has long been linked to the discipline in question (e.g., in the behavioral sciences, the focus has been on explaining behavior and testing theoretical models, whereas bioinformaticians, for example, have been more interested in prediction), but it seems that ML methods are increasingly blurring these boundaries.

More recently, Yarkoni and Westfall (2017) point out in an influential paper that psychologists ultimately have to choose between developing simplified models that may be theoretically elegant but cannot adequately predict human behavior, and complex models that can predict human behavior but are not easily understood or communicated. They even break down the replication crisis in psychology (e.g., Open Science Collaboration, 2015) to overfit of explanatory models, that is, models that were supposedly able to explain a phenomenon failed to perform similarly when tested on new samples. In contrast, in predictive modeling, the performance of a model should be measured by its ability to predict an outcome in new data, not by its performance in the sample on which it was trained, which is one of the key conceptual differences between the two approaches. In terms of the bias-variance-tradeoff in modeling, explanatory approaches clearly seek to minimize bias (Yarkoni & Westfall, 2017), whereas in predictive modeling using ML, the goal is to minimize prediction error by finding an optimal balance between introducing bias (using regularization techniques) and reducing explained variance (in model training) to some extent.

### **Model Training and (Cross-)Validation**

Depending on the specific ML algorithm, different so-called hyperparameters can be tuned during model training in order to maximize the predictive performance. In the

following, I will give a brief overview of the three ML algorithms that I used in the analyses included in this thesis and that are also widely used in the literature on predictive modeling in psychological research: elastic net regression, random forests, and gradient boosting machines.

Elastic net regression is a regression variant that incorporates an additional penalty term, balancing between ridge and least absolute shrinkage and selection operator (LASSO) regression (Zou & Hastie, 2005). In ridge regression, the regression coefficients are constrained by adding the sum of the squared weights weighted by the tunable shrinkage factor  $\lambda$ . In doing so, coefficients of less important predictors are shrunk toward (but not set to) 0, and coefficients of correlated predictors are shrunk toward each other. In LASSO regressions, the sum of the absolute weights is used for the penalty term (as opposed to squared weights), and features can be eliminated from the model by settings coefficients to exactly 0, resulting in potentially more parsimonious models. By using a tunable penalty parameter  $\alpha$  (ranging from 0 to 1), elastic net regression compromises between the two approaches in a data-driven manner to improve predictive performance. Note that elastic net regressions are often (and also in this thesis) used as a regularized point of comparison to unregularized linear regressions, but it would also be possible to additionally include any other form of association one is interested in (i.e., nonlinear or interaction effects) in the model formula, potentially obviating the need for more complex models if it would be doable to specify any relevant effects a priori.

Random forests is an ensemble algorithm that constructs a multitude of decision trees during model training (e.g., Probst et al., 2019). Like individual decision trees, it allows nonlinear and interaction effects to be incorporated into modeling without having to specify them in advance. The "random" refers to two aspects: the random selection of data points for each tree and the random subset of features considered when splitting nodes. Each tree in the

forest independently predicts the outcome, and the final prediction is often the average (for regression tasks) or the majority vote (for classification tasks) of all the trees. Tunable hyperparameters include the number of trees, the number of variables (often called *mtry*) and sample size to be considered at each split, the minimum node size, and the splitting rule. The optimal settings for these parameters depend on the respective data, but the number of trees should be set sufficiently high (a reasonable strategy would be to test different numbers and use the one where the predictive performance does not increase with an increasing number of trees, since more trees also increase the computational complexity of the analyses).

Gradient boosting machines (GBM) is another tree-based ensemble technique that also incorporates nonlinear and interaction effects into the modeling, but instead of constructing multiple trees independently, it constructs trees sequentially, with each subsequent tree aiming to fit the residual errors of the previous ones. It minimizes a loss function using a technique called boosting, in which weak learners (usually shallow trees) are combined to create a strong learner (James et al., 2017). Unlike in random forests, the entire data set is considered when splitting nodes. Due to the sequential approach, models tend to become complex which can be mitigated (in order to avoid overfitting) by judicious tuning of hyperparameters such as the number of trees, learning rate, or minimum leaf size (McNamara et al., 2022).

According to Hong et al. (2020), comparisons between increasingly complex models can be used to determine whether nonlinear or interaction effects are present in the data. In the analyses included in this thesis, I compare at least unregularized regressions, elastic net regression and an additional tree-based algorithm to be, in principle, able to make such statements about the incremental value of allowing for more complex associations. However, this approach is only valid if the data are suitable for it, that is, if the sample size is sufficiently large and the indicators are reliably measured so that complex patterns can be detected at all (Jacobucci & Grimm, 2020).

## ***Model Validation***

In predictive modeling, there are different approaches to model development and validation, ranging in its rigor from developing a model using all available data and evaluating its performance on the same data to validating the model using completely unseen data from an independent study (Collins et al., 2024; Dwyer et al., 2018). Often, the latter is not feasible due to limited resources or lack of truly independent samples, so researchers resort to workarounds such as splitting the sample into training and testing data, using k-fold-cross-validation or a combination of both. In order to strictly separate the hyperparameter tuning from the model validation, it is recommended to use some form of nested resampling approach (Bischl et al., 2012; Pargent et al., 2023). Figure 1 illustrates a simplified version<sup>1</sup> of such a nested resampling process: In general, the full dataset is split into a training (e.g., 80%) and a testing data set (the remaining 20%) in each iteration of the outer validation loop. Any data preprocessing steps (such as imputation of missing values) are performed separately on the training and testing data to avoid biasing the model validation by introducing information leakage. In the inner validation loop, the model is trained using k-fold-cross-validation (5-fold in Figure 1).

The tuned hyperparameters taken from the inner loop model training are then used to estimate the predictive performance in the unseen testing data of the outer validation loop. Finally, the outer loop test performances are averaged across all iterations to obtain a realistic estimate of the model's predictive performance in unseen data (assuming that these unseen data are from the same population). Nested resampling has two advantages over other validation schemes such as using the averaged testing performances from k-fold-cross-validation: First, hyperparameter tuning and model validation are strictly separated and

---

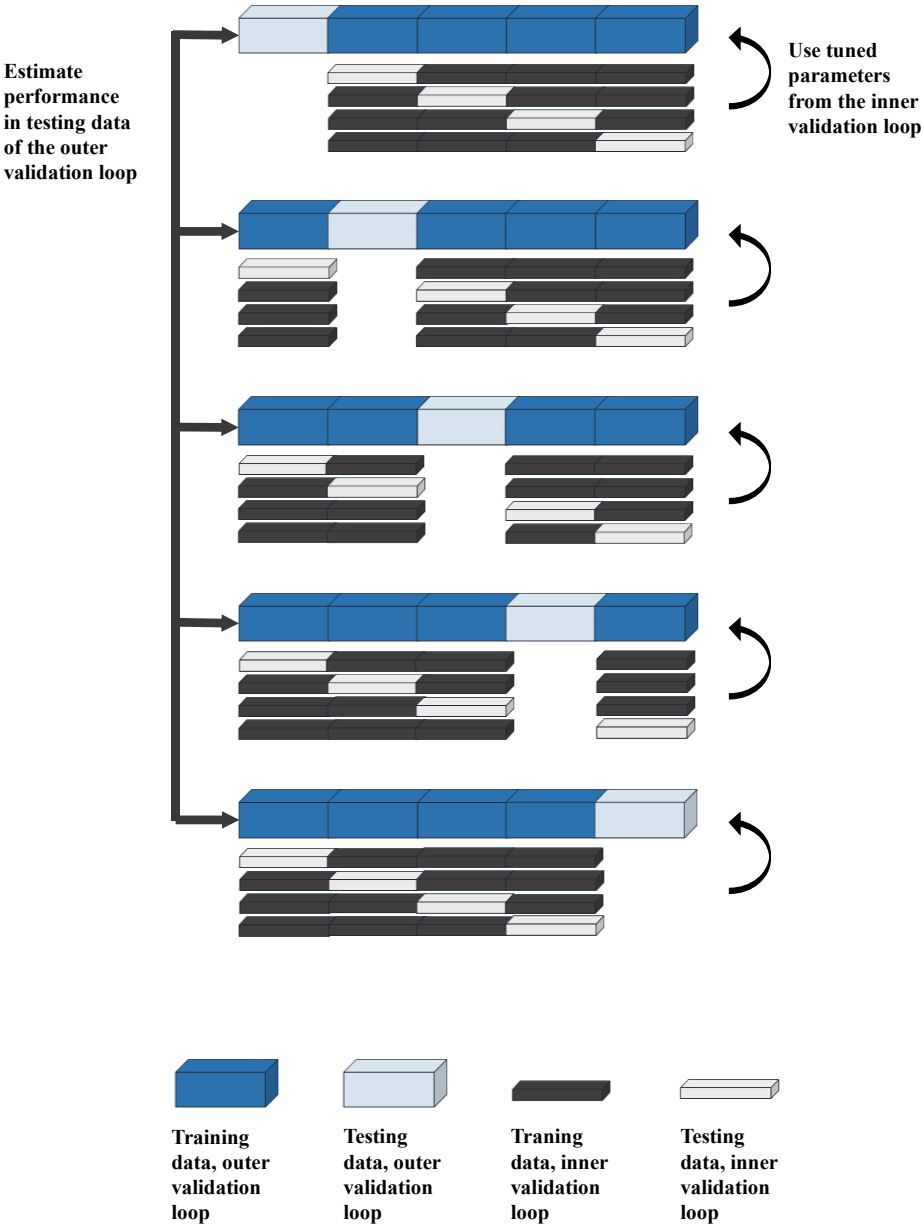
<sup>1</sup> For example, in our analyses, we used 10-fold-cross-validation in the inner validation loop and at least 100 up to 1,000 iterations of the outer validation loop.



second, it is possible to depict the variation of predictive performance due to person sampling in the outer loop. However, in my reading of the literature on ML modeling in psychological research, with a few exceptions, model validation is often done in a less extensive or thoughtful manner. In the following section, which introduces the four manuscripts included in this thesis, I will elaborate on this point for each research area.

**Figure 1**

*Schematic Illustration of a Nested Resampling Process Including 5 Folds in the Inner Validation Loop and 5 Iterations of the Outer Validation Loop*



## **Predicting Attrition in Longitudinal Studies**

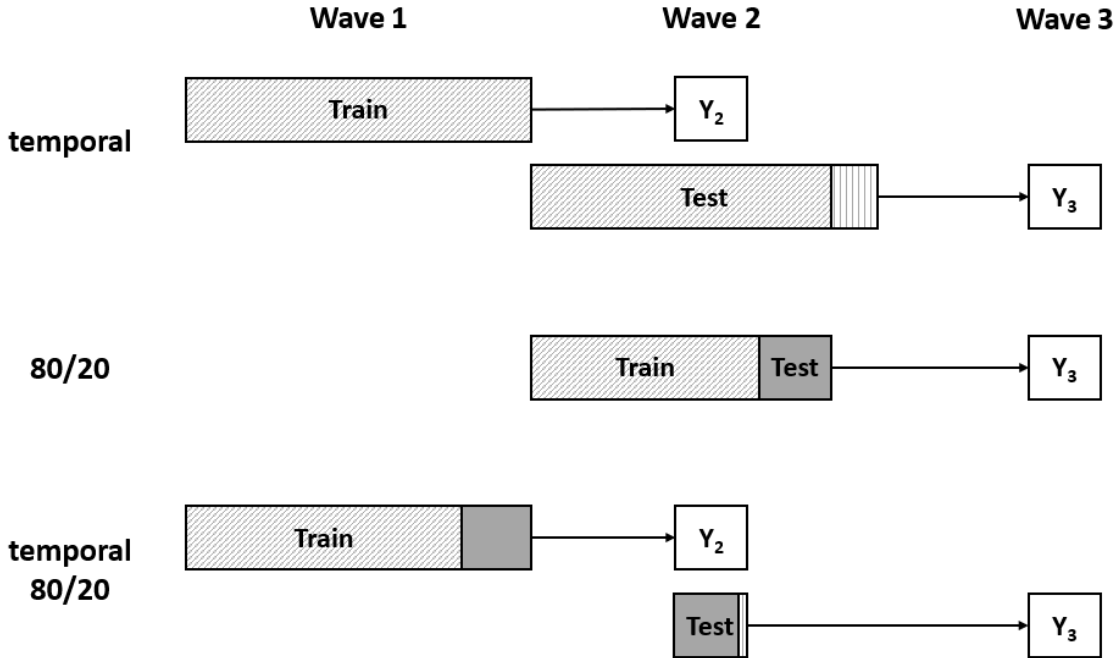
Attrition, especially if it is systematic and not properly accounted for in the analysis of data, poses a serious threat to the validity of findings from longitudinal studies (Little & Rubin, 2002). Common approaches to dealing with attrition in longitudinal studies require either strong assumptions about the underlying attrition-generating process (e.g., multiple imputation; Schafer & Graham, 2002) or additional resources (e.g., refreshment sampling; Deng et al., 2013). Since the best solution to the problem of missing data “is not to have missing data” (e.g., Little, 2021, p.105), it would be ideal to know in advance which participants are at risk of dropping out of the study in order to be able to implement retention measures. Therefore, study attrition represents a fitting use case for algorithms that focus on prediction rather than explanation of phenomena, since any model of attrition would only be useful in real-world applications if it could improve predictive performance with respect to future behavior in subsequent measurement waves.

There have been a few studies that have used ML algorithms to predict longitudinal attrition with promising results (e.g., Jacobsen et al., 2021; Kern et al., 2021; Zinn & Gnamb, 2020). In particular, the superior predictive performance of complex models over simple regression analyses suggests that ML could add real value to survey research, prompting me to take a closer look at model training and validation in these studies. For example, Kern et al. (2021) used different sets of predictors with various ML algorithms (i.e., penalized logistic regression, decision trees, random forest, extremely randomized trees, and extreme gradient boosting) to predict longitudinal attrition within the GESIS panel (Bosnjak et al., 2018). To validate their prediction models, the authors performed temporal cross-validation across the survey waves (see the top row in Figure 2) as follows: A prediction model was built using data from all active participants at survey wave 1 to predict the participation status at wave 2. The resulting model was then tested using all active wave 2 participants to predict

participation status at wave 3. Participation status was a dichotomous variable distinguishing active participants from temporary or permanent dropouts. Using baseline variables (including sociodemographics and variables about survey cooperation), the highest predictive accuracy with an average Area Under the Curve (AUC) of .759 across all 18 survey waves (from 2014 to 2017) was reported for the random forest algorithms. In comparison, using logistic regressions yielded an average AUC of .645.

**Figure 2**

*Different Longitudinal Cross-Validation Approaches*



*Note.* Excluded participants at Wave 2 are represented by the vertically dashed sections of the rectangle.

On the face of it, Kern et al.’s argument for trying to predict the unknown participation status of a future wave is perfectly reasonable, but the potential problems with this model validation lie in the details: Due to the temporal cross-validation scheme used by the authors, most participants and their values of all baseline predictor variables in the test data are the same as in the training data for all waves. The only differences between the training and

testing data are a slightly smaller sample size in the testing data (because all permanently inactive participants were excluded across time / survey waves) and a change in the outcome for a minority of participants (e.g., 11% between wave 3 and 4 in 2015). Thus, it is reasonable to assume that the superior performance of the random forests was inflated in the same way as in studies using upsampling before splitting the data into testing and validation data (e.g., Vandewiele et al., 2021).

To examine the effects of three different validation strategies for predicting longitudinal attrition in more detail, I reanalyzed data from the six 2015 GESIS waves. The first strategy I used was the temporal cross-validation performed on the full datasets also used by Kern et al. (2021). A second strategy was to discard the temporal validation approach and split the data from each wave into a training data (80 %) and testing data (20 %; see the second row in Figure 2), thus testing the model trained on wave 2 data to predict participation status at wave 3. A third strategy combined temporal validation and splitting data into disjoint training and testing data (see the bottom row in Figure 2). For this, the model was trained on 80 % of the data from survey wave 1 to predict status at wave 2, and tested on the active participants from the remaining 20 % of wave 2 to predict participation status at wave 3. I used the same baseline variables from the recruitment interview as Kern et al. (2021) and compared elastic net regressions and random forest algorithms across these three validation approaches.

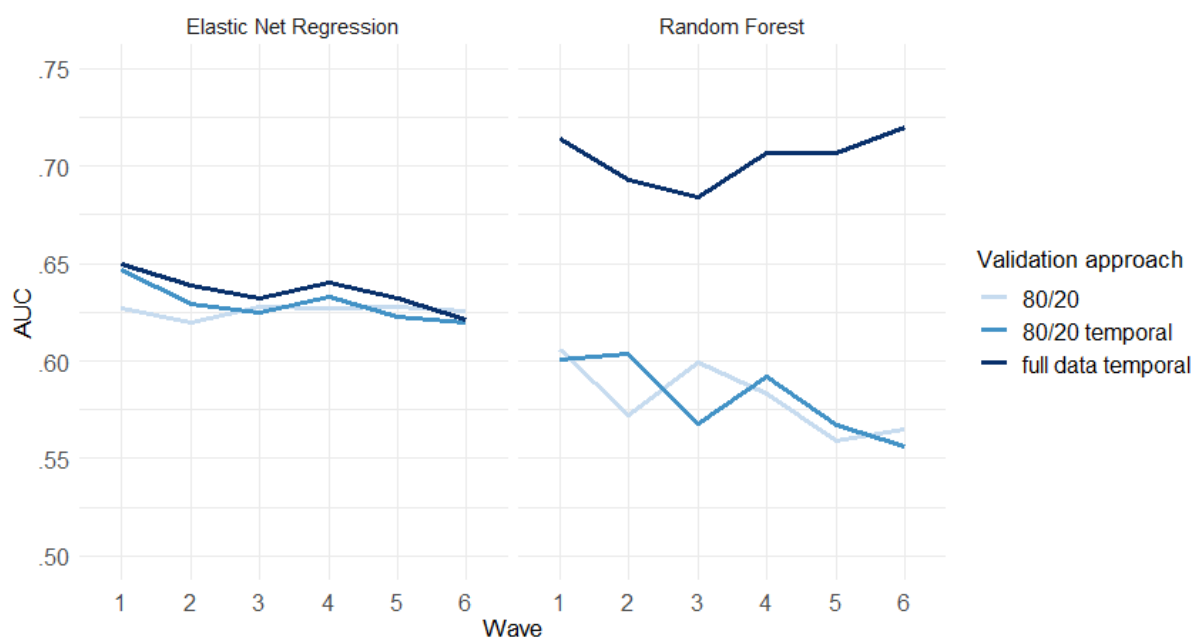
Figure 3 shows two main findings: The reproduced temporal method leads to a large overestimation of predictive accuracy for tree-based algorithms such as random forests in every wave I analyzed. This spuriously higher accuracy may lead researchers to infer interactions or nonlinear effects where none exist. Second, the upwardly biased AUC estimates are not caused by the temporal cross-validation approach per se—as the AUC are

not systematically different between the 80/20 and the temporal 80/20 approach—but rather by the failure to use independent holdout data for model testing.

In summary, validation in a longitudinal setting carries the risk of using overlapping samples for the training and the test data, which should be avoided. The intention to develop and empirically evaluate models in a longitudinal setting is reasonable and the literal goal of prediction, as it should yield more realistic/ecologically valid results than cross-sectional model evaluation. To achieve this, temporal cross-validation can be applied, but only as a possible complement to the splitting of the sample into disjoint training and validation data.

**Figure 3**

*AUC for the Different Longitudinal Validation Approaches of Each GESIS Wave of 2015*



In **Manuscript 1: Validation and generalizability of machine learning prediction models on attrition in longitudinal studies**, to further demonstrate the application of the two model validation strategies described above (rows 2 and 3 of Figure 2), we apply them to validate models for predicting longitudinal attrition in two survey panels, namely the Midlife in the United States (MIDUS; Brim et al., 2004) and Panel Analysis of Intimate Relationships

and Family Dynamics (pairfam; Huinink et al., 2011). Using baseline indicators from both panel studies (including demographics, health indicators, and personality) as predictor variables, we compare the accuracy of logistic regressions and GBM. We also discuss common approaches to dealing with panel attrition, where they may fall short, and the ways in which predictive modeling can provide added value. Additionally, we examine the generalizability of our prediction models across both studies, methods, and measurement waves.

### **Predicting Suicidal Behavior**

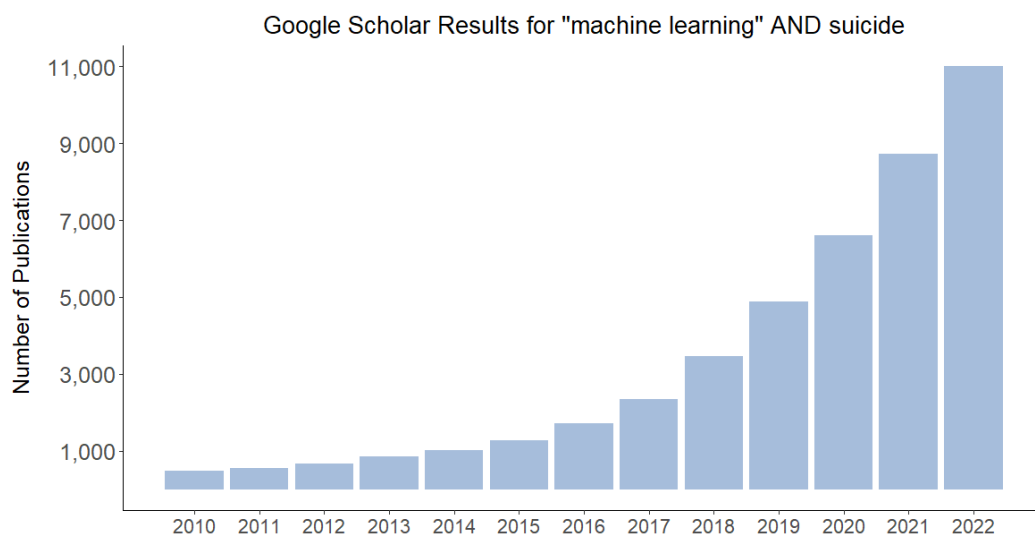
In recent studies on the prediction of suicidal behavior, authors often argue for the use of complex ML algorithms. Many point to the comprehensive meta-analysis of risk factors for suicidal thoughts and behaviors by Franklin et al. (2017) which summarized 365 studies from the last 50 years and found that the individual effects of 16 different predictor categories were small, that the prediction of suicide attempts was only better than chance, and that few studies combined multiple assumed risk factors for their modeling. Franklin et al. (2017) prominently stated that their results “suggest the need for a shift in focus from risk factors to ML-based risk algorithms” (p. 187). Whether or not they were actually responsible, it is fair to say that the authors' call was heard. Figure 4 shows the clearly increasing number of publications related to ML-based modeling and suicide.

Some studies using ML to predict suicidal thoughts and behaviors have found algorithms that allow for interactions and nonlinear effects, such as random forests, to be substantially superior to “traditional” methodological approaches such as logistic regression. For example, Walsh et al. (2018) predicted suicide attempts in adolescents and found a  $\Delta$ AUC of  $> .20$  for each reported comparison between logistic regressions and random forests (using which models classifying suicidal adolescents versus a general hospital control group at 7 days prior to the respective suicide attempts achieved an impressive AUC of .97). Fox et al.

(2019) predicted self-harming behaviors in an adult sample and also reported better predictive performance of random forests (AUC = .87-.90 across different time points) compared to logistic regressions (AUC = .70-.72). Huang et al. (2020) distinguished between suicide ideators and suicide attempters; across five samples, logistic regression models achieved AUC values between .65 and .72 as compared to random forests with AUC values ranging between .87 and .90. It is not surprising that these results were widely recognized, since it seems entirely plausible that suicidal behaviors may be complex, and such high-performing predictive models would represent a breakthrough in the creation of clinically useful risk assessment tools.

#### **Figure 4**

*Number of Publications From 2010 to 2022 Related to the Search Terms “Machine Learning” and Suicide*



A reanalysis and simulation study by Jacobucci et al. (2021) showed that these results were, unfortunately, due to a similar model validation problem as discussed above for predicting attrition, that is, non-distinct training and testing data led to information leakage that fully explained the better performance of the random forests. Specifically, in the

respective studies, the authors used *optimism bootstrap sampling* (Harrell et al., 1996), which involves building a model on the full sample, then building it on several bootstrap samples—evidently containing some observations multiple times—and then testing each of these models again on the full original sample. The averaged difference between these “test” performances and the accuracy obtained with the bootstrap samples is called “optimism”. In a final step, the optimism estimate is subtracted from the accuracy of the first model built using the full sample. While optimism bootstrap sampling leads to significant overestimation of predictive performance in tree-based ML algorithms (e.g., random forests or gradient boosting), simulations show that it has a less dramatic effect on (regularized) linear regression models (e.g., Tantithamthavorn et al., 2017). Thus, the falsely high predictive accuracy of tree-based models could lead to the erroneous assumption that complex interactions or nonlinear relationships exist in the data. Note that this bias is not only present in small samples but was recently demonstrated by Coley et al. (2023) for suicide prediction models incorporating data from more than 13 million health care visits.

Apart from studies using optimism bootstrap sampling, more reliable statements about the incremental value of ML algorithms for predicting suicidal behaviors are generally hard to come by. For example, according to a review of ML-based suicide prediction by Burke et al. (2019), studies that explicitly aimed to improve predictive performance showed an overall higher accuracy than the previous literature using less-complex methods. However, the authors also highlight the overall nontransparent and inconsistent reporting of model training and validation approaches, which—combined with the rarity of studies presenting simpler regression analyses as performance baselines—complicates comparisons between algorithms and across studies. In addition, it has been found that the majority of ML-based predictive models use adult clinical samples (e.g., Bernert et al., 2020). On the one hand, this is to be expected, as the average suicide risk in these samples is expected to be higher than in



community samples, rendering them particularly relevant for prevention efforts. On the other hand, from an early prevention perspective (as past suicide ideation and suicidal behavior are the strongest predictors of subsequent suicidal behavior; Beckmann et al., 2018; Geulayov et al., 2019), it would be worthwhile to examine the utility of ML algorithms more closely for predicting suicidal behaviors in younger community samples.

**In Manuscript 2: Predicting lifetime suicide attempts in a community sample of adolescents using machine learning algorithms**, we predict self-reported lifetime suicide attempts among 17-year-olds in the Millennium Cohort Study ( $N = 7,347$ ). We review recent studies using ML algorithms for the prediction of suicide ideation and suicidal behaviors in community samples of adolescents and young adults and discuss their strengths and limitations. For the prediction of lifetime suicide attempts, we include a diverse set of self- or other-reported (usually by the adolescent's caregiver) predictor variables (638 in total) across 14 different categories (e.g., attitudes, demographics, drug use, mental health, offenses, or victimization). By comparing the predictive accuracy of two ML algorithms (elastic net regressions and GBM) to logistic regressions, we investigate to what extent it is possible to predict self-reported lifetime suicide attempts in a community sample of adolescents and whether ML algorithms can be used to improve predictive performance. We also compare the predictions of models using either variables measured 3 years prior to the measurement wave in which the suicide attempts were reported or using variables taken from the same measurement wave. In doing so, we aim to assess the possible effect of different time intervals between the measurement of predictors and outcome on model performance and the effect of different adolescent developmental stages on variable importances.

### **Predicting Psychotherapy Dropout and Treatment Response**

Predictive modeling via ML algorithms has been increasingly used in psychotherapy research, for example, to personalize treatment (Cohen et al., 2020; Gómez Penedo et al.,

2022; Schwartz et al., 2020), to predict treatment response (Hilbert et al., 2021; Webb et al., 2020) or therapy dropout (e.g., Bennemann et al., 2022), and to build feedback systems based on these predictions (e.g., Lutz et al., 2019). The overall goal of these personalized predictions (often summarized under the interdisciplinary umbrella term precision medicine) is to be able to prevent treatment failures and to provide valuable feedback on the therapeutic process to patients and therapists alike. Predicting therapy dropout represents a classification task similar to predicting longitudinal attrition (and logically includes the same ultimate goal of reducing the number of events, i.e., dropouts), however, depending on the setting, there are additional challenges. First, in clinical contexts, researchers may be more prone to mislabeling the outcome or creating a group of “dropouts” that is too heterogeneous. In longitudinal attrition, the case is fairly straightforward, a participant either provides responses or is absent at a given time. In psychotherapies, termination of therapies may be initiated by the therapist, the patient, or both, and dropout is usually (but not always) defined as the termination of a therapy by the patient against the therapist’s advice. There is also the question of whether a person who drops out after one week can be grouped with another person who almost successfully completes the outpatient therapy or a stay at a clinic for the initially planned length of time. Therefore, when reporting results and comparing findings across studies, it is important to take into account exactly how the outcome was labeled in each study. Second, studies that use baseline indicators to predict therapy outcomes often have small samples and a relatively large number of predictor variables (Chekroud et al., 2021) which—in combination with a rare event being predicted—can lead to overfitting and nongeneralizable models. While this is an issue that has received some attention in the medical community (e.g., Van Calster et al., 2020), the magnitude of the problem has not yet been fully recognized within psychological research (e.g., Giesemann et al., 2023).

**In Manuscript 3: First impressions count: Therapists' impression on patients' motivation and helping alliance predicts psychotherapy dropout**, we predict therapy dropout in two German inpatient psychotherapy clinics ( $N = 1,691$  in Sample 1 and  $N = 12,473$  in Sample 2) using baseline indicators (e.g., demographics or variables on previous treatments and symptom severity). We compare the predictive accuracy of linear regressions, elastic net regressions, and GBM. Using data from two different clinics, we also look for similarities and differences in predictive accuracy and variable importances across settings to be able to make statements about model generalizability. In additional analyses, we examine the interrelated and potentially detrimental effects of unequal group sizes, number of events (i.e., dropouts), and number of predictor variables on predictive accuracy in classification tasks and explore the extent to which regularization by ML algorithms can mitigate them. As a sensitivity check, we also compare predictions for patients who dropped out of the therapy within the first week of treatment with predictions for patients who dropped out at a later date.

In addition to criticisms on poorer interpretability of complex ML models (Siddaway et al., 2020) compared to simpler models and a general lack of external model validation (Wilkinson et al., 2020), which dampen the precision medicine hype in psychotherapy research, there are also reviews reporting a lack of transparency and model presentation norms that make it difficult to evaluate ML models in clinical research (Lee et al., 2018).

**In Manuscript 4: Predicting treatment response using machine learning: A Registered Report**, we aim to address these criticisms by conducting one of the first (and to our knowledge, the first in clinical psychology) registered reports using ML algorithms in psychological research. The process of a registered report differs from other studies in that the theoretical background and proposed analyses are reviewed in a first step, and the study including all analyses is conducted only after the submitted manuscript has been accepted in principle. In this way, the exploitation of the researchers' degrees of freedom should be

minimized and the modeling steps should ideally be presented in a transparent and reproducible way. In stage 1, we proposed to compare linear regressions, elastic net regressions, and GBM for predicting treatment response (as operationalized by the Patient Health Questionnaire Anxiety and Depression Scale) in a German inpatient sample ( $N = 723$ ) using a variety of baseline indicators. The first part of the review process led to a fine-tuning of the algorithms' hyperparameter grids and to the creation of four predictor variable groups (demographics, variables on physical and mental health, and treatment-related variables) based on which we will perform sub-analyses (using all possible group combinations) in order to examine the unique and joint contribution of the predictor categories to the predictive performance.

## References

- Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, 73(7), 899–917.  
<https://doi.org/10.1037/amp0000190>
- Ali, F., & Ang, R. P. (2022). Predicting how well adolescents get along with peers and teachers: A machine learning approach. *Journal of Youth and Adolescence*, 51(8), 1241–1256. <https://doi.org/10.1007/s10964-022-01605-5>
- Beckman, K., Mittendorfer-Rutz, E., Waern, M., Larsson, H., Runeson, B., & Dahlin, M. (2018). Method of self-harm in adolescents and young adults and risk of subsequent suicide. *Journal of Child Psychology and Psychiatry*, 59(9), 948-956.  
<https://doi.org/10.1111/jcpp.12883>
- Behr, A., Giese, M., Tegum K, H. D., & Theune, K. (2020). Early prediction of university dropouts – A Random Forest approach. *Jahrbücher für Nationalökonomie und Statistik*, 240, 743-789. <https://doi.org/10.1515/jbnst-2019-0006>
- Bennemann, B., Schwartz, B., Gieseemann, J., & Lutz, W. (2022). Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *The British Journal of Psychiatry*, 1–10. Advance online publication.  
<https://doi.org/10.1192/bjp.2022.17>
- Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnoui, F. (2020). Artificial intelligence and suicide prevention: A systematic review of machine learning investigations. *International Journal of Environmental Research and Public Health*, 17(16), 5929. <https://doi.org/10.3390/ijerph17165929>
- Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation*, 20(2), 249–275. [https://doi.org/10.1162/EVCO\\_a\\_00069](https://doi.org/10.1162/EVCO_a_00069)

- Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review*, 36(1), 103-115. <https://doi.org/10.1177/0894439317697949>
- Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3). <https://doi.org/10.1214/ss/1009213726>
- Brim O. G., Ryff C. D., Kessler R. C. (2004). The MIDUS national survey: An overview. In Brim O. G., Ryff C. D., Kessler R. C. (Eds.), *How healthy are we? A national study of well-being at midlife* (pp. 1–34). University of Chicago Press.
- Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of Affective Disorders*, 245, 869-884. <https://doi.org/10.1016/j.jad.2018.11.073>
- 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>
- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110, 12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>

- Cohen, Z. D., Kim, T. T., Van, H. L., Dekker, J. J. M., & Driessen, E. (2020). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research*, 30(2), 137–150.  
<https://doi.org/10.1080/10503307.2018.1563312>
- Coley, R. Y., Liao, Q., Simon, N., & Shortreed, S. M. (2023). Empirical evaluation of internal validation methods for prediction in large-scale clinical data with rare-event outcomes: a case study in suicide risk prediction. *BMC Medical Research Methodology*, 23(1), 33. <https://doi.org/10.1186/s12874-023-01844-5>
- Collins, G. S., Dhiman, P., Ma, J., Schlüssel, M. M., Archer, L., Van Calster, B., et al. (2024). Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*, 384, e074819. <https://doi.org/10.1136/bmj-2023-074819>
- Deng Y., Hillygus D. S., Reiter J. P., Si Y., Zheng S. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science*, 28(2), 238–256. <https://doi.org/10.1214/13-sts414>
- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14(1), 91–118.  
<https://doi.org/10.1146/annurev-clinpsy-032816-045037>
- Fox, K. R., Huang, X., Linthicum, K. P., Wang, S. B., Franklin, J. C., & Ribeiro, J. D. (2019). Model complexity improves the prediction of nonsuicidal self-injury. *Journal of Consulting and Clinical Psychology*, 87(8), 684–692.  
<https://doi.org/10.1037/ccp0000421>

- Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, *143*(2), 187-232. <https://doi.org/10.1037/bul0000084>
- Geulayov, G., Casey, D., Bale, L., Brand, F., Clements, C., Farooq, B., Kapur, N., Ness, J., Waters, K., Tsiachristas, A., & Hawton, K. (2019). Suicide following presentation to hospital for non-fatal self-harm in the multicentre study of self-harm: a long-term follow-up study. *The Lancet Psychiatry*, *6*(12), 1021-1030. [https://doi.org/10.1016/S2215-0366\(19\)30402-X](https://doi.org/10.1016/S2215-0366(19)30402-X)
- Gieseemann, 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>
- Gómez Penedo, J. M., Schwartz, B., Gieseemann, J., Rubel, J. A., Deisenhofer, A. K., & Lutz, W. (2022). For whom should psychotherapy focus on problem coping? A machine learning algorithm for treatment personalization. *Psychotherapy Research*, *32*(2), 151–164. <https://doi.org/10.1080/10503307.2021.1930242>
- Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361–387. [https://doi.org/10.1002/\(SICI\)1097-0258\(19960229\)15:4<361::AID-SIM168>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4)



- Hilbert, K., Jacobi, T., Kunas, S. L., Elsner, B., Reuter, B., Lueken, U., & Kathmann, N. (2021). Identifying CBT non-response among OCD outpatients: A machine-learning approach. *Psychotherapy Research*, 31(1), 52–62. <https://doi.org/10.1080/10503307.2020.1839140>
- Hong, M., Jacobucci, R., & Lubke, G. (2020). Deductive data mining. *Psychological Methods*, 25(6), 691–707. <https://doi.org/10.1037/met0000252>
- Huang, X., Ribeiro, J. D., & Franklin, J. C. (2020). The differences between suicide ideators and suicide attempters: Simple, complicated, or complex? *Journal of Consulting and Clinical Psychology*, 88(6), 554–569. <https://doi.org/10.1037/ccp0000498>
- Huinink J., Brüderl J., Nauck B., Walper S., Castiglioni L., Feldhaus M. (2011). Panel Analysis of Intimate Relationships and Family Dynamics (pairfam): Conceptual Framework and Design. *Zeitschrift für Familienforschung*, 23(1), 77–101. <https://madoc.bib.uni-mannheim.de/30017/>
- Jacobsen, E., Ran, X., Liu, A., Chang, C. H., & Ganguli, M. (2021). Predictors of attrition in a longitudinal population-based study of aging. *International Psychogeriatrics*, 33(8), 767–778. <https://doi.org/10.1017/S1041610220000447>
- Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, 15(3), 809–816. <https://doi.org/10.1177/1745691620902467>
- Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of inflated prediction performance: A commentary on machine learning and suicide research. *Clinical Psychological Science*, 9(1), 129–134. <https://doi.org/10.1177/2167702620954216>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning*. Springer.

- Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). Causal Machine Learning: a survey and open problems. arXiv (Cornell University).  
<https://doi.org/10.48550/arxiv.2206.15475>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804.  
<https://doi.org/10.1016/j.patter.2023.100804S>
- Kern, C., Weiß, B., & Kolb, J.-P. (2021). Predicting nonresponse in future waves of a probability-based mixed-mode panel with machine learning. *Journal of Survey Statistics and Methodology*, 11(1), 100–123. <https://doi.org/10.1093/jssam/smab009>
- Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, 24(1), 134–161. <https://doi.org/10.1093/ectj/utaa014>
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), 5802–5805.  
<https://doi.org/10.1073/pnas.1218772110>
- 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., 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>
- Little, R. J. (2021). Missing data assumptions. *Annual Review of Statistics and Its Application*, 8(1), 89-107. <https://doi.org/10.1146/annurev-statistics-040720-031104>
- Little R. J. A., Rubin D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.

- Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A. K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour Research and Therapy*, *120*, 103438. <https://doi.org/10.1016/j.brat.2019.103438>
- Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., & Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study. *Scientific Reports*, *8*(1). <https://doi.org/10.1038/s41598-018-25953-0>
- McNamara, M. E., Zisser, M., Beevers, C. G., & Shumake, J. (2022). Not just “big” data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions. *Behaviour Research and Therapy*, *153*, 104086. <https://doi.org/10.1016/j.brat.2022.104086>
- Möttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*(3), 474–490. <https://doi.org/10.1037/pspp0000100>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. <https://doi.org/10.1126/science.aac4716>
- Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3). <https://doi.org/10.1177/25152459231162559>
- Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3). <https://doi.org/10.1002/widm.1301>

- Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated R code. *Social and Personality Psychology Compass*. <https://doi.org/10.1111/spc3.12579>
- Schafer J. L., Graham J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Schwartz, B., Cohen, Z. D., Rubel, J. A., Zimmermann, D., Wittmann, W. W., & Lutz, W. (2020). Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychotherapy Research*, 31(1), 33–51. <https://doi.org/10.1080/10503307.2020.1769219>
- Seeboth, A., & Möttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*, 32(3), 186–201. <https://doi.org/10.1002/per.2147>
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3). <https://doi.org/10.1214/10-sts330>
- Siddaway, A. P., Quinlivan, L., Kapur, N., O'Connor, R. C., & de Beurs, D. (2020). Cautions, concerns, and future directions for using machine learning in relation to mental health problems and clinical and forensic risks: A brief comment on “model complexity improves the prediction of nonsuicidal self-injury” (Fox et al., 2019). *Journal of Consulting and Clinical Psychology*, 88(4), 384–387. <https://doi.org/10.1037/ccp0000485>
- Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality*, 34(5), 613–631. <https://doi.org/10.1002/per.2257>

- Stewart, R. D., Möttus, R., Seeboth, A., Soto, C. J., & Johnson, W. (2022). The finer details? The predictability of life outcomes from Big Five domains, facets, and nuances. *Journal of personality, 90*(2), 167–182. <https://doi.org/10.1111/jopy.12660>
- Tantithamthavorn, C., Mcintosh, S., Hassan, A. E., & Matsumoto, K. (2017). An empirical comparison of model validation techniques for defect prediction models. *IEEE Transactions on Software Engineering, 43*(1), 1–18. <https://doi.org/10.1109/tse.2016.2584050>
- Van Calster, B., van Smeden, M., De Cock, B., & Steyerberg, E. W. (2020). Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research, 29*(11), 3166–3178. <https://doi.org/10.1177/0962280220921415>
- Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J., Van Hoecke, S., & Demeester, T. (2021). Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine, 111*, 101987. <https://doi.org/10.1016/j.artmed.2020.101987>
- Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *Journal of Child Psychology and Psychiatry, 59*(12), 1261-1270. <https://doi.org/10.1111/jcpp.12916>
- Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology, 88*(1), 25–38. <https://doi.org/10.1037/ccp0000451>

- Wilkinson, J., Arnold, K. F., Murray, E. J., van Smeden, M., Carr, K., Sippy, R., de Kamps, M., Beam, A., Konigorski, S., Lippert, C., Gilthorpe, M. S., & Tennant, P. W. G. (2020). Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health*, 2(12), e677–e680.  
[https://doi.org/10.1016/S2589-7500\(20\)30200-4](https://doi.org/10.1016/S2589-7500(20)30200-4)
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12, 1100–1122. <https://doi.org/10.1177/1745691617693393>
- Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, 112(4), 1036–1040.  
<https://doi.org/10.1073/pnas.1418680112>
- Zinn, S., & Gnambs, T. (2020). Analyzing nonresponse in longitudinal surveys using Bayesian Additive Regression Trees: A nonparametric event history analysis. *Social Science Computer Review*. <https://doi.org/10.1177/0894439320928242>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.  
<https://doi.org/10.1111/j.1467-9868.2005.0050>

# Validation and generalizability of machine learning prediction models on attrition in longitudinal studies

Kristin Jankowsky<sup>1</sup>, Ulrich Schroeders<sup>1</sup>

1: University of Kassel

Status – accepted

Jankowsky, K. & Schroeders, U. (2022). Validation and generalizability of machine learning prediction models on attrition in longitudinal studies. *International Journal of Behavioral Development*, 46(2), 169–176. <https://doi.org/10.1177/01650254221075034>

# Validation and generalizability of machine learning prediction models on attrition in longitudinal studies

International Journal of  
Behavioral Development  
2022, Vol. 46(2) 169–176  
© The Author(s) 2022



Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/01650254221075034  
journals.sagepub.com/home/jbd



Kristin Jankowsky<sup>1</sup> and Ulrich Schroeders<sup>1</sup>

## Abstract

Attrition in longitudinal studies is a major threat to the representativeness of the data and the generalizability of the findings. Typical approaches to address systematic nonresponse are either expensive and unsatisfactory (e.g., oversampling) or rely on the unrealistic assumption of data missing at random (e.g., multiple imputation). Thus, models that effectively predict who most likely drops out in subsequent occasions might offer the opportunity to take countermeasures (e.g., incentives). With the current study, we introduce a longitudinal model validation approach and examine whether attrition in two nationally representative longitudinal panel studies can be predicted accurately. We compare the performance of a basic logistic regression model with a more flexible, data-driven machine learning algorithm—gradient boosting machines. Our results show almost no difference in accuracies for both modeling approaches, which contradicts claims of similar studies on survey attrition. Prediction models could not be generalized across surveys and were less accurate when tested at a later survey wave. We discuss the implications of these findings for survey retention, the use of complex machine learning algorithms, and give some recommendations to deal with study attrition.

## Keywords

Machine learning, attrition, longitudinal studies, predictive modeling, generalizability

Data of longitudinal panel surveys constitute an important resource for educational, psychological, sociological, and health-related research (e.g., Behr et al., 2020; Rackoff & Newman, 2020). In contrast to cross-sectional data, longitudinal data allow to study developmental trajectories or within-person change in addition to between-person differences (Voelkle et al., 2014). However, the strength of longitudinal designs—assessing the same individuals at multiple occasions—also entails the risk of attrition, which is defined as temporary or permanent dropout of participants. High attrition rates are a major problem in longitudinal research affecting the validity of conclusions drawn from such data (Schoeni et al., 2012). More precisely, systematic dropout of participants sharing common characteristics (e.g., low socioeconomic status) renders the remaining sample unrepresentative, which in turn can lead to biased results (Heffetz & Reeves, 2019; Little & Rubin, 2002). For example, a longitudinal study on the effects of counseling on depression in which participants with the highest depression scores are most likely to drop out of the sample would falsely indicate a therapy to be more effective (Nicholson et al., 2017).

With the current study, we try to predict attrition using data-driven machine learning algorithms. Insights about relevant predictors can then be used to take potentially more effective measures to anticipate and prevent attrition such as targeted incentives for at-risk participants (Lynn, 2017; Pforr et al., 2015). We compare the predictive accuracy of logistic regressions models with a machine learning algorithm, namely, gradient boosting machines (GBM; Friedman, 2001) in two longitudinal panel

studies: Midlife in the United States (MIDUS) and Panel Analysis of Intimate Relationships and Family Dynamics (pairfam). Finally, we evaluate our results in terms of generalizability across studies and survey waves, respectively.

## Strategies in Dealing With Panel Attrition

In the following, we will shortly present methods that are used to ensure the representativeness of the sample—(a) statistical modeling, (b) poststratification weights, or (c) oversampling/refreshment samples—and discuss their strengths and limitations. First, to address wave nonresponse, that is, participants' data completely missing for a study wave in longitudinal studies, one could use the same procedures that are recommended in the missing data literature for item nonresponse (e.g., Enders, 2010; Little & Rubin, 2002). However, imputation-based or model-based approaches rely on the assumption of *missing at random* (Schafer & Graham, 2002), that is, the occurrence of missing values does not depend on the expression of the variable itself or

<sup>1</sup> University of Kassel, Germany

### Corresponding author:

Kristin Jankowsky, Psychological Assessment, Institute of Psychology, University of Kassel, Holländische Str. 36-38, 34127 Kassel, Germany.  
Email: Jankowsky@psychologie.uni-kassel.de



on the expression of other variables in the data set after controlling for other observed variables. This prerequisite is problematic, as participants' most likely drop out systematically (*missing not at random*) and variables that are associated with this process are often unknown in advance or difficult to measure. However, recently promising approaches on handling non-random missing data have been developed (for an overview, see Kleinke et al., 2020; Van Buuren, 2018). Researchers often try to reduce potential bias by incorporating relevant auxiliary variables in multiple imputation that might produce robust results despite common concerns (Mustillo & Kwon, 2014), but not in all cases (Hardt et al., 2012). Simpler methods such as listwise or pairwise deletion are used regularly and often lead to biased estimates (Jeličić et al., 2009).

A second approach to compensate for attrition bias is to use poststratification weights. Groups or individuals are assigned weights according to their inversed probability of participation (Seaman & White, 2013). Thus, the usefulness of weighting hinges on whether all relevant predictors of attrition are integrated into the statistical model that is used to calculate these probabilities (Gelman, 2007). As weighting does not replace missing values and requires complete data, any occurring item nonresponse must be addressed beforehand (e.g., using multiple imputation). Consequently, the later waves' sample sizes of a longitudinal study still lack statistical power. Also, weights often lead to an increased variance of estimators (Schmidt & Woll, 2017) and must be adjusted depending on which study waves or variables are analyzed.

A third approach is oversampling, which refers to the countermeasure of recruiting more participants who are likely to drop during a longitudinal study. Oversampling recognizes attrition as inevitable and tries to buffer the unavoidable unrepresentativeness of the data and to reduce selection bias by starting with an unbalanced sample at baseline. Following a similar logic, refreshment samples consist of new participants added at subsequent measurement occasions that are often sampled using the same sampling procedure as for the initial recruitment (Deng et al., 2013). Whereas additional participants generally enhance statistical power, it has been advised to select refreshment participants who share characteristics with nonrespondents to avoid introducing bias (Dorsett, 2010). Additional negative aspects of using oversampling or refreshment samples are their high costs and that they often not sufficiently compensate bias and therefore have to be combined with other strategies.

### Drawbacks of Common Approaches to Analyzing Panel Attrition

Previous studies often examined attrition with different variables that are routinely collected at baseline such as demographic variables using logistic regressions (Eisner et al., 2018). This research repeatedly reported that males, singles, people with migration background, less educated, and urban living participants are at higher risk of becoming nonrespondents (Radler & Ryff, 2010; Young et al., 2006). Given that longitudinal studies usually focus on a specific topic and that panels are time-restricted, the breadth and depth of these variables are somewhat limited. But it is plausible to assume that the decision to (regularly) take part in longitudinal studies can be influenced by several factors beyond

demographics such as personality (e.g., Lugtig, 2014) or health (e.g., Jacobsen et al., 2021). However, studies on personality or health focus on specific sets of variables, neglecting others.

Taken together, the selection and quantity of predictors used in previous research to predict attrition are often limited. Moreover, the assumption of exclusively linear effects on attrition is questionable. Radler and Ryff (2010) showed that, for example, age interacted with subjective health when predicting attrition in the second study wave of MIDUS: Elderly participants only had a higher attrition probability when they also rated their subjective health as poorly, whereas older participants in excellent health showed significantly lower attrition rates. Not addressing such interaction effects may result in less accurate models.

Another common drawback of traditional attrition modeling approaches is that it is unclear whether their results are generalizable. The ability of a model to provide accurate and generalizable predictions is especially essential in applied research (Rocca & Yarkoni, 2020; Shmueli, 2010) such as study retention. To enable panel administrators to employ effective retention strategies (e.g., person-specific incentives at future waves), a prediction model also has to hold in future waves. In general, to quantify the unbiased predictive accuracy, any model must be evaluated on new data, which is often achieved by splitting a data set into a training-validation and a testing data set. However, the question whether a model predicting attrition will also hold in future waves or across different longitudinal studies goes beyond this form of internal cross-validation. Rather, it aims at the *generalizability* of the results. Generalizability concerns the extent to which the study results apply across different items assessing the same construct (item sampling), across different participants (person sampling), across different measurement occasions (time sampling), and across different analytical methods (method sampling). As these aspects of longitudinal testing are of particular interest for study planning, researchers should ask to what extent their prediction models generalize across them.

### Predicting Panel Attrition Using Machine Learning

A few recent studies have picked up on the notion of temporally validating their models of attrition and including nonlinear and interaction effects by using machine learning algorithms to predict attrition in longitudinal studies (Jacobsen et al., 2021; Kern et al., 2019; Zinn & Gnams, 2020). Machine learning algorithms are often recommended to efficiently deal with extensive data, collinearity of predictors, and complex relations between predictors and outcomes (e.g., Zou & Hastie, 2005). The assumption in these studies is that the reasons for participants to drop are complex and that the complexity of the method should match this causal complexity. For example, Kern et al. (2019) used different sets of predictors with various machine learning algorithms to predict attrition in a longitudinal German panel study. To validate their prediction models, the authors performed temporal cross-validation, which consisted of the following steps: A prediction model was built using data of all participants present at Wave 1 to predict the participation status at Wave 2. The resulting model is then tested using all active participants of Wave 2 to predict participation status at Wave 3. This validation approach was repeated for all 18 survey waves.

Using baseline variables and information on previous response behavior, a random forest algorithm achieved the highest predictive accuracy with an average *area under the curve* (AUC) of .875.<sup>1</sup> However, these promising results must be taken with a grain of salt. First, participants were automatically excluded from the panel when they were inactive for three waves in a row which is problematic because the outcome is logically dependent on a set of predictors, leading to inflated accuracies. Second, due to the temporal cross-validation scheme, most participants in the training data remain in the test data at later waves. Although this might seem justified at first glance since the study results do not have to generalize to other participants outside the given study sample, from a statistical point of view, an overlap of participants in training and test data leads to inflated accuracies, especially for tree-based algorithms (e.g., Jacobucci et al., 2021).

## The Present Study

The present study has three main objectives: First, we aim to empirically test the notion that attrition can be predicted more accurately by means of machine learning algorithms that are able to incorporate nonlinear or interaction effects of heterogeneous predictors. To this end, we compare the predictive accuracy of a tree-based machine learning algorithm, GBM, and a logistic regression model. GBM sequentially combine multiple single decision trees that usually have a comparably poor predictive accuracy (Breiman, 2001). One advantage of GBM is that researchers do not have to a priori parameterize the relationship between an outcome and its predictors, which makes them popular for supervised classification tasks (e.g., Schroeders et al., 2022).

Second, we are interested in the longitudinal predictive accuracy of models on attrition. To validate prediction models, we employ a temporal validation approach with strictly disjoint training and testing data. This model validation strategy represents a stricter and more realistic test of predictive accuracy for future survey waves that are not bound to a specific group of participants. The third goal of this study is to tackle this issue of generalizability. Thus, we compare the prediction of attrition across two longitudinal large-scale studies that differ greatly in their study aims, sample, time frame: While one study is primarily concerned with midlife development of health and well-being in the United States with one wave every 9 years, the other is an annual German survey on partnership and family dynamics. Both studies measure similar constructs in their baseline assessment albeit sometimes using slightly different items. In terms of dimensions of generalizability, the items, persons, and time frame differ to a substantial degree allowing to gauge the generalizability of results across studies.

## Method

### Sample and Design

**MIDUS.** MIDUS is an American national survey carried out by the MacArthur Midlife Research Network (Brim et al., 2004). Each survey wave consists of a phone interview and additional questionnaires that participants have to send back. Starting in 1995, there was a random digit dialing sample of 4,244 participants as well as siblings of some of these participants ( $N=950$ ) and a twin sample ( $N=1,914$ ). Subsequent survey waves of

MIDUS were conducted 9 years later in 2004 (second wave) and in 2013 (third wave). More information about MIDUS and the data of the first three waves can be found at <http://www.midus.wisc.edu/data/>. We consider participants as responding if they completed all parts of a survey wave. Therefore, we only use the subset of participants who completed all parts of survey at the first study wave ( $N=6,325$ ).

**Pairfam.** Pairfam is an annually conducted national survey on partnership and family dynamics in Germany (Huinink et al., 2011). It started in 2008 with a sample of 12,402 participants from three age cohorts (1971–1973, 1981–1983, and 1991–1993). Information about the participants are gathered via computer-assisted personal interviewing. Participants who were nonresponding in a previous wave, but did not explicitly decline their participation, are contacted again. After two nonresponses in a row, participants are excluded from the panel. The scientific use file and more information can be accessed at <https://www.pairfam.de/>. The following analyses were conducted on a subset of  $N=11,875$ , because we excluded 527 participants with implausible values ( $BMI > 50$ ).

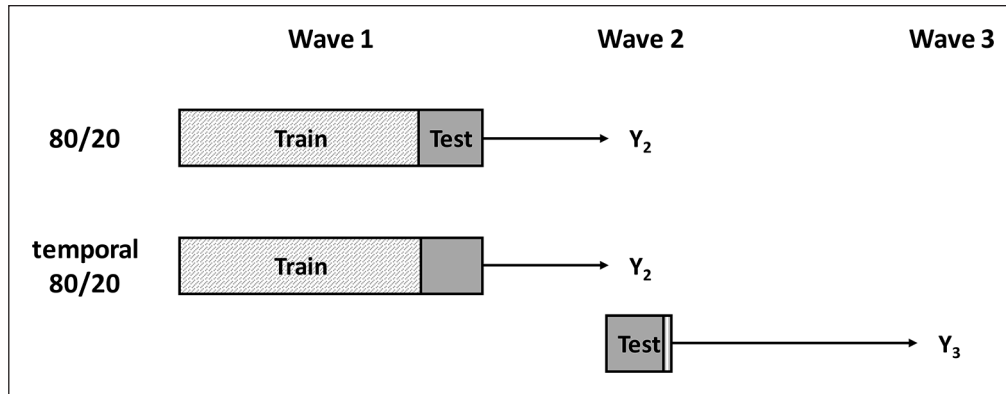
### Measures

We used core demographics, health, and personality related variables that have been shown to correlate with longitudinal attrition in previous studies and were available at baseline, except for personality in the pairfam study (see Supplemental Table S1 at <https://osf.io/usjr7/>). All categorical variables were dummy coded prior to the analysis using the first category as reference. The outcome participation status was dichotomously coded, irrespective of the reason.

### Statistical Analyses

The current analyses are prediction models based on logistic regressions and gradient boosted machines. Irrespective of the algorithm, one important issue of any prediction is to reduce overfit, that is, to reduce the tendency of “statistical models to mistakenly fit sample-specific noise as if it were signal” (Yarkoni & Westfall, 2017, p. 3) while obtaining the highest predictive accuracy possible. To quantify the “true” or unbiased predictive accuracy, any prediction model has to be evaluated on new data—also called test data or withhold sample (Rocca & Yarkoni, 2020; Yarkoni & Westfall, 2017). Validating a prediction model with new data of an independent study is the most rigorous way of testing its generalizability (Dwyer et al., 2018). However, this is not always a feasible option and researchers often resort to workarounds such as multiple splitting their data into a training and testing data set to obtain robust estimates that resolve overfitting.

We used the following two validation strategies for the first three survey waves of MIDUS and pairfam, respectively: First, we ignored the temporal aspect of predicting future events and split the data into training data (80%) and testing data (20%; see the upper part of Figure 1), that is, training and testing the predictive model was done using the same measurement occasion (Wave 2). Second, we added a temporal validation strategy in which the aforementioned splitting of the data in strictly disjoint training and testing data is combined with temporal model



**Figure 1.** Different Cross-Validation Approaches in a Longitudinal Study Context. Excluded participants at Wave 2 are represented by the white section of the rectangle.

validation (see the lower part of Figure 1). More precisely, we trained the model on 80% of the data at Wave 1 to predict status at Wave 2 and tested the resulting model using the active participants of the remaining 20% at Wave 2 to predict participation status at Wave 3. In doing so, we avoided any overlap of training and testing data and were also able to validate the prediction of the participation status of a future Wave 3.

To avoid biased predictions due to highly unbalanced data, we used up-sampling to match the sample size of nonrespondents to respondents in the training data. The testing data were not affected by this procedure. Missing values were imputed separately for the training and testing data (i.e., after the 80/20 split) using the  $k$ -nearest neighbors algorithm implemented in *caret*. Nearest neighbor imputation procedures are hot-deck imputations in which a given number ( $k$ ) of observations that are similar to the observation with a missing value (according to a distance metric, in this analysis the Euclidean distance) are used to replace missing values (e.g., Beretta & Santaniello, 2016). We used the default settings for imputation which were mean values of  $k=5$ . For training the models, we used 10-fold cross-validation. To evaluate the classification into respondents and nonrespondents, we report balanced accuracy, that is, the mean of sensitivity and specificity. Sensitivity represents the ratio of correctly identified nonrespondents to all nonrespondents; specificity represents the ratio of correctly identified respondents to all respondents. Balanced accuracy was calculated for each testing data set of the 1,000 iterations.

All analyses were conducted using the R package *caret* (Kuhn, 2008) as an interface for modeling and prediction. We compared the predictive accuracy of a logistic regression and the GBM algorithm of the R package *gbm* (Version 2.1.5; Greenwell et al., 2019). We used the following default settings for the *gbm* tuning parameters: interaction depth of 1, 2, or 3; a minimum leaf size of 10; a shrinkage of .10; and number of trees 50, 100, or 150. As a sensitivity check of so-called hyperparameters on study results, we compared the default settings with a larger grid (interaction depth of 1, 2, 3, or 4, a minimum leaf size of 10, 20, or 50, a sequence of shrinkage values between .001 and .201 using steps of .01, and the number of trees 50, 100, 150, 300, or 500). The overall number of combinations in the larger grid was 1,260 as opposed to nine in the default settings. Considering that we split the data 1,000 times, we estimated 1,260,000 models with

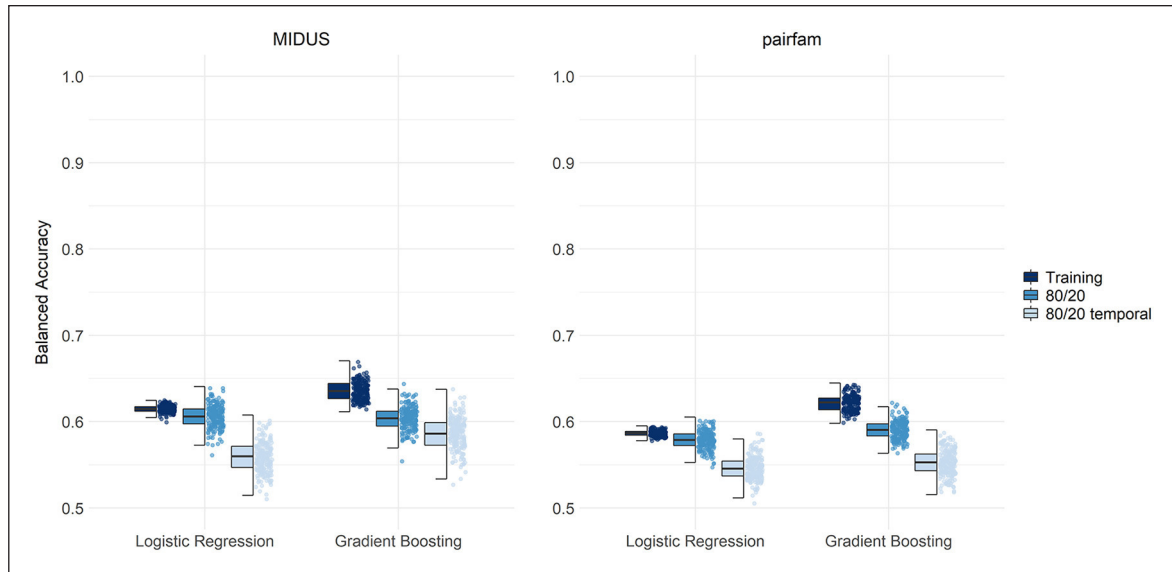
the larger grid compared with 9,000 with the default grid. Supplemental Figure S1 (see at <https://osf.io/usjr7/>) shows the balanced accuracies for MIDUS and pairfam and both validation strategies for both grids. The results show that the larger grid did not lead to any substantial improvement in the predictive accuracy. Thus, we focus the presentation and discussion of our results on those of the default grid. Annotated analyses scripts are available at <https://osf.io/usjr7/>.

## Results

Following a suggestion of an anonymous reviewer, we checked whether the quality of the data at hand is eligible to be analyzed with the proposed methods. Results of this kind of “prestudy” showed that the data can be analyzed with logistic regression and GBM, that is, that the prediction accuracy can be reproduced given a known missing procedure. More information on these analyses can be found in a supplement in the OSF project at <https://osf.io/usjr7/>.

Both samples differ with respect to persons studied, items administered, and time frame considered. For example, participants of MIDUS were on average 21 years old, had an 11 percentage points lower share of migration background, and were more than twice as likely married than participants of the pairfam study. Education level and occupation status were measured differently across both studies and MIDUS had more information on chronic health conditions and personality than pairfam. With respect to attrition, in MIDUS 38% dropped from first to second wave (i.e., 2,396 of initially 6,325 participants) and another 20% from the second to third wave (1,283). In pairfam, 27% dropped out from first to second wave (i.e., 3,174 of the initial 11,875 participants) and another 9% from second to third wave. We provide an extensive Supplemental Table S1 showing descriptive statistics of all predictor variables for MIDUS and pairfam, respectively, and correlation plots of all predictor variables and participation status in the OSF project.

Figure 2 shows the balanced accuracies of 1,000 iterations for the logistic regressions and the GBM models for both studies and both validation approaches. Overall, it was not possible to accurately differentiate between nonrespondents and respondents. In the following, we will consider the results of the traditional 80/20



**Figure 2.** Balanced Accuracies for Predicting Attrition in MIDUS and pairfam. The boxplots represent the interquartile range, the solid line represents the median, and the whiskers 1.5 times the interquartile range. Balanced accuracy values of 200 randomly selected values are displayed as jittered distribution on the right with outliers as triangles.

validation approach first. The amount of overfit (i.e., difference in the balanced accuracies between training and testing sample) was less pronounced for the logistic regressions (a difference in balanced accuracies of  $<.01$  for MIDUS and  $.01$  for pairfam) than for the GBM ( $.04$  for MIDUS and  $.03$  for pairfam). In general, both algorithms yielded almost identical balanced accuracies.

Next, we focus on the disjoint temporal cross-validation. As to the question whether GBM outperforms logistic regression, the findings are mixed: Logistic regression yielded averaged balanced accuracies of  $.56$  (MIDUS) and  $.55$  (pairfam), and GBM achieved  $.59$  (MIDUS) and  $.55$  (pairfam) in the 80/20 temporal validation. Considering the much higher computational effort, the more complex (and ambiguous) model interpretation in GBM, and the mediocre overall balanced accuracies, the differences were—as in the traditional 80/20 validation approach—rather small and negligible.

To evaluate whether the respective models can be used for predicting attrition in future waves, the comparison of accuracies across both approaches are of particular interest. A decline in accuracies between the traditional 80/20 and the disjoint 80/20 approach was observed: For MIDUS, the averaged balanced accuracies of the 80/20 approach were higher (logistic regression:  $.61$ , GBM:  $.60$ ) than those of the 80/20 temporal validation approach ( $.56$  and  $.59$ , respectively). For pairfam, the nontemporal approach yielded higher averaged balanced accuracy values of  $.58$  (logistic regression) and  $.59$  (GBM) than the temporal validation with both  $.55$ . In summary, the already inaccurate prediction models lost further predictive accuracy when validated in a longitudinal framework.

The corresponding specificities and sensitivities for all models can be found as Supplemental Figures S3 and S4 in the OSF project (see at <https://osf.io/usjr7/>). For MIDUS, the averaged sensitivities were  $.53$  (logistic regression) and  $.55$  (GBM) and thus lower than the averaged specificities ( $.59$  and  $.62$ , respectively). For pairfam, the averaged sensitivities were  $.55$  (logistic

regression) and  $.53$  (GBM), hence nearly the same as the averaged specificities ( $.54$  and  $.57$ , respectively). To conclude, these differences are rather small, but for MIDUS, the group of respondents could be detected slightly more accurately compared with the nonrespondents. These sensitivities translate to positive predictive values (i.e., the proportion of true nonrespondents of all participants who were flagged as nonrespondents) of  $.39$  (logistic regression) and  $.41$  (GBM) for MIDUS and  $.21$  (logistic regression) and  $.22$  (GBM) for pairfam.

### Which Variables Predict Attrition?

For an overview of variable importances, we present the standardized regression coefficients of the logistic regression models averaged across all 1,000 iterations in Table 1. Overall, there was little consistency in regression coefficients across both surveys. For example, in MIDUS, the highest level of education was the predictor with the largest effect on attrition, whereas the level of education was not among the most important predictor variables in pairfam. Age had a negative effect on attrition in MIDUS (i.e., older participants were more likely to participate again) and a positive one in pairfam. In pairfam, the migration background was the second-most important variable, whereas in MIDUS migration background played no significant role in predicting attrition.

### Discussion

High rates of systematic attrition can lead to biased results of studies using longitudinal data (Heffetz & Reeves, 2019; Little & Rubin, 2002). We argued that the optimal way to deal with attrition is to prevent it as best as possible, for example, with target-specific incentives. To achieve this goal, predicting attrition in future survey waves is more important than explaining possible underlying causal relationships of attrition. Thus, we focus on the prediction of attrition using machine learning algorithms in a



**Table 1.** Averaged Standardized Coefficients of the Logistic Regression Models.

No.	MIDUS			pairfam		
	Variable	M	SD	Variable	M	SD
1	Highest level of education	-.30	.02	Full-time employment	.24	.03
2	Age	-.27	.03	Migration background	.19	.02
3	Sex	-.25	.02	Homemaker	.15	.02
4	Instrumental activities of daily living	.19	.02	Number of household members	-.13	.02
5	Widow or widower	.16	.02	Age	.13	.03
6	Separated	.15	.02	Vocational training	.11	.01
7	Agreeableness	.15	.02	Self-employed	.11	.02
8	Conscientiousness	-.15	.02	Unemployed	.11	.02
9	Physical health, self-evaluated	-.14	.02			
10	Divorced	.13	.02			
11	Never been married	.12	.02			
12	Current employment—Retired	-.12	.03			
13	BMI	-.12	.02			

Note. MIDUS: Midlife in the United States; pairfam: Panel Analysis of Intimate Relationships and Family Dynamics. Regression coefficients <.10 are not displayed.

longitudinal validation framework. The results of this study showed that the issue of attrition cannot be easily solved by applying more complex statistical models, that is, GBM did not outperform logistic regression analyses in predictive accuracy.

From a practical point of view, a central question is which strategy in dealing with attrition—target-specific incentives, equal distribution of incentives, over- or refreshment sampling—is most promising or cost-effective. The answer to this seemingly straightforward question depends on several parameters. For the following thought experiment, we focus on three of these parameters: (a) the overall available resources, (b) the percentage of participants who remained instead of dropping out, and (c) the positive predictive value of a prediction model. Let us assume that there is a budget of €20,000 available to implement retention measures to retain as much as possible of 1,000 (of 4,000) participants that are at risk of dropping out at a next survey wave. As a first strategy, one could prophylactically provide all 4,000 participants with incentives worth €5 such as sending thank you and birthday cards. With small investments per person, assuming a persuading effect of 5%, 50 of 1,000 at-risk participants could be converted.

A second approach could be to incentivize only those participants identified at risk of dropping out by a predictive model with €50 and assume that this will have the desired effect (staying active participants in the study) on 50% of them. The success of this second strategy depends on the predictive accuracy of the model. Within the budget of €20,000, using a perfect prediction model (positive predictive value=1), it would be possible to persuade 200 participants to stay in the study (i.e., €20,000 / €50=400 participants, all of them get correctly flagged and funded, and half of them get convinced to stay). A model with a positive predictive value of .40 (as in our results for MIDUS) would still result in 80 participants (i.e., €20,000 / €50=400 participants, 40% of them get correctly flagged and funded, and half of them get convinced to stay). With a dropout rate of 25%, a model that is as accurate as random guessing would have a positive predictive value of .25 and result in 50 convinced participants (i.e., €20,000 / €50=400 participants, 25% of them get correctly flagged and funded, and half of them get convinced to

stay). Thus, even small increments in positive predictive value translate into more successful retention of participants. However, there is no one-size-fits-all strategy that researchers must apply, rather the conditions of the individual longitudinal study have to be taken into account.

A third approach to deal with attrition could be to renounce the attempt of persuading participants and to sample new participants to replace all dropouts (refreshment). The cost of this approach depends on the number of waves a participant has been active (because the participants' "value" accumulates across study waves) and on the resources needed for an assessment (e.g., online surveys are more economical than extensive examinations by medical professionals). However, retaining participants is always preferable over recruiting new ones (e.g., for analyzing intraindividual trajectories).

### On the Generalizability of Prediction Models

The results concerning the variable importance were not generalizable across studies. In the introduction, we proposed four dimensions of generalizability: item sampling, person sampling, time sampling, and method sampling. First, different items and operationalizations of the same constructs (e.g., education and occupation) could have led to differences in variable importances. But also different cultural contexts could have a moderating effect. For example, although the participants' migration background was defined in the same way in both studies, it could have a diverging effect due to different cultural and political implications in the United States and Germany (e.g., Berry et al., 2006). Second, the participants of MIDUS and pairfam already differed from each other at the respective baseline assessments. These different populations combined with the different topics of the panels also contributed to the nongeneralizability of effects: MIDUS is primarily concerned with midlife development of health and well-being, maybe leading to higher responding rates in older participants. In pairfam, younger nonsingle participants are more likely to participate again, which fits in with the fact that pairfam is a survey on partnership and family dynamics. Third, in MIDUS the survey waves are 9 years apart, whereas

pairfam has annual survey waves and therefore places a higher burden on the resources of participants. However, regardless of the mechanisms underlying these differences, a model developed using MIDUS data cannot be used to predict attrition in pairfam and vice versa.

In addition to this nongeneralizability across items and persons, which also is true for cross-sectional studies, the nongeneralizability across measurement occasions is a specific that complicates matters in longitudinal studies. There is a very plausible explanation for this: If participants with certain characteristics drop out more likely, some of them will be no longer active participants at the next survey wave, altering the population for which nonresponse is to be predicted at a following survey wave. Either the same predictors also contribute to the prediction of nonresponse for the remaining individuals at future waves or their effects and importance also shift. The results of this study support the latter notion, that is, the *reasons* why people dropout change jointly with the participants. However, if one and the same model does not apply to or fit equally well for multiple survey waves, it is not useful for proactively planning survey retention strategies.

### More Complex Models Are Not Better Suited to Predict Attrition

With respect to the last dimension of generalizability, the method sampling, the results are intriguing: The more complex data-driven models did not lead to substantial incremental in predictive accuracy in comparison with simple, logistic models. From this, one can conclude that the effects are mostly linear and that for reasons of parsimony a less complex model is preferable over computationally extensive and harder to interpret algorithms. The question arises, however, why other recent studies using machine learning algorithms to predict survey attrition reported relatively high predictive accuracies (e.g., Kern et al., 2019; Zinn & Gnambs, 2020). There are two reasons: First, in studies reporting higher accuracies, the previous response status was used as a predictor variable that, on one hand, was the most important predictor variable. However, on the other hand, this information is not available in longitudinal surveys without temporal nonrespondents (i.e., participants coming back at later study waves) as in this study. Second, it has been found that machine learning algorithms outperforming more simple models is often due to an insufficient distinction between training and testing samples (e.g., Jacobucci et al., 2021). In this study and in contrast to the traditional validation approach, we used a validation approach that also guarantees disjoint training and testing samples in a longitudinal context. Consequently, our predictive accuracies were lower.

To sum up, our rather strict approach at testing the accuracy of attrition models involving different survey occasions, two greatly differing longitudinal studies, and the comparison of a more basic modeling approach with a complex machine learning algorithm shed light on seldom asked, let alone solved problems within survey retention research. Since attrition models could not be generalized across studies and measurement occasions and their predictive accuracies were low in general, there is no clear answer to the question how to best tackle the issue of longitudinal attrition. However, under specific assumptions, even models with relatively low accuracies could be a useful tool for targeted incentives and for survey planning.

### Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Kristin Jankowsky  <https://orcid.org/0000-0002-4847-0760>

Ulrich Schroeders  <https://orcid.org/0000-0002-5225-1122>

### Supplemental Material

Supplemental material for this article is available online.

### Note

1. In the current case of classification, area under the curve (AUC) values range from .50 to 1.00, the former indicating an accuracy as good as a random guess and the latter a perfect discrimination between groups.

### References

- Behr, A., Giese, M., Tegum Kamdjou, H. D., & Theune, K. (2020). Early prediction of university dropouts—A Random Forest approach. *Jahrbücher für Nationalökonomie und Statistik*, *240*(6), 743–789. <https://doi.org/10.1515/jbnst-2019-0006>
- Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, *16*, Article 74. <https://doi.org/10.1186/s12911-016-0318-z>
- Berry, J. W., Phinney, J. S., Sam, D. L., & Vedder, P. (2006). Immigrant youth: Acculturation, identity, and adaptation. *Applied Psychology*, *55*(3), 303–332. <https://doi.org/10.1111/j.1464-0597.2006.00256.x>
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brim, O. G., Ryff, C. D., & Kessler, R. C. (2004). The MIDUS national survey: An overview. In O. G. Brim, C. D. Ryff, & R. C. Kessler (Eds.), *How healthy are we? A national study of well-being at midlife* (pp. 1–34). University of Chicago Press.
- Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., & Zheng, S. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science*, *28*(2), 238–256. <https://doi.org/10.1214/13-sts414>
- Dorsett, R. (2010). Adjusting for nonignorable sample attrition using survey substitutes identified by propensity score matching: An empirical investigation using labour market data. *Journal of Official Statistics*, *26*(1), 105–125.
- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. <https://doi.org/10.1146/annurev-clinpsy-032816-045037>
- Eisner, N. L., Murray, A. L., Eisner, M., & Ribeaud, D. (2018). A practical guide to the analysis of nonresponse and attrition in longitudinal research using a real data example. *International Journal of Behavioral Development*, *43*(1), 24–34. <https://doi.org/10.1177/0165025418797004>

- Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153–164. <https://doi.org/10.1214/088342306000000691>
- Greenwell, B., Boehmke, B., & Cunningham, J., & GBM Developers. (2019). *gbm: Generalized boosted regression models* (Version 2.1.5) [Computer software]. <https://CRAN.R-project.org/package=gbm>
- Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology*, 12(1), Article 184. <https://doi.org/10.1186/1471-2288-12-184>
- Heffetz, O., & Reeves, D. B. (2019). Difficulty of reaching respondents and nonresponse Bias: Evidence from large government surveys. *Review of Economics and Statistics*, 101(1), 176–191. [https://doi.org/10.1162/rest\\_a\\_00748](https://doi.org/10.1162/rest_a_00748)
- Huinink, J., Brüderl, J., Nauck, B., Walper, S., Castiglioni, L., & Feldhaus, M. (2011). Panel Analysis of Intimate Relationships and Family Dynamics (pairfam): Conceptual Framework and Design. *Zeitschrift für Familienforschung*, 23(1), 77–101. <https://madoc.bib.uni-mannheim.de/30017/>
- Jacobsen, E., Ran, X., Liu, A., Chang, C.-C. H., & Ganguli, M. (2021). Predictors of attrition in a longitudinal population-based study of aging. *International Psychogeriatrics*, 33, 767–778. <https://doi.org/10.1017/s1041610220000447>
- Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of inflated prediction performance: A commentary on machine learning and suicide research. *Clinical Psychological Science*, 9(1), 129–134. <https://doi.org/10.1177/2167702620954216>
- Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, 45(4), 1195–1199. <https://doi.org/10.1037/a0015665>
- Kern, C., Weiss, B., & Kolb, J.-P. (2019). A longitudinal framework for predicting nonresponse in panel surveys. *arXiv:1909.13361*
- Kleinke, K., Reinecke, J., Daniel, S., & Spiess, M. (2020). *Applied multiple imputation: Advantages, pitfalls, new developments and applications in R*. Springer.
- Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5), 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.
- Lutig, P. (2014). Panel attrition: Separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods & Research*, 43(4), 699–723. <https://doi.org/10.1177/0049124113520305>
- Lynn, P. (2017). From standardised to targeted survey procedures for tackling nonresponse and attrition. *Survey Research Methods*, 11(1), 93–103. <https://doi.org/10.18148/srm/2017.v11i1.6734>
- Mustillo, S., & Kwon, S. (2014). Auxiliary variables in multiple imputation when data are missing not at random. *The Journal of Mathematical Sociology*, 39(2), 73–91. <https://doi.org/10.1080/0022250x.2013.877898>
- Nicholson, J. S., Deboeck, P. R., & Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development*, 41(1), 143–153. <https://doi.org/10.1177/0165025415618275>
- Pffor, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Fräbldorf, M., Hajek, K., Helmschrott, S., Kleinert, C., Koch, A., Kreiger, U., Kroh, M., Saßenroth, D., Schmiedeberg, C., Trüdinger, E. -M., & Rammstedt, B. (2015). Are incentive effects on response rates and nonresponse bias in large-scale, face-to-face surveys generalizable to Germany? Evidence from ten experiments. *Public Opinion Quarterly*, 79(3), 740–768. <https://doi.org/10.1093/poq/nfv014>
- Rackoff, G. N., & Newman, M. G. (2020). Reduced positive affect on days with stress exposure predicts depression, anxiety disorders, and low trait positive affect 7 years later. *Journal of Abnormal Psychology*, 129(8), 799–809. <https://doi.org/10.1037/abn0000639>
- Radler, B. T., & Ryff, C. D. (2010). Who participates? Accounting for longitudinal retention in the MIDUS national study of health and well-being. *Journal of Aging and Health*, 22(3), 307–331. <https://doi.org/10.1177/0898264309358617>
- Rocca, R., & Yarkoni, T. (2020, November 12). *Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction*. <https://doi.org/10.31234/osf.io/e437b>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Schmidt, S., & Woll, A. (2017). Longitudinal drop-out and weighting against its bias. *BMC Medical Research Methodology*, 17, Article 164. <https://doi.org/10.1186/s12874-017-0446-x>
- Schoeni, R. F., Stafford, F., Mcgonagle, K. A., & Andreski, P. (2012). Response rates in national panel surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 60–87. <https://doi.org/10.1177/0002716212456363>
- Schroeders, U., Schmidt, C., & Gnams, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, 82(1), 29–56. <https://doi.org/10.1177/00131644211004708>
- Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3), 278–295. <https://doi.org/10.1177/0962280210395740>
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://projecteuclid.org/euclid.ss/1294167961>
- Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press.
- Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research*, 49(3), 193–213. <https://doi.org/10.1080/00273171.2014.889593>
- 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>
- Young, A. F., Powers, J. R., & Bell, S. L. (2006). Attrition in longitudinal studies: Who do you lose? *Australian and New Zealand Journal of Public Health*, 30(4), 353–361. <https://doi.org/10.1111/j.1467-842x.2006.tb00849.x>
- Zinn, S., & Gnams, T. (2020). Analyzing nonresponse in longitudinal surveys using Bayesian Additive Regression Trees: A nonparametric event history analysis. *Social Science Computer Review*. Advance online publication. <https://doi.org/10.1177/0894439320928242>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

# **Predicting lifetime suicide attempts in a community sample of adolescents using machine learning algorithms**

Kristin Jankowsky<sup>1</sup>, Diana Steger<sup>2</sup>, Ulrich Schroeders<sup>1</sup>

1: University of Kassel

2: Leibniz-Institut für Bildungsverläufe

Status – accepted

Jankowsky, K., Steger, D., & Schroeders, U. (2023). Predicting lifetime suicide attempts in a community sample of adolescents using machine learning algorithms. *Assessment*, Advance online publication. <https://doi.org/10.1177/10731911231167490>



# Predicting Lifetime Suicide Attempts in a Community Sample of Adolescents Using Machine Learning Algorithms

Assessment

1–17

© The Author(s) 2023



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/10731911231167490

[journals.sagepub.com/home/asm](https://journals.sagepub.com/home/asm)

Kristin Jankowsky<sup>1</sup> , Diana Steger<sup>1</sup> , and Ulrich Schroeders<sup>1</sup> 

## Abstract

Suicide is a major global health concern and a prominent cause of death in adolescents. Previous research on suicide prediction has mainly focused on clinical or adult samples. To prevent suicides at an early stage, however, it is important to screen for risk factors in a community sample of adolescents. We compared the accuracy of logistic regressions, elastic net regressions, and gradient boosting machines in predicting suicide attempts by 17-year-olds in the Millennium Cohort Study ( $N = 7,347$ ), combining a large set of self- and other-reported variables from different categories. Both machine learning algorithms outperformed logistic regressions and achieved similar balanced accuracies (.76 when using data 3 years before the self-reported lifetime suicide attempts and .85 when using data from the same measurement wave). We identified essential variables that should be considered when screening for suicidal behavior. Finally, we discuss the usefulness of complex machine learning models in suicide prediction.

## Keywords

suicide prediction, suicide risk screening, adolescents, machine learning

According to the latest report of the World Health Organization (WHO, 2021) on adolescent mental health, suicide is the fourth leading cause of death in people aged 15 to 29 years worldwide. A recent cross-cultural meta-analysis including 686,672 children and adolescents estimated the lifetime prevalence for suicide attempts to be 6%, and for suicide ideation, 18% (Lim et al., 2019). Even at the early age of 9–10 years, 8.4% of 7,944 children interviewed in the US-based *Adolescent Brain and Cognitive Development* study reported having past or current suicide ideation and 1.3% confirmed attempted suicides (Janiri et al., 2020). Non-fatal self-harm and previous suicide attempts considerably increase the risk of subsequent suicide attempts (Beckman et al., 2018; Geulayov et al., 2019; Iorfino et al., 2020), making it crucial to better understand the factors associated with suicidal behaviors at an early age to prevent enhanced risk of deaths by suicides. However, previous efforts to predict suicide were often unsatisfactory, mainly because effect sizes of individual factors that have been shown to correlate with suicidal behaviors were only small to moderate (Franklin et al., 2017).

To address the multitude of small factors contributing to suicidal behavior, more recent studies argued for the use of advanced modeling techniques in the

prediction of suicidal behaviors and thoughts (e.g., Fox et al., 2019; Huang et al., 2020). Machine learning (ML) algorithms can incorporate many and potentially collinear predictors and, therefore, constitute a useful tool for reflecting or condensing the complex processes including a myriad of factors and multiple phases that lead to suicide. Most previous studies using ML algorithms to predict suicidal behaviors included adult or clinical samples, whereas studies trying to predict suicide risk in non-clinical adolescent samples are rare (e.g., Bernert et al., 2020). However, for assessing the occurrence of and potentially preventing further escalations of adolescents' suicidal behaviors, it would be pivotal to understand which factors are associated with suicidal behavior in the general population of adolescents rather than those in treatment or hospitalized due to self-harm (e.g., Beauchaine et al., 2019). With this study, we try to predict self-reported lifetime suicide attempts in 17-year-old adolescents using a representative community

<sup>1</sup>University of Kassel, Germany

## Corresponding Author:

Kristin Jankowsky, Psychological Assessment, Institute of Psychology, University of Kassel, Holländische Str. 36-38, 34127 Kassel, Germany. Email: [jankowsky@psychologie.uni-kassel.de](mailto:jankowsky@psychologie.uni-kassel.de)

sample from the longitudinal *Millennium Cohort Study UK* (MCS; Joshi & Fitzsimons, 2016). Besides logistic regression as a baseline model, we used ML algorithms, namely elastic net regressions (e.g., Zou & Hastie, 2005) and gradient boosting machines (GBMs; Friedman, 2001) to integrate a large number of variables in our prediction model to account for the multitude of potential risk factors and compare classification accuracy and stability across different time intervals. Furthermore, we examine whether specific variable categories such as mental health, drug use, personality, and so forth are important for the prediction of lifetime suicide attempts, which would render them informative for public screening.

### **Suicide as a Complex Interplay of Various Facilitating Factors and Acute Risk**

In the following, we will discuss aspects impeding an accurate prediction (and potential prevention) of suicidal behaviors in adolescents: (a) the very heterogeneous pool of individual risk factors with small effects, (b) the low prevalence of suicide attempts and deaths by suicide, and (c) the difficulty in exactly pinpointing the timing of suicidal behaviors. Within a comprehensive meta-analysis on risk factors for suicidal thoughts and behaviors by Franklin et al. (2017), previous self-harm has been shown to be the strongest predictor of suicide attempts across all age groups. The authors summarized the research of the last 50 years, stating that individual effects of various predictor variables across 16 categories (including internal and external psychopathology, normative personality, demographics, physical illness or social factors) have been small, and the prediction of suicide attempts was unsatisfactory (with weighted area under the curve [AUC] values of .49–.61). Nevertheless, there is ample research including multiple meta-analyses on so-called warning signs or risk factors of suicidal behaviors. Factors that have been meta-analytically shown to enhance the risk of suicidal behaviors include accumulated childhood adversity (Björkenstam et al., 2017), perfectionism (Smith et al., 2018), sleep problems (Kearns et al., 2020), hopelessness and depression (Ribeiro et al., 2018), anxiety sensitivity (Stanley et al., 2018), as well as mental disorders in general (Too et al., 2019).

Hawton et al. (2012) assigned previously identified risk factors of self-harm and suicide in adolescents into three broad categories: sociodemographic and educational (e.g., sexual orientation or low socioeconomic status), negative life events and family adversity (e.g., parental death, parental mental disorder, or bullying), and psychiatric and psychological (e.g., low self-esteem,

perfectionism, or hopelessness). Similar factors have been reported for suicides of adolescents occurring between April 2019 and April 2020 based on death reviews (i.e., professional mandatory data collection on causes of death of all children younger than 18 years in England), namely household functioning, mental and physical health, loss of or conflict with key relationships, risk-taking behavior, drug misuse, problems with the law, abuse and neglect, bullying, problems at school, social media and internet use, and sexual orientation/gender identity (National Child Mortality Database [NCMD], 2021). The combination of small effects and a low prevalence rate of suicides in children and adolescents (e.g., 1.8 per 100,000 in children aged 9–17 years in 2019 in England; NCMD, 2021) renders modeling suicide risk complicated, often resulting in low positive predictive values, that is, the proportion of true suicides or suicide attempts of all cases that are classified as such (e.g., Belsher et al., 2019). The prediction of suicide is further complicated by the low specificity of potential risk factors such as bullying, so that even among the group of adolescents affected by multiple risk factors, the vast majority will not attempt or die by suicide.

Another aspect worth considering in the prediction of suicide attempts in adolescents is that rates of suicidal behaviors differ across age within childhood and adolescence. For example, in England during the period of April 2019 to April 2020, 46% of all children's and adolescents' suicides occurred in 17-year-olds (compared to 16% in 15- and 16-year-olds and 22% in the group of 14-year-olds and all younger children). An overview of global deaths by suicide showed a similar trend (Naghavi et al., 2019). In 10- to 14-year-olds, the rate was 1.3 per 100,000. This number increases to 8.4 per 100,000 in 15- to 19-year-olds, demonstrating a sharp increase in self-harming and suicidal behaviors in the phase of later adolescence. Research on the developmental course of risk factors of suicide attempts suggests that while there are factors (e.g., previous self-harm, psychological distress) that enhance suicide risk across age groups, there may also be factors that are particularly relevant for specific age groups. For example, Lear et al. (2020) examined whether risk factors for suicide ideation and suicide attempts differed between middle school (aged 11–14 years) and high school students (aged 14–18 years) and found that general psychological distress was associated with suicide attempts in both groups, whereas feeling unsafe at school, lacking family involvement, and community disorganization were only significantly associated with suicide attempts in middle school students. The authors propose a higher dependence on family or other adults in general for middle school students as an explanation for this finding.

## Predicting Suicidal Behaviors Using Machine Learning Algorithms

The usefulness of ML algorithms in suicide research is controversial (e.g., Siddaway et al., 2020). On one hand, there is evidence for an enhanced predictive accuracy by ML models on suicide compared to less complex models such as logistic regressions (Burke et al., 2019). In addition, a recent meta-analysis encourages the notion that ML algorithms outperformed theory-driven predictions of suicide ideation, attempts, and deaths by suicide in longitudinal studies (Schafer et al., 2021). On the other hand, although statistical predictions overall slightly outperform clinical predictions (e.g., Ægisdóttir et al., 2006; Grove et al., 2000), this increased predictive accuracy comes at the cost of lower interpretability. Moreover, the fact that positive predictive values are still rather low, even in models using high-risk, clinical samples (Belsher et al., 2019; Kessler et al., 2019), raises questions about their practical usability. In addition, some researchers emphasized the methodological pitfalls that can occur in estimating complex models including non-linear and interaction effects, possibly leading to biased results (e.g., Jacobucci et al., 2021). In summary, ML algorithms might be useful for regularization in data-driven analyses when there is a plethora of (intertwined) variables, but it is necessary to ensure the generalizability of the ML-based results.

Table 1 provides an overview of five recent studies that use ML algorithms to predict suicide attempts in community samples including adolescents or young adults. Some characteristics of these studies make it difficult to compare and summarize the results. First, the studies operationalize the outcome “suicide attempt” differently in terms of time frame (e.g., last week, last 12 months, or lifetime) or whether attempts are combined with suicidal thoughts. Second, although we solely included studies with adolescent or young adult samples, age varies widely within studies (up to 16 years) and across studies. Third, different methodological approaches regarding model validation, handling of missing data, and handling of unequal group sizes further complicate an unbiased comparison. Bearing these constraints in mind, one main finding across all five studies was the importance of previous suicidal thoughts, suicide attempts, or self-harm when predicting suicidal behaviors, which is in line with recent meta-analyses (e.g., Franklin et al., 2017). Variables reported as important predictors beyond previous self-harm largely depended on the settings of the respective study. For example, van Mens et al. (2020) stated that their variable set was limited insofar as it only comprises psychological risk factors although demographics, lifestyle behaviors, or victimization have been shown to be influential in the other four studies.

Predictive accuracies varied across studies, modeling approaches, and outcome type (e.g., balanced accuracies between .51 and .87), and there was no clear picture as to whether more complex algorithms that allow for non-linear or interaction effects (e.g., random forest or gradient boosting) necessarily lead to more accurate predictions than (regularized) linear regressions. This was often due to a lack of explicit comparison between the respective ML algorithms and simpler approaches. Another methodological shortcoming of four out of five studies presented in Table 1 concerns the lack of independent or unbiased model validation. The most rigorous way of quantifying the predictive accuracy of a model—validation with independent data of an entirely new study (Dwyer et al., 2018)—was not used in any of the studies. In three of the five studies, the full data were split into training and testing data sets, a less optimal but still much-used strategy in ML (e.g., Christodoulou et al., 2019). However, in two of those studies, the authors divided their sample only once. In these cases, ML models tend to fit to the noise in the training data, especially when the features-sample size ratio is high (e.g., Vabalas et al., 2019). Thus, predictive performance and regression coefficients might unduly depend on chance, that is, on which persons were sampled into training or testing data. To minimize this bias and obtain a more robust estimate of predictive accuracy, the procedure needs to be repeated many times.

## The Present Study

In this study, we aim to predict self-reported lifetime suicide attempts by 17-year-old adolescents using data from the longitudinal MCS. We rely on a large set of predictor variables from self-reports and other reports by adolescents and their families covering 14 categories including physical and mental health, drug use, victimization, personality, or future goals (see Table 2). Generally, we use the term *predict* in a statistical fashion, that is, as a statistical abstraction of the relation between a set of variables (predictor variables) and an outcome. We do not imply a causal relationship or a strict temporal order.<sup>1</sup> This study pursues three goals.

First, we investigated to what extent it is possible to predict self-reported lifetime suicide attempts using ML algorithms. In doing so, we compared the predictive accuracy of elastic net regressions and GBM with logistic regressions. The overarching goal of regularized regressions is to avoid overfitting by penalizing too complex models (e.g., Cox et al., 2020). Elastic net regression finds a compromise between least absolute shrinkage and selection operator and ridge regressions to strike a balance between minimizing the sum of squared weights (assigning variables small but non-zero weights) and the

**Table 1.** Overview of Studies Predicting Suicide Using Machine Learning in Adolescents and Young Adults.

Study	Sample	Outcome and prevalence	Design and method	Predictor variables	Predictive accuracy	Important predictors
Burke et al. (2020)	Primary care patients from Pennsylvania ( $N = 13,325$ ) between 14 and 24 years of age ( $M_{\text{age}} = 17.06$ , $SD_{\text{age}} = 2.61$ )	(a) Suicide attempt in the last week ( $n = 39$ , 0.3%) (b) Lifetime suicide attempt ( $n = 608$ , 4.6%)	Cross-sectional, ridge regressions and random forests. 200 imputed data sets split into training (75%) and testing (25%) samples	107 Variables covering demographics, school, family, safety, substance use, nutrition, safety, sexual risk, medical history, depression, anxiety, psychosis, trauma, bullying, gun access	Random forest: .84 (lifetime) .76 (last week) Ridge regressions: .87 (lifetime) .77 (last week)	Active and passive suicidal ideation, suicide planning, and non-suicidal self-injury, physical abuse
Hill et al. (2019)	National (USA) Longitudinal Study of Adolescent to Adult Health ( $N = 4,834$ ), $M_{\text{age}} = 16.15$ , $SD_{\text{age}} = 1.63$ at wave 1	Number of attempted suicides in the past 12 months, coded 0 for 0 and 1 for $> 0$ ( $n = 192$ , 3.97%)	Longitudinal, classification trees, no validation using independent test data	345 Variables covering mental health, victimization, negative life events, family, peer and school functioning, and community engagement	Results of different classification trees are presented, tree 15 with the highest accuracy: .80	Past suicide ideation, depression, mother's education and work, risky behaviors, sexual behavior and sexually transmitted diseases, substance use, expectations about romantic relationships
Macalli et al. (2021)	French i-Share cohort, volunteer student sample ( $N = 5,066$ ), $M_{\text{age}} = 20.7$ , $SD_{\text{age}} = 2.6$	Participants having occasional or frequent suicidal thoughts and/or reported suicide attempts ( $n = 874$ , 17.3%)	Longitudinal, random forests, 10-fold cross-validation, no validation using independent test data	70 Variables covering demographics, lifestyle, family, physical health, substance use, psychiatric disorders, lifetime suicide attempts, suicidal thoughts in the last 12 months, depression, anxiety, self-esteem, perceived stress, and impulsivity	AUC of .84 for girls, .82 for boys	Suicidal thoughts at baseline, self-esteem, trait anxiety, and depression symptoms
van Mens et al. (2020)	Scottish wellbeing study ( $N = 3,508$ ) between 18 and 34 years old ( $M_{\text{age}} = 20.7$ , $SD_{\text{age}} = 2.6$ )	Participants reported a suicide attempt in the last 12 months ( $n = 50$ , 2.0%)	Longitudinal, multiple machine learning algorithms, dataset split once into training (70%) and testing (30%) samples, 10 $\times$ 10-fold cross-validation	211 Items assessing psychological risk factors such as depression, stress, wellbeing, defeat, entrapment, social support, interpersonal needs, goal activation, optimism, resilience, acquired capability, impulsivity, death-related mental imagery, and history of suicidal ideation and suicide attempts	Logistic regression: .51, k-nearest neighbors: .64, classification tree: .71, random forest: .66, gradient boosting: .69, support vector machine: .65	Acquired capability, defeat, depressive symptoms, and a history of suicide attempts
Van Vuuren et al. (2021)	Dutch community sample ( $N = 8,998$ ) Students in the 2nd and 4th year of secondary education (13- to 14-year-olds and 15- to 16-year-olds)	Ask Suicide-Screening Questionnaire—Revised, coded as 1 if any of 4 questions (on recent suicide ideation and lifetime attempts) were affirmed ( $n = 732$ , 8.3%)	Longitudinal, comparison of random forest and LASSO algorithms to a decision rule that classifies every student as "at risk" who affirmed suicidal behaviors at baseline, dataset split once into training (70%) and testing (30%) samples, 10 $\times$ 10-fold cross-validation	Demographics, lifestyle behaviors, physical and mental health, (un)safe environment, and whether made use of information or help on the provided websites or from a school nurse	Decision rule: .64 Random forest: .65 LASSO: .68	Suicide-screening score at baseline, nutrition, drug abuse

Note. All samples are community samples. For Burke et al. (2020), we only present the results of primary care patients (not emergency care patients). Unless otherwise stated, accuracy values are balanced accuracies. If balanced accuracy values were not reported within a study, we calculated them (mean of specificity and sensitivity). For studies with multiple outcomes, we present accuracy values referring to the outcome that includes suicide attempts. Important predictors are presented for the most accurate model predicting suicide attempts.  
AUC = area under the curve; LASSO = least absolute shrinkage and selection operator.



**Table 2.** Predictor Categories With Example Items.

Category	Topics of example items	Number of predictors	
		Model 1/2	Model 3
Activities	Going to the cinema	24	31
Attitudes	Attitudes concerning gender equality	12	19
Behavior	Misbehaving in lessons	22	18
Demographics	Current legal marital status	37	39
Drug use	Smoking e-cigarettes	16	22
Emotion & motivation	Hating oneself	34	40
Family	Parent working long hours	50	48
Future goals	Estimated likelihood of attending university	3	12
Mental health	Currently treated for depression or anxiety	16	38
Offenses (illegal)	Ever been arrested	21	27
Personality	Conscientiousness	39	83
Physical health	Having diabetes	59	58
Sexuality	Having had sexual intercourse with another young person	17	20
Victimization	Being hurt or picked on by other children	7	17

sum of absolute weights (leading to models with many variables given weights of zero). In contrast to logistic regressions and elastic net regressions, GBM algorithms allow for the integration of non-linear and interaction effects without making a priori assumptions on specific functions between predictor variables and the outcome (James et al., 2017; Schroeders et al., 2022). When predicting suicidal behaviors, many potential moderators are conceivable. King et al. (2014), for example, advocated the necessity of compiling and validating suicide screenings separated by gender because they only found a relation for suicide ideation and subsequent suicide attempts within a year after hospitalization for girls. In addition, initial findings from the MCS suggested that there are inequalities in the prevalence of psychological distress and suicidal behaviors across gender, ethnicity, sexual orientation, and socioeconomic status for 17-year-old adolescents (Patalay & Fitzsimons, 2020). In this report, differences between gender groups and sexual orientation were highlighted: More than twice as many girls than boys confirmed a previous suicide attempt (10.6% vs. 4.3%), and for LGB+ adolescents, the rate was even higher (21.7%). It, therefore, seems worthwhile to incorporate these and other potential moderating effects into the prediction of suicidal behaviors in adolescents.

Second, we also investigated if it is possible to predict lifetime suicide attempts 3 years before they were reported. Although there was no information available on the exact timing of the lifetime suicide attempts, given the prevalences described above, we assume that the large majority of suicide attempts occurred after the adolescents turned 14 years of age. For the prediction of future suicide attempts based on electronic health records, Walsh et al. (2017) found that accuracy

improved as the assessment of predictors and suicide attempts was closer in time. It is likely that this finding generalizes to other longitudinal studies such as the MCS. Thus, we also examined the extent to which predictions get more accurate when predictors and outcome were assessed at the same measurement wave as compared to 3 years ahead of time.

Third, we provided an overview of the relative variable importances across models and discussed potential patterns or groupings of predictors of suicide attempts that could inform future theory-building or systematic screenings. We also scrutinized if variable importances shift between models using different variable sets and time frames. Using a longitudinal data set, we avoided typical problems of cross-sectional data, that is, potential changes in the variable importance would indicate that models on suicide need to account for the specific developmental stages of children and adolescents.

This study was not preregistered. All analyses are exploratory; we had no specific prior hypotheses regarding the three aforementioned goals or research questions apart from the respective rationales we presented (see also Wagenmakers et al., 2012).

## Method

### Sample

The MCS UK (Joshi & Fitzsimons, 2016) is a longitudinal cohort study comprising 18,818 children born in the United Kingdom in 2000–2001. Participants were sampled from clusters of electoral wards, with ethnic minorities being disproportionally stratified. We used data from the sixth (conducted in 2015,  $n = 11,872$ ) and seventh (conducted in 2018;  $n = 10,757$ ) measurement

waves in which the participants were 14 and 17 years old, respectively. Both measurement waves have been approved by research ethics committees (ref 13/LO/1786 and 17/NE/0341). We used all cases for which the adolescents' interview at the seventh measurement wave was available ( $n = 10,345$ ), included only one child per family ( $n = 10,238$ ), and discarded all cases in which the outcome variable ("Have you ever hurt yourself on purpose in an attempt to end your life?") was missing (resulting in  $n = 9,723$ ). Finally, we only included adolescents whose self-reports and other reports (in 99.4% reported by one of the parents) were available in both Waves 6 and 7 to avoid issues of attrition (Jankowsky & Schroeders, 2022). This resulted in a total of  $n = 7,347$  participants. To check whether there were systematic differences between these 7,347 and the excluded cases, we compared the correlations of all variables in the adolescents' interview to the outcome across both disjoint samples. The averaged absolute difference between person correlation coefficients was .03 ( $SD = .02$ ). Absolute differences ranged from 0 to .31 with a median of .02. About 83% of the correlation differences across samples were smaller than .05%, and 99% were smaller than .10, which is often used as a rough upper limit for small effect sizes (Gignac & Szodorai, 2016).

The gender ratio of this subsample was balanced; 3,571 (48.6%) of the adolescents were male. Families with higher incomes were overrepresented within the sample: In the sixth measurement wave, 10.6% of the families were categorized in the lowest income quantile, 13.9% in the second-lowest, 19.3% in the middle, 26.4% in the second-highest, and 29.7% in the highest. The majority of adolescents (86.3%) were White, 1.0% were of mixed race, 2.6% were Indian, 3.8% were Pakistani, 1.4% were Bangladeshi, 1.0% were Black Caribbean, 1.7% were Black African, and the remaining 1.9% were summarized into an "other" category. For a more detailed overview on participants' demography, please see Supplemental Table S1.

## Measures

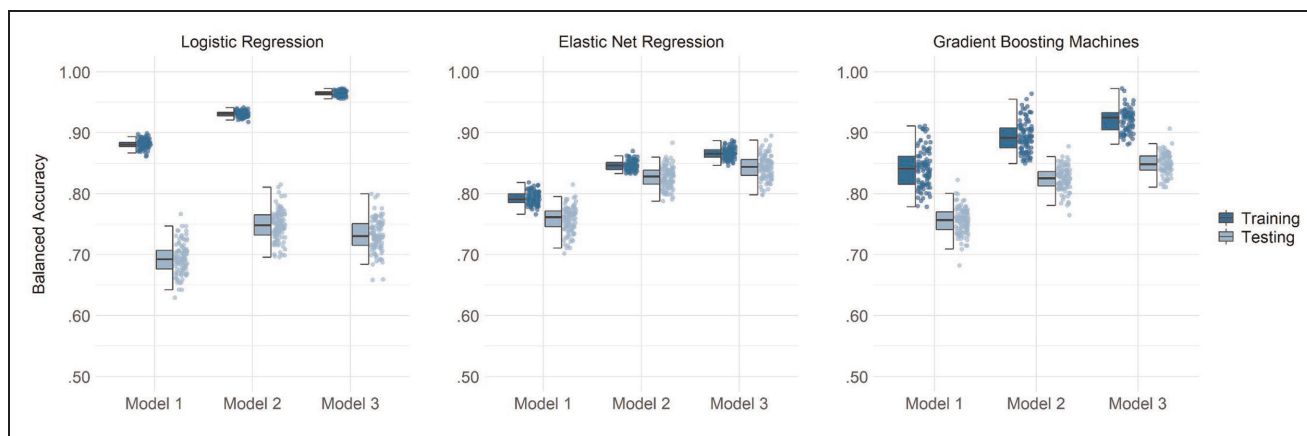
We selected a broad range of variables to predict self-reported lifetime suicide attempts (overall  $N = 638$ ). For an overview, we present the categorization scheme used to classify the predictor variables, together with some example items and the overall number of predictors in each category (Table 2). We used the raw data at item level wherever available to fully capture any potential item effects in suicide prediction (e.g., McClure et al., 2021). We dummy-coded all categorical variables before the analysis using the first category as reference. The outcome measure we used in the present analysis was a single variable of the seventh wave assessing

lifetime suicide attempts ("Have you ever hurt yourself on purpose in an attempt to end your life?") coded as 0 for *no* and 1 for *yes*.

## Statistical Analyses

We compared three different models for the prediction of lifetime suicide attempts: The first included 357 variables from the sixth survey wave of the MCS in which adolescents were 14 years old. Of these variables, 49.02% were answered by the adolescents' parents, of which 42.86% were other reports about their children. In the second model, we updated wherever possible the information of the seventh survey wave in which the adolescents were 17 years old (i.e., for 153 of the 357 variables of the first model or 42.86%). In case a variable was not surveyed again, the original values of the sixth wave were taken. In the seventh survey wave, some variables previously answered by parents were answered by the adolescents so that the share of variables answered by parents decreased to 43.31%, out of which 35.57% were other reports. The third model comprised the same variables as the second model supplemented with 139 variables that were only available within the seventh wave (which were all self-reports by adolescents, decreasing the share of variables answered by parents within the third model to overall 31.57%). These newly available variables predominantly fall into the categories future goals, mental health, personality, and victimization (see also Table 2). By including both the second and the third models in our analyses, we were able to disentangle the effects of time (i.e., more current information likely being more predictive assuming the respective suicide attempts were recent) and content (i.e., incremental validity of newly added variables). All analyses were conducted using the R package *caret* (Kuhn, 2008) as a wrapper interface for modeling and prediction. We compared the predictive accuracy of logistic regressions, elastic net regressions (using the package *glmnet*; Friedman et al., 2001), and GBM (using the package *gbm*; Greenwell et al., 2019). All supplemental materials including analyses scripts, supplemental figures, and a list of all categorized variables are available at <https://osf.io/bycvd/>. The data reported in this manuscript are publicly available data from the MCS and can be accessed after registration with the UK Data Service.

For an unbiased model evaluation, we split the full data into a training data set (80%) and an independent testing data set (20%). Missing values were imputed separately for the training and testing data sets (i.e., after the 80/20 split) using the  $k$ -nearest neighbors algorithm implemented in *caret*. For most of the variables (about 80%), the amount of missingness was small ( $< 5\%$ ), and the average missingness was 4% across all variable sets.



**Figure 1.** Balanced Accuracies for Predicting Suicide Attempts in Adolescents

Note. The boxplot reflects the interquartile range, the solid line represents the median, and the whiskers 1.5-times the interquartile range of 100 iterations. Balanced accuracies are displayed as jittered distribution on the right.

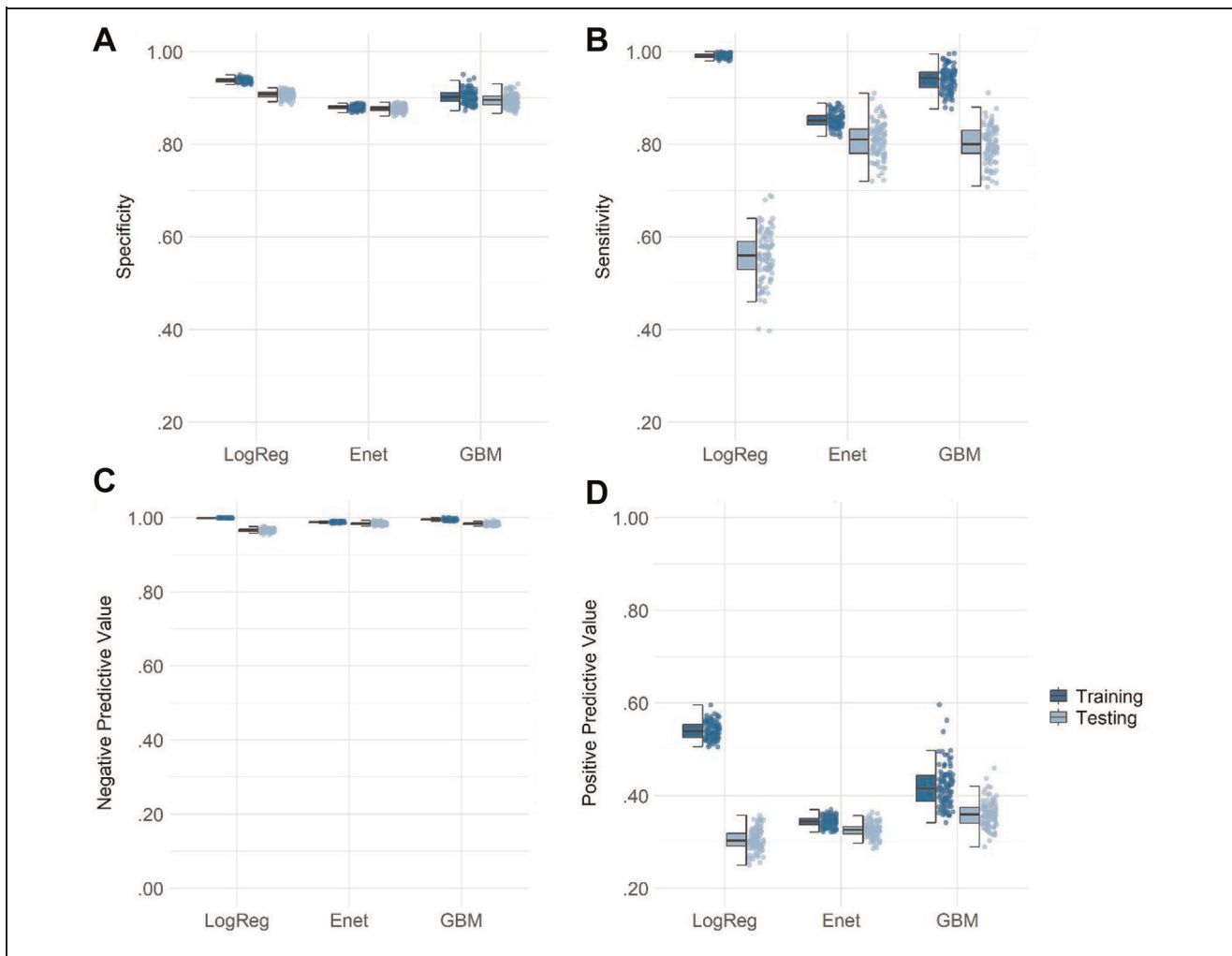
For model training, we used 10-fold cross-validation with upsampling, which means that persons from the minority group (i.e., suicide attempters) were upsampled to match the size of the non-attempters in the training data set. Upsampling is a common and robust procedure often used to handle imbalanced data sets (e.g., García et al., 2012). Parameters for the elastic net regressions (the shrinkage parameter  $\lambda$  and the penalty parameter  $\alpha$ ; Zou & Hastie, 2005) were tuned using a tuning length of 21. GBM are tree-based ML algorithms that sequentially combine multiple decision trees, also called “weak learner,” into an ensemble. Every new tree aims at fitting the residual error of the previous one, leading to a potentially better predictive performance. However, they also run the risk of overfitting, which should be counteracted with sensible hyperparameter tuning (e.g., McNamara et al., 2022). For the gbm tuning parameters, we used the following settings: interaction depth of 1, 2, 3, or 4; a minimum leaf size of 5, 10, 20, or 50; a sequence of shrinkage values between .051 and .201 using steps of .01; and five different numbers of trees (50, 100, 150, 300, and 500).

To evaluate the classification into adolescents who ever attempted suicide vs. adolescents who never attempted suicide, we report the balanced accuracy (the mean of sensitivity and specificity), sensitivity, specificity, and the positive predictive value. In our analyses, sensitivity represents the ratio of correctly identified attempters to all attempters. Specificity represents the ratio of correctly identified non-attempters to all non-attempters, and the positive predictive value represents the proportion of true attempters out of all adolescents who were flagged as attempters. All indices were calculated for each testing data set across 100 iterations of splitting the data into training and testing data.

## Results

Overall, 502 of the 7,347 seventeen-year-old participants (6.83%) indicated that they had hurt themselves on purpose in an attempt to end their life at some point in their life. In the following paragraphs, we first present how accurately this outcome could be predicted by the three different models described above. In a subsequent step, we examine which variables predict these self-reported lifetime suicide attempts and whether the set of the most predictive variables varied across models.

Figure 1 shows the balanced accuracies of 100 iterations of logistic regressions, elastic net regressions, and GBM for three models. The first one used data from the sixth wave of the MCS including 14-year-olds (Model 1: 14 years). In the second model (Model 2: 14 years updated), all variables of the first model were updated if the information was available in the seventh wave, otherwise the original variable was kept in the data set. Finally, the third model (Model 3: 17 years) used variables of the seventh wave that were not available in the previous assessment in addition to the variables of Model 2. Considering only the predictive accuracy within the testing data sets across all models, elastic net regressions and GBM models achieved similar averaged balanced accuracies (ABAs) (.76 and .76 for Model 1, .83 and .82 for Model 2, and .84 and .85 for Model 3, respectively), whereas the predictions with logistic regressions were clearly less accurate (ABA of .69 for Model 1, .75 for Model 2, and .73 for Model 3). This can be explained by the fact that, even with a relatively large sample (with a testing sample of 1,470 adolescents), the logistic regressions were highly overfitted with differences of .19, .18, and .23 in ABA between training and testing (see Figure 1, panel A). Overfit was



**Figure 2.** Specificity, Sensitivity, and Negative and Positive Predictive Values for Predicting Suicide Attempts in Adolescents in Model 3  
*Note.* LogReg = logistic regression, Enet = elastic net regression, GBM = gradient boosting machines. The boxplot reflects the interquartile range, the solid line represents the median, and the whiskers 1.5-times the interquartile range. Specificities (A), sensitivities (B), negative predictive values (C), and positive predictive values (D) are displayed as jittered distribution on the right.

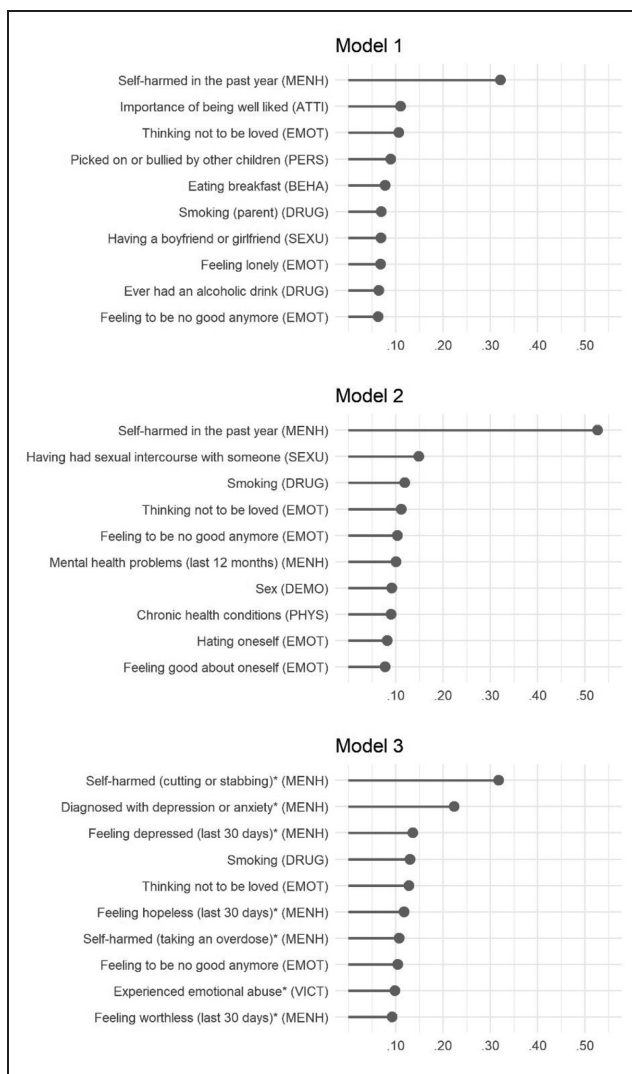
less pronounced for the GBM models (differences of .07 for all models; Figure 1, panel C) and smallest for the elastic net regressions (differences of .03, .02, and .03; Figure 1, panel B), showing that both ML algorithms efficiently used regularization to handle the large number of predictor variables.

Irrespective of the specific algorithm, the above-described ABAs also show that predictions were more accurate for models using variables that were assessed at the same time as the self-report information on lifetime suicide attempts than for models that relied on data from the 14-year-olds only. There was only a negligible difference in the averaged predictive accuracy between Model 2 and Model 3 which both used data from the 17-year-olds. Temporal proximal predictors thus led to higher accuracies than distal ones (again, assuming that

the lifetime suicide attempts were more closer in time to the second measurement wave we used in this study). Adding variables that were only available in the assessment of the 17-year-olds—predominantly variables about future goals, mental health, personality, and victimization—had overall little incremental predictive value.

Figure 2 shows a comparison of the specificities, sensitivities, negative predictive values, and positive predictive values across the three modeling algorithms for Model 3 (for a similar overview of Models 1 and 2, see Figure S1 and S2). Averaged specificities in the testing data set (.91 for logistic regression, .88 for elastic net regression, and .90 for GBM) were generally higher than sensitivities (.56 for logistic regression, .81 for elastic net regressions, and .80 for GBM), meaning that the group





**Figure 3.** Averaged Standardized Regression Coefficients as Indicators of Variable Importance

Note. The variable’s category is given in parentheses following the respective label. MENH = mental health; EMOT = emotion and motivation; ATTI = attitudes; PERS = personality; BEHA = behavior; DRUG = drug use; DEMO = demographics; SEXU = sexuality; PHYS = physical health; VICT = victimization. The asterisk (\*) denotes newly added variables in Model 3.

of non-attempters could be detected more accurately than those who did report suicide attempts irrespective of the modeling approach. Although there were minor differences in specificity between algorithms, sensitivity was much lower for the logistic regression. Averaged negative predictive values were high irrespective of modeling algorithm (.97 for logistic regression, .97 for elastic net regressions, and .98 for GBM), indicating that nearly all who were classified as not at risk were correctly identified as such. The average positive predictive value or precision was lowest for the logistic regressions (.30),

higher for the elastic net regressions (.33), and highest for GBM (.36), indicating that, even in the best model, only about one third of all flagged respondents is correctly identified as suicidal (i.e., the flagging was false in two thirds of the cases).

To sum up, using all available variables for 17-year-olds, it was possible to detect over 80% of adolescents who ever attempted suicide. Moreover, there were no significant differences in predictive accuracy between the elastic net models and GBM models in our study, contradicting earlier findings in suicide prediction. Because the training model of the elastic net regressions showed the least overfit (or in other words, is likely most transferable to unseen data) and tree-based ensemble models tend to be less straightforward to interpret (Cox et al., 2020), in the following paragraphs, we will focus on the variable importance of the elastic net regressions. However, we will also present variable importance of the GBM models for comparison and as a robustness check of our results.

### Important Variables in Predicting Suicide Attempts

In Figure 3, we show the 10 largest averaged standardized regression coefficients of the different elastic net regressions to indicate their importance in the prediction (for the 50 most important predictor variables per model, see Figure S3–S5). Because the differences between the effects of individual variables were often small, the rank order should not be given too much weight. Nonetheless, we provide a short overview of which categories were most predictive. In the following paragraphs, we will summarize three major points: First, the most important variables across the three models by far were indicators of previous self-harm. The question arises as to what extent algorithms that can incorporate several hundreds of variables have incremental value over a simple decision rule that classifies every adolescent who ever showed previous self-harming behavior as “at risk” (e.g., Van Vuuren et al., 2021). Using such a single-item decision rule (i.e., classifying every 17-year-old who confirmed previous self-harm at 14 years of age as “at risk”), balanced accuracy was only slightly lower than that in the first model (.74 vs. .76), but sensitivity was substantially lower (.59 vs. .69). Thus, in our study, such a simplified decision rule would be less sensitive than using ML algorithms with all information available.

Second, if we categorize the most important predictors across all models, these were (in descending order), mental health, emotion and motivation, drug use, sexuality, demography, victimization, physical health, personality, attitudes, and behavior. Out of the category emotion and motivation, “thinking not to be loved” and

“feeling to be no good anymore” as indicators for loneliness and low self-esteem were among the 10 most important predictors irrespective of model. Of the 30 most important predictors (10 per model), all but two variables of the first model (“smoking” and “bullying”) were self-reports by adolescents. Third, for some variables, a shift in importances across the three models can be detected, that is, different categories of variables were specifically important for the prediction of lifetime suicide attempts at a specific point in time. For example, in Model 3, variables of the categories mental health and emotion and motivation were among the 10 most important predictors, while these variables were not as important at an earlier developmental stage. This shift can probably be explained by the more fine-grained and reliable assessment of self-harm (cutting or stabbing and taking an overdose were ranked first and third) and the inclusion of the six items of the Kessler Psychological Distress Scale (Kessler et al., 2002) in the assessment of the 17-year-olds (see the variables asking about feelings within the last 30 days in Figure 3). Looking at the 50 most important variables in each model (for an overview, see Supplemental Figures S3–S5), some variables of the categories sexuality, drug use, illegal offenses, or victimization also tended to gain significance across adolescence. It should be noted that this is also due to the fact that the prevalences of these behaviors significantly increase within the given age range (e.g., sexual intercourse with a peer).

Supplemental Figures S6–S9 show the 10 or 50 most important variables for the GBM models. Overall, the overlap in important variables was large across modeling approaches: Out of the 10 most important variables of the elastic net models, 60% (Model 1), 70% (Model 2), and 90% (Model 3) were also among the 10 most important variables of the GBM models, and all those 10 (for all models) were among the 50 most important variables of the GBM models. Regarding the 50 most important variables, overlap between elastic net and GBM models was at 68% (Model 1), 80% (Model 2), and 70% (Model 3), with more deviations for lesser important variables (which can be expected due to the very small differences in effects and, thus, unstable rank order among the lesser important variables). All in all, the conclusions we derived of the elastic net results about important variable categories equally apply to the GBM models.

## Discussion

Every death by suicide in adolescence is a tragedy. Being able to accurately model and better understand suicidal behaviors in adolescents is literally a vital goal as factors relevant in predicting lifetime suicide attempts of 17-

year-olds could also be relevant for preventive screening tools when implemented at earlier ages. Beauchaine et al. (2019, p. 643) argued for interventions to take place even in childhood since “early starters exhibit greater frequency of non-suicidal self-injury, use more diverse and dangerous methods, and are hospitalized more often than later starters.” Screenings with clear-cut decision rules such as the Oxford Mental Illness and Suicide Tool (Fazel et al., 2019) have been specifically developed for adult and clinical samples. However, they mainly address samples with severe mental disorders (e.g., schizophrenia spectrum or bipolar disorder), and they usually require some clinical expertise. In this study, we tried to narrow down the range of potentially relevant variables in predicting lifetime suicide attempts in 17-year-olds in the United Kingdom using self-reports and other reports as typically administered in large-scale panel studies. Generally, the predictive accuracy in the current investigation was higher than that in most other similar studies analyzing adolescent community samples (see Table 1 for a point of comparison). Overall, we evidenced that it is possible to accurately model lifetime suicide attempts using data from longitudinal (household) panels although the variables were not specifically included for this purpose (i.e., for assessing constructs known or hypothesized to affect suicidal behaviors, such as in van Mens et al., 2020). Results of studies like ours can provide valuable information on what variables might enhance lifetime suicide attempts screenings for adolescents.

In any classification task, researchers have to decide whether to consider sensitivity and specificity equally—as we did in the present study—or to optimize one metric at the cost of the other. Regarding suicide attempts by adolescents, undoubtedly the false negative rate (= miss rate) should be as low as possible, which is equal to having a high sensitivity. Overall, we found that non-attempters could be predicted more accurately than attempters, that is, specificity was higher than sensitivity for all models. Adding new variables from the assessment of the 17-year-olds (Model 2 vs. Model 3) did not significantly change sensitivity but led to higher specificity, thus reducing the number of false alarms. Whether false alarms are problematic depends on the consequences that will be drawn from the modeling. For instance, low-cost brief-contact interventions or information materials could easily be provided broadly to adolescents with a high risk score, and false alarms might pose less of an issue in these low-threshold offers. However, it has been meta-analytically shown that brief-contact interventions only slightly reduced the overall number of repeated self-harm incidents per person and not the odds of death by suicide (Milner et al., 2015). In contrast, more extensive (and possibly more effective) interventions such as individual cognitive therapies

(Zalsman et al., 2016) are more expensive and are only available to some individuals. To use personnel and financial resources adequately, it is thus essential to offer help to the most vulnerable adolescents, which is analogous to high specificity and a positive predictive value. On a side note, if researchers were to inform parents of each adolescent with a (very) high risk score, false alarms could also lead to additional burden on adolescents and parents, irritation, or subsequent underreporting of suicidal behaviors (see Kleiman et al., 2019 for a similar argument in real-time monitoring).

### *The How and the When of Suicide Screening*

In the present analyses, our prediction could not draw upon data of a strict temporal order. Rather we modeled any adolescents' suicide attempts (depending on the model either fully or partly past) and examined variable categories that were especially informative, making them promising candidates for the inclusion in screening instruments for overall lifetime suicide risk. Our results are in line with previous research on suicide risk (e.g., Franklin et al., 2017; Hawton et al., 2012), with the most informative variables including previous self-harm, mental and physical health problems, victimization, lack of future goals, drug misuse, atypical or negative sexual experiences (in relation to a specific developmental phase), and psychological constructs covering distress including feelings of hopelessness, loneliness, thwarted belongingness, and low self-esteem. Reassuringly, the predictors that were important in our study are overall similar to those already included in established tools for the assessment of suicide risk in adolescents (such as the Tool for Assessment of Suicide Risk Adolescent Version Modified; Kutcher, 2013), but our results also offer some pointers as to how screenings for suicide risk in adolescents could be improved. Broad screening tools differ from typical suicide risk assessment: Risk assessments in mental health facilities are often filled out by clinicians. Moreover, the evaluation as to whether, for example, an adolescent shows signs of anger or impulsivity is often done with a two- or three-level clinical rating scale. In contrast, we would recommend the inclusion of a more fine-grained and often continuous assessment of risk factors in self-report screening instruments. Furthermore, including (more) open questions about specific self-harming behaviors in broad screening instruments might also add valuable information as we found that different self-harming behaviors were associated with different risks of suicide attempts: Variables such as cutting or stabbing oneself or taking an overdose of pills were more informative than, for example, burning or bruising oneself. In line with this, Beckman et al.

(2018) also showed different odds for subsequent suicide attempts depending on previous self-harming behaviors (i.e., more "violent" methods such as hanging, drowning, or jumping from a height led to higher subsequent suicidal risk). In addition, the frequency of self-harming behaviors should also be assessed as Ammermann et al. (2017) found differences in relations to psychopathology symptomology between individuals that self-harmed more severely (i.e., more often) and those who report five or less self-harming acts, even among individuals who engaged in self-harming behaviors at least once.

The results of the present analyses are also interesting with respect to what was not found: Apart from the mental and physical health of parents, information given by or about members of the adolescents' family played only a negligible role in the prediction of adolescents' lifetime suicide attempts. The information provided through other reports is thus not decisive in modeling lifetime suicide attempts in adolescents, which is in line with DeVille et al. (2020) who showed that caregivers underreport suicidality and self-harming behavior in their children: Eighty-eight percent of the caregivers had no knowledge that their 9- to 10-year-olds reported previous suicide attempts. Thus, although self-reports per se are prone to typical biases such as impression management, they are often a better source of information when it comes to internal or self-evaluative processes, especially in the absence of additional hard facts such as clinical records. Previous studies on adults repeatedly showed that diagnosed mental disorders are prominent risk factors of suicide (e.g., Cavanagh et al., 2003), which was not the case for the present sample of 14-year-olds. A possible explanation is that mental disorders are usually not diagnosed at this young age, a situation which especially holds true for the 14-year-olds and might change in the future with the *Diagnostic and Statistical Manual of Mental Disorders, fifth edition*, and a stronger developmental perspective of mental disorders (Clark et al., 2017). But it also shows that the significance of the predictor sets varies with age. Accordingly, preventions should be tailored to specific phases of development. Although there were many similarities in the variable importances across models, specific variables on sexuality, victimization, offenses, or drug misuse were more important for the 17-year-olds or only available for 17-year-olds. These behaviors have largely different prevalence rates and, therefore, different meanings for 14-year-olds compared to 17-year-olds. Thus, it might be promising to include questions that predominantly concern older adolescents (from a normative view) in the assessment of younger adolescents because they might point to an unusual or problematic behavior.



### ***Model Complexity Does Not Revolutionize Suicide Prediction***

A frequently highlighted advantage of tree-based ML algorithms is the possibility of including non-linear effects or interactions into prediction models without the need for specifying a priori theoretical assumptions about specific predictor variables and their relation to the respective outcome. The idea that these algorithms could lead to an increment in the prediction of suicide behaviors is compelling and has already been put forward by several recent studies (e.g., Fox et al., 2019; Walsh et al., 2017). In our study, both ML algorithms that use regularization (and are thus better equipped to handle overfitting) achieved a higher predictive accuracy than a simple logistic regression with all predictors. Our results, thus, also emphasize that ML algorithms can enhance accuracy in predicting suicide attempts. However, using models that are capable of including non-linear effects or higher-order interactions such as GBM did not result in a more accurate prediction.

The lacking superiority of more complex models might be attributed to different factors: First, we used a model validation approach that strictly separates the training data from an independent holdout testing data, preventing information leakage between both data sets. Jacobucci et al. (2021) convincingly demonstrated that the higher predictive accuracies for suicidal behaviors in tree-based algorithms such as random forests that have previously been reported in the literature were largely based on a specific validation approach called “optimism bootstrapping” which can lead to inflated predictive accuracies. In other words, previous results demonstrating an advantage for more complex models in suicide prediction should be treated cautiously, and presumably the relations between predictor variables and suicide attempts are mostly linear. Second, simulation studies showed a measurement error can impact the ability of tree-based algorithm to accurately depict the non-linear effects or interactions contained in the true model of simulated data. Accordingly, GBM models did not achieve a higher prediction accuracy than linear regressions when the measurement error was high (e.g., Jacobucci & Grimm, 2020; McNamara et al., 2022). Given that we used many “fuzzy” psychological predictors that are affected by measurement error (e.g., indicators of emotions, feeling, or personality), similar issues may have occurred in our data.

At first glance, our results might seem discouraging for researchers aiming to use more-complex ML algorithms to further the prediction of suicidal behaviors. In fact, they only show that more complex models are no silver bullet that will always guarantee enhanced predictive accuracy but that “human” tasks such as trying to

reliably assess constructs of interest, conducting appropriate model validation, or selecting relevant predictors are still important decisions for the researcher.

### ***Limitations***

We would like to discuss two limitations of our study regarding the outcome variable, namely the single self-reported indicator of lifetime suicide attempts (“Have you ever hurt yourself on purpose in an attempt to end your life?”). The first limitation concerns the fact that lifetime suicide attempts were solely assessed for the 17-year-olds without an indication of the exact time point. Thus, it is not possible to rule out that some of these self-reported suicide attempts happened even before the first measurement wave we used for prediction, that is, when participants were younger than 14 years. However, suicide attempts become much more common at later stages of adolescence (e.g., Naghavi et al., 2019). For example, between 2019 and 2020, 78% of all adolescents’ deaths by suicide in England occurred in 15- to 17-year-olds; thus, the majority of attempts in childhood and adulthood will have occurred between the ages of 14 and 17 years. Although Models 2 and 3 (which use information only available after the respective suicide attempts) clearly predict past suicide attempts, we assume that Model 1 (which uses information from the assessment of the 14-year-olds) largely predicts future suicides. In addition, even predicting or rather modeling past suicide attempts in adolescents can be a worthwhile goal on its own and useful for prevention as it is a robust finding that previous self-harming and suicidal behavior is the strongest predictor of subsequent suicidal behaviors and, thus, enhances the risk of eventual deaths by suicide.

The second limitation refers to using a single self-report item. Hom et al. (2016) found that while all participants of their study sample endorsed a previous suicide attempt on a single-item self-report, only 60% had an actual suicide attempt history when a multi-item assessment and an in-person interview were conducted at follow-up. In contrast, there are also studies highlighting that some participants choose to not disclose their suicide history but are at risk (e.g., Podlogar et al., 2016) or underreport previous self-harming behavior (e.g., Khazem et al., 2021). Thus, correct disclosure of sensible personal information in self-reports also impacts the validity of the outcome variable. In addition, if suicidal behaviors are assessed with a single item, the item wording can substantially impact overall endorsement rates. Ammerman et al. (2021) found that item wording impacted endorsement consistency across a range of questions on suicide ideation, planning, and attempts as well as across different time frames. However, asking

about lifetime suicide attempts such in our study was least affected by item wording.

### Potential Future Research Directions

With respect to the aforementioned limitations, it would clearly be desirable to include additional information such as a detailed medical record or the history of previous suicide attempts of an individual into the model. Such information was not available in the present data set but might exist for samples at risk of attempting suicide (e.g., patients with post-traumatic stress disorder, Bryan, 2016, or military personnel, Rozek et al., 2020). In addition, regarding the prediction of the exact time of suicidal behavior, the 3 years of the present study are too wide an interval to achieve the ultimate goal of preventing deaths by suicide on an individual level. However, we argue that suicide prevention in children or adolescents should be understood as a multi-stage process. At a first stage, screening tools for broad community samples are helpful in providing a rough estimate of the overall risk on an interindividual level. The present study hopefully adds to the knowledge about influential predictors in such an assessment. However, a screening tool cannot be used for individual diagnoses at a specific moment in time. Predicting a narrow time frame with an elevated risk of suicidal behavior for an individual requires a different assessment relying on intraindividual data including fluctuating emotional states, interpersonal problems, triggering situations, or access to common means of suicide. The main reason for this is that while there are some relatively robust factors enhancing the risk at general, acute risk of death by suicide is defined by high levels of heterogeneity in individual circumstances, calling for a more personalized modeling of states across shorter time intervals (Kaurin et al., 2022).

Thus, after selecting at-risk participants in community screenings, a closer and more tailored monitoring could be initiated at a second stage. For clinical samples, adolescents have shown high adherence to ecological momentary assessments perceiving them as positive and helpful rather than burdensome (e.g., Glenn et al., 2022). Also in these cases, prediction models that integrate dynamic risk factors of suicidal thoughts in recently discharged suicidal adolescence have evidenced promising results. For example, Czyz et al. (2021) used different combinations of the mean and variance of six risk factors (e.g., hopelessness, connectedness, and psychological pain) assessed via daily diaries for the detection of suicidal crisis. At a 1-month follow-up visit after the discharge of adolescent psychiatric inpatients, they achieved high classification performance ( $AUC = .91$ ). The extent to which those results can be transferred to

adolescents who have not been hospitalized after a recent suicide attempt but have been flagged by broader risk screenings, however, remains an open question that could be addressed in future research.

### Conclusion

The present analyses of longitudinal panel data showed that it is possible to predict lifetime suicide attempts in a community sample of adolescents. The results of such a screening could help to find relevant factors that can be used in an initial step of a two-stage suicide risk assessment to initiate a more fine-grained evaluation or to provide information on how to seek help. Besides previous self-harm, indicators of poor mental health and negative emotions were most indicative for lifetime suicide attempts. Because using more complex ML algorithm did not lead to improvements in predictive performance, reliably assessing relevant longitudinal information seems more promising for the improvement of suicide prediction than using even more complex statistical models. Furthermore, our results indicated shifts of variable importances across different stages of adolescence, suggesting that the assessment should be tailored to the developmental phase.


### Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Kristin Jankowsky  <https://orcid.org/0000-0002-4847-0760>

Diana Steger  <https://orcid.org/0000-0002-5282-6934>

Ulrich Schroeders  <https://orcid.org/0000-0002-5225-1122>

### Supplemental Material

Supplemental material for this article is available online at <https://osf.io/bycvd/>.

### Note

1. Because the information on lifetime suicide attempts and the predictor variables were assessed at the same time point (for the models using data from the 17-year-olds), the modeling identifies the best indicators of past attempts, rather than assessing future behavior.

### References

- Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N.,

- Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341–382. <https://doi.org/10.1177/0011000005285875>
- Ammerman, B. A., Burke, T. A., Jacobucci, R., & McClure, K. (2021). How we ask matters: The impact of question wording in single-item measurement of suicidal thoughts and behaviors. *Preventive Medicine, 152*, Article 106472. <https://doi.org/10.1016/j.ypmed.2021.106472>
- Ammerman, B. A., Jacobucci, R., Kleiman, E. M., Muehlenkamp, J. J., & McCloskey, M. S. (2017). Development and validation of empirically derived frequency criteria for NSSI disorder using exploratory data mining. *Psychological Assessment, 29*(2), 221–231. <https://doi.org/10.1037/pas0000334>
- Beauchaine, T. P., Hinshaw, S. P., & Bridge, J. A. (2019). Non-suicidal self-injury and suicidal behaviors in girls: The case for targeted prevention in preadolescence. *Clinical Psychological Science, 7*(4), 643–667. <https://doi.org/10.1177/2167702618818474>
- Beckman, K., Mittendorfer-Rutz, E., Waern, M., Larsson, H., Runeson, B., & Dahlin, M. (2018). Method of self-harm in adolescents and young adults and risk of subsequent suicide. *Journal of Child Psychology and Psychiatry, 59*(9), 948–956. <https://doi.org/10.1111/jcpp.12883>
- Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., Morgan, R. L., Evatt, D. P., Tucker, J., & Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *Journal of the American Medical Association Psychiatry, 76*(6), 642–651. <https://doi.org/10.1001/jamapsychiatry.2019.0174>
- Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnoui, F. (2020). Artificial intelligence and suicide prevention: A systematic review of machine learning investigations. *International Journal of Environmental Research and Public Health, 17*(16), Article 5929. <https://doi.org/10.3390/ijerph17165929>
- Björkenstam, C., Kosidou, K., & Björkenstam, E. (2017). Childhood adversity and risk of Suicide: Cohort study of 548 721 adolescents and young adults in Sweden. *British Medical Journal, 357*, Article j1334. <https://doi.org/10.1136/bmj.j1334>
- Bryan, C. J. (2016). Treating PTSD within the context of heightened suicide risk. *Current Psychiatry Reports, 18*(8), Article 73. <https://doi.org/10.1007/s11920-016-0708-z>
- Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of Affective Disorders, 245*, 869–884. <https://doi.org/10.1016/j.jad.2018.11.073>
- Burke, T. A., Jacobucci, R., Ammerman, B. A., Alloy, L. B., & Diamond, G. (2020). Using machine learning to classify suicide attempt history among youth in medical care settings. *Journal of Affective Disorders, 268*, 206–214. <https://doi.org/10.1016/j.jad.2020.02.048>
- Cavanagh, J. T., Carson, A. J., Sharpe, M., & Lawrie, S. M. (2003). Psychological autopsy studies of suicide: A systematic review. *Psychological Medicine, 33*(3), 395–405. <https://doi.org/10.1017/s0033291702006943>
- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>
- Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three Approaches to Understanding and Classifying Mental Disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychological Science in the Public Interest, 18*(2), 72–145. <https://doi.org/10.1177/1529100617727266>
- Cox, C. R., Moscardini, E. H., Cohen, A. S., & Tucker, R. P. (2020). Machine learning for suicidology: A practical review of exploratory and hypothesis-driven approaches. *Clinical Psychology Review, 82*, Article 101940. <https://doi.org/10.1016/j.cpr.2020.101940>
- Czyz, E., Koo, H., Al-Dajani, N., King, C., & Nahum-Shani, I. (2021). Predicting short-term suicidal thoughts in adolescents using machine learning: Developing decision tools to identify daily level risk after hospitalization. *Psychological Medicine*. Advance online publication. <https://doi.org/10.1017/S0033291721005006>
- DeVile, D. C., Whalen, D., Breslin, F. J., Morris, A. S., Khalsa, S. S., Paulus, M. P., & Barch, D. M. (2020). Prevalence and family-related factors associated with suicidal ideation, suicide attempts, and self-injury in children aged 9 to 10 years. *Journal of the American Medical Association Network Open, 3*(2), Article e1920956. <https://doi.org/10.1001/jamanetworkopen.2019.20956>
- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*(1), 91–118. <https://doi.org/10.1146/annurev-clinpsy-032816-045037>
- Fazel, S., Wolf, A., Larsson, H., Mallett, S., & Fanshawe, T. R. (2019). The prediction of suicide in severe mental illness: Development and validation of a clinical prediction rule (OxMIS). *Translational Psychiatry, 9*(1), Article 98. <https://doi.org/10.1038/s41398-019-0428-3>
- Fox, K. R., Huang, X., Linthicum, K. P., Wang, S. B., Franklin, J. C., & Ribeiro, J. D. (2019). Model complexity improves the prediction of nonsuicidal self-injury. *Journal of Consulting and Clinical Psychology, 87*(8), 684–692. <https://doi.org/10.1037/ccp0000421>
- Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin, 143*(2), 187–232. <https://doi.org/10.1037/bul0000084>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. <https://doi.org/10.1214/aos/101320345>
- García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with



- different levels of class imbalance. *Knowledge-Based Systems*, 25(1), 13–21. <https://doi.org/10.1016/j.knosys.2011.06.013>
- Geulayov, G., Casey, D., Bale, L., Brand, F., Clements, C., Farooq, B., Kapur, N., Ness, J., Waters, K., Tsiachristas, A., & Hawton, K. (2019). Suicide following presentation to hospital for non-fatal self-harm in the Multicentre Study of Self-harm: A long-term follow-up study. *The Lancet Psychiatry*, 6(12), 1021–1030. [https://doi.org/10.1016/S2215-0366\(19\)30402-X](https://doi.org/10.1016/S2215-0366(19)30402-X)
- Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102, 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>
- Glenn, C. R., Kleiman, E. M., Kearns, J. C., Santee, A. C., Esposito, E. C., Conwell, Y., & Alpert-Gillis, L. J. (2022). Feasibility and acceptability of ecological momentary assessment with high-risk suicidal adolescents following acute psychiatric care. *Journal of Clinical Child and Adolescent Psychology*, 51(1), 32–48. <https://doi.org/10.1080/15374416.2020.1741377>
- Greenwell, B., Boehmke, B., & Cunningham, J., & GBM Developers. (2019). *gbm: Generalized boosted regression models* (Version 2.1.5) [Computer software]. <https://CRAN.R-project.org/package=gbm>
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19–30. <https://doi.org/10.1037/1040-3590.12.1.19>
- Hawton, K., Saunders, K. E., & O'Connor, R. C. (2012). Self-harm and suicide in adolescents. *The Lancet*, 379(9834), 2373–2382. [https://doi.org/10.1016/S0140-6736\(12\)60322-5](https://doi.org/10.1016/S0140-6736(12)60322-5)
- Hill, R. M., Oosterhoff, B., & Do, C. (2019). Using machine learning to identify suicide risk: A classification tree approach to prospectively identify adolescent suicide attempters. *Archives of Suicide Research*, 24(2), 218–235. <https://doi.org/10.1080/13811118.2019.1615018>
- Hom, M. A., Joiner, T. E., & Bernert, R. A. (2016). Limitations of a single-item assessment of suicide attempt history: Implications for standardized suicide risk assessment. *Psychological Assessment*, 28(8), 1026–1030. <https://doi.org/10.1037/pas0000241>
- Huang, X., Ribeiro, J. D., & Franklin, J. C. (2020). The differences between suicide ideators and suicide attempters: Simple, complicated, or complex? *Journal of Consulting and Clinical Psychology*, 88(6), 554–569. <https://doi.org/10.1037/ccp0000498>
- Iorfino, F., Ho, N., Carpenter, J. S., Cross, S. P., Davenport, T. A., Hermens, D. F., Yee, H., Nichles, A., Zmicerevska, N., Guastella, A., Scott, E., & Hickie, I. B. (2020). Predicting self-harm within six months after initial presentation to youth mental health services: A machine learning study. *PLOS ONE*, 15(12), Article e0243467. <https://doi.org/10.1371/journal.pone.0243467>
- Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, 15(3), 809–816. <https://doi.org/10.1177/1745691620902467>
- Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of inflated prediction performance: A commentary on machine learning and suicide research. *Clinical Psychological Science*, 9(1), 129–134. <https://doi.org/10.1177/2167702620954216>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning*. Springer.
- Janiri, D., Doucet, G. E., Pompili, M., Sani, G., Luna, B., Brent, D. A., & Frangou, S. (2020). Risk and protective factors for childhood suicidality: A U.S. population-based study. *The Lancet Psychiatry*, 7(4), 317–326. [https://doi.org/10.1016/s2215-0366\(20\)30049-3](https://doi.org/10.1016/s2215-0366(20)30049-3)
- Jankowsky, K., & Schroeders, U. (2022). Validation and generalizability of machine learning prediction models on attrition in longitudinal studies. *International Journal of Behavioral Development*, 46(2), 169–176. <https://doi.org/10.1177/01650254221075034>
- Joshi, H., & Fitzsimons, E. (2016). The Millennium Cohort Study: The making of a multi-purpose resource for social science and policy. *Longitudinal and Life Course Studies*, 7(4). <https://doi.org/10.14301/lcs.v7i4.410>
- Kaurin, A., Dombrovski, A. Y., Hallquist, M. N., & Wright, A. G. C. (2022). Integrating a functional view on suicide risk into idiographic statistical models. *Behaviour Research and Therapy*, 150, 104012. <https://doi.org/10.1016/j.brat.2021.104012>
- Kearns, J. C., Coppersmith, D., Santee, A. C., Insel, C., Pigeon, W. R., & Glenn, C. R. (2020). Sleep problems and suicide risk in youth: A systematic review, developmental framework, and implications for hospital treatment. *General Hospital Psychiatry*, 63, 141–151. <https://doi.org/10.1016/j.genhosppsych.2018.09.011>
- Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, 32(6), 959–976. <https://doi.org/10.1017/s0033291702006074>
- Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2019). Suicide prediction models: A critical review of recent research with recommendations for the way forward. *Molecular Psychiatry*, 25(1), 168–179. <https://doi.org/10.1038/s41380-019-0531-0>
- Khazem, L. R., Rufino, K. A., Rogers, M. L., Gallyer, A. J., Joiner, T. E., & Anestis, J. C. (2021). Underreporting on the MMPI-2-RF extends to extra-test measures of suicide risk. *Psychological Assessment*, 33(8), 789–794. <https://doi.org/10.1037/pas0001034>
- King, C. A., Jiang, Q., Czyz, E. K., & Kerr, D. C. (2014). Suicidal ideation of psychiatrically hospitalized adolescents has one-year predictive validity for suicide attempts in girls only. *Journal of Abnormal Child Psychology*, 42(3), 467–477. <https://doi.org/10.1007/s10802-013-9794-0>
- Kleiman, E. M., Glenn, C. R., & Liu, R. T. (2019). Real-time monitoring of suicide risk among adolescents: Potential barriers, possible solutions, and future directions. *Journal of Clinical Child and Adolescent Psychology*, 48(6), 934–946. <https://doi.org/10.1080/15374416.2019.1666400>

- Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5). <https://doi.org/10.18637/jss.v028.i05>
- Kutcher, S. (2013). *Tool for Assessment of Suicide Risk Adolescent Version Modified (TASR-Am)*. <https://mentalhealthliteracy.org/product/tool-assessment-suicide-risk-adolescent-version-modified-tasr/>
- Lear, M. K., Perry, K. M., Stacy, S. E., Canen, E. L., Hime, S. J., & Pepper, C. M. (2020). Differential suicide risk factors in rural middle and high school students. *Psychiatry Research*, 284, Article 112773. <https://doi.org/10.1016/j.psychres.2020.112773>
- Lim, K. S., Wong, C. H., McIntyre, R. S., Wang, J., Zhang, Z., Tran, B. X., Tan, W., Ho, C. S., & Ho, R. C. (2019). Global lifetime and 12-month prevalence of suicidal behavior, deliberate self-harm and non-suicidal self-injury in children and adolescents between 1989 and 2018: A meta-analysis. *International Journal of Environmental Research and Public Health*, 16(22), Article 4581. <https://doi.org/10.3390/ijerph16224581>
- Macalli, M., Navarro, M., Orri, M., Tournier, M., Thiébaud, R., Côté, S. M., & Tzourio, C. (2021). A machine learning approach for predicting suicidal thoughts and behaviours among college students. *Scientific Reports*, 11(1), 11363. <https://doi.org/10.1038/s41598-021-90728-z>
- McClure, K., Jacobucci, R., & Ammerman, B. A. (2021). Are items more than indicators? An examination of psychometric homogeneity, item-specific effects, and consequences for structural equation models. *PsyArXiv*. <https://doi.org/10.31234/osf.io/n4mxx>
- McNamara, M. E., Zisser, M., Beevers, C. G., & Shumake, J. (2022). Not just “big” data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions. *Behaviour Research and Therapy*, 153, Article 104086. <https://doi.org/10.1016/j.brat.2022.104086>
- Milner, A., Carter, G., Pirkis, J., Robinson, J., & Spittal, M. (2015). Letters, green cards, telephone calls and postcards: Systematic and meta-analytic review of brief contact interventions for reducing self-harm, suicide attempts and suicide. *British Journal of Psychiatry*, 206(3), 184–190. <https://doi.org/10.1192/bjp.bp.114.147819>
- Naghavi, M., & Global Burden of Disease Self-Harm Collaborators. (2019). Global, regional, and national burden of suicide mortality 1990 to 2016: Systematic analysis for the Global Burden of Disease Study 2016. *British Medical Journal*, 364, Article 94. <https://doi.org/10.1136/bmj.194>
- National Child Mortality Database. (2021). *Suicide in children and young people*. <https://www.ncmd.info/wp-content/uploads/2021/11/NCMD-Suicide-in-Children-and-Young-People-Report.pdf>
- Patalay, P., & Fitzsimons, E. (2020). *Mental ill-health at age 17 in the UK: Prevalence of and inequalities in psychological distress, self-harm and attempted suicide*. Centre for Longitudinal Studies.
- Podlogar, M. C., Rogers, M. L., Chiurliza, B., Hom, M. A., Tzoneva, M., & Joiner, T. (2016). Who are we missing? Non-disclosure in online suicide risk screening questionnaires. *Psychological Assessment*, 28(8), 963–974. <https://doi.org/10.1037/pas0000242>
- Ribeiro, J. D., Huang, X., Fox, K. R., & Franklin, J. C. (2018). Depression and hopelessness as risk factors for suicide ideation, attempts and death: Meta-analysis of longitudinal studies. *The British Journal of Psychiatry*, 212(5), 279–286. <https://doi.org/10.1192/bjp.2018.27>
- Rozek, D. C., Andres, W. C., Smith, N. B., Leifker, F. R., Arne, K., Jennings, G., Dartnell, N., Bryan, C. J., & Rudd, M. D. (2020). Using machine learning to predict suicide attempts in military personnel. *Psychiatry Research*, 294, Article 113515. <https://doi.org/10.1016/j.psychres.2020.113515>
- Schafer, K. M., Kennedy, G., Gallyer, A., & Resnik, P. (2021). A direct comparison of theory-driven and machine learning prediction of suicide: A meta-analysis. *PLOS ONE*, 16(4), Article e0249833. <https://doi.org/10.1371/journal.pone.0249833>
- Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, 82(1), 29–56. <https://doi.org/10.1177/00131644211004708>
- Siddaway, A. P., Quinlivan, L., Kapur, N., O'Connor, R. C., & de Beurs, D. (2020). Cautions, concerns, and future directions for using machine learning in relation to mental health problems and clinical and forensic risks: A brief comment on “Model complexity improves the prediction of nonsuicidal self-injury” (Fox et al., 2019). *Journal of Consulting and Clinical Psychology*, 88(4), 384–387. <https://doi.org/10.1037/ccp0000485>
- Smith, M. M., Sherry, S. B., Chen, S., Saklofske, D. H., Mushquash, C., Flett, G. L., & Hewitt, P. L. (2018). The perniciousness of perfectionism: A meta-analytic review of the perfectionism-suicide relationship. *Journal of Personality*, 86(3), 522–542. <https://doi.org/10.1111/jopy.12333>
- Stanley, I. H., Boffa, J. W., Rogers, M. L., Hom, M. A., Albanese, B. J., Chu, C., Capron, D. W., Schmidt, N. B., & Joiner, T. E. (2018). Anxiety sensitivity and suicidal ideation/suicide risk: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 86(11), 946–960. <https://doi.org/10.1037/ccp0000342>
- Too, L. S., Spittal, M. J., Bugeja, L., Reifels, L., Butterworth, P., & Pirkis, J. (2019). The association between mental disorders and suicide: A systematic review and meta-analysis of record linkage studies. *Journal of Affective Disorders*, 259, 302–313. <https://doi.org/10.1016/j.jad.2019.08.054>
- Vabalas, A., Gowen, E., Poliakov, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11), Article e0224365. <https://doi.org/10.1371/journal.pone.0224365>
- Van Mens, K., de Schepper, C., Wijnen, B., Koldijk, S. J., Schnack, H., de Looft, P., Lokkerbol, J., Wetherall, K., Cleare, S. C., O'Connor, R., & de Beurs, D. (2020). Predicting future suicidal behaviour in young adults, with different machine learning techniques: A population-based longitudinal study. *Journal of Affective Disorders*, 271, 169–177. <https://doi.org/10.1016/j.jad.2020.03.081>
- Van Vuuren, C. L., van Mens, K., de Beurs, D., Lokkerbol, J., van der Wal, M. F., Cuijpers, P., & Chinapaw, M. J. M.



- (2021). Comparing machine learning to a rule-based approach for predicting suicidal behavior among adolescents: Results from a longitudinal population-based survey. *Journal of Affective Disorders*, 295, 1415–1420. <https://doi.org/10.1016/j.jad.2021.09.018>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632–638. <https://doi.org/10.1177/1745691612463078>
- Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, 5(3), 457–469. <https://doi.org/10.1177/2167702617691560>
- World Health Organization. (2021, November 17). *Adolescent mental health*. <https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health>
- Zalsman, G., Hawton, K., Wasserman, D., van Heeringen, K., Arensman, E., Sarchiapone, M., Carli, V., Höschl, C., Barzilay, R., Balazs, J., Purebl, G., Kahn, J. P., Sáiz, P. A., Lipsicas, C. B., Bobes, J., Cozman, D., Hegerl, U., & Zohar, J. (2016). Suicide prevention strategies revisited: 10-year systematic review. *The Lancet Psychiatry*, 3(7), 646–659. [https://doi.org/10.1016/S2215-0366\(16\)30030-X](https://doi.org/10.1016/S2215-0366(16)30030-X)
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

# **First impressions count: Therapists’ impression on patients’ motivation and helping alliance predicts psychotherapy dropout**

Kristin Jankowsky<sup>1</sup>, Zimmermann, J.<sup>1</sup>, Jaeger, U.<sup>2</sup>, Mestel, R.<sup>3</sup>, & Schroeders, U.<sup>1</sup>

1: University of Kassel

2: Asklepios Clinic Tiefenbrunn

3: Vamed Clinic, Bad Grönenbach

Status – submitted

Jankowsky, K., Zimmermann, J., Jaeger, U., Mestel, R., & Schroeders, U. (2023, September 27). First impressions count: Therapists’ impression on patients’ motivation and helping alliance predicts psychotherapy dropout. Retrieved from [psyarxiv.com/nhs6c](https://psyarxiv.com/nhs6c)

## **Abstract**

### **Objective**

With meta-analytically estimated rates of about 25%, dropout in psychotherapies is a major concern for individuals, clinicians, and the healthcare system at large. To be able to counteract dropout in psychotherapy, accurate insights about its predictors are needed.

### **Method**

We compared logistic regression models with two machine learning algorithms (elastic net regressions and gradient boosting machines) in the prediction of therapy dropout in two large inpatient samples ( $N = 1,691$  and  $N = 12,473$ ) using patient- and therapist-reported variables collected at the time of admission to the clinic.

### **Results**

Predictive accuracies of the two machine learning algorithms were similar and higher than for logistic regressions: Therapy dropout could be predicted with an AUC of .73 and .83 for Sample 1 and 2, respectively. The initial evaluation of patients' motivation and the therapeutic alliance rated by the respective therapist were the most important predictors of dropout.

### **Conclusions**

Therapy dropout in naturalistic inpatient settings can be predicted to a considerable degree by using baseline indicators. Feature selection via regularization leads to higher predictive performances whereas non-linear or interaction effects are dispensable. The most promising point of intervention to reduce therapy dropouts seems to be patients' motivation and the therapeutic alliance.

**Keywords:** therapy dropout; predictive modeling; machine learning; inpatients; helping alliance

**First impressions count: Therapists' impression on patients' motivation and helping alliance predicts psychotherapy dropout**

More than one in four people are affected by mental illness at some point in their life (e.g., Steel et al., 2014). To reduce individual suffering and problematic consequences for society as a whole (e.g., due to a high number of sick leaves), it is particularly important to offer the best possible treatment. To this end, psychotherapy has been shown to be effective (Barth et al., 2013; Kamenov et al., 2017; Leichsenring et al. 2022), under the condition of its regular completion. However, with meta-analytically estimated rates of about 20-25%, dropout (usually defined as the termination of psychotherapy initiated by the patient against the therapist's advice) is a major issue in psychotherapies (Fernandez et al., 2015; Hans & Hiller, 2013). On an individual level, dropout is problematic because patients respond worse to premature discontinued treatments (e.g., Barrett et al., 2008), which in turn might lead to disease chronification. On a societal level, dropout implies that limited health care resources are not optimally spent.

Previous studies found several patient characteristics to be associated with therapy dropout such as clinical variables (e.g., substance abuse, comorbidities, baseline symptom severity), demographics (e.g., lower education, younger age, being male, unemployment), personality traits, and attitudes or motivation toward therapy (Bucher et al., 2019; Fernandez et al., 2015; Karterud et al., 2003; Schmidt et al., 2019; Swift & Greenberg, 2012; Zimmermann et al., 2017). To reduce dropout rates, more recent studies proposed to predict patients' risk of eventual dropout using the oftentimes large amount of patient data usually gathered within baseline documentations of psychotherapies (e.g., Bennemann et al., 2022; Gonzalez Salas Duhne et al., 2022). Complex machine learning algorithms might help to enhance the accuracy of such predictions because they handle multicollinearity between predictor variables and allow to integrate non-linear and (higher order) interaction effects without the need to a priori specify the associations between variables and the respective

outcome (James et al., 2017). Bennemann et al. (2022) recently compared 21 different machine learning algorithms for the prediction of therapy dropout of cognitive-behavioral therapy in an outpatient sample and found that an ensemble of a Random Forest and Nearest Neighbor Model achieved the overall moderate best predictive accuracy with an Area Under the Curve (AUC) of .66. Important predictors were education, age, and subscales of the Personality Style and Disorder Inventory (PSDI; Kuhl & Kazén, 2009) as well as the Brief Symptom Inventory (BSI; Derogatis, 1982). However, outpatients differ from inpatients regarding the frequency of therapy sessions, overall treatment timeframe, and dropout rates (e.g., Fernandez et al., 2015). Hence, it is unclear whether these results can be generalized across treatment settings.

In this study, we aim to answer three main questions on the prediction of therapy dropout in inpatients of two German psychotherapy clinics: First, we will investigate how accurately therapy dropout can be predicted for inpatient samples using variables collected within the standard baseline documentation. Second, we examine which self- or therapist-reported variables are particularly indicative of subsequent therapy dropout. Third, we also study whether complex machine learning algorithms are superior to simpler unregularized logistic regression analyses at predicting therapy dropout. We will compare our results across two inpatient samples and discuss them regarding previous research on outpatients to be able to evaluate the generalizability of predictive accuracies and variable importances across samples and therapy settings.

## **Method**

### **Samples**

We reanalyzed anonymized routine outcome monitoring data assessed at admission from two psychotherapy clinics in [blinded for reviews]. All patients gave written informed consent. We assert that all procedures contributing to this work comply with the Helsinki Declaration of 1975, as revised in 2013. Sample 1 comprised 1,691 patients treated between

2007 and 2011 (62.12% women, age ranging from 17 to 71 years;  $M = 35.49$ ,  $SD = 11.88$ ). Length of stay was between 1 to 252 days ( $M = 79.81$ ,  $SD = 35.75$ ). Most common diagnoses were depressive or major depressive disorder, anxiety or stress-related disorders, and personality disorders. Patients suffering from acute psychosis, acute suicidality, dementia or withdrawal symptoms were generally not admitted. Treatment was tailored to the individuals, but psychoanalytical-interactional individual and group therapies were central elements of the clinic's treatment concept (Leichsenring et al., 2016; Streeck & Leichsenring, 2010). Patients received on average two individual therapy sessions and three group therapy sessions per week.

Sample 2 comprised 12,473 patients treated between 1995 and 2010 (72.18% women, age ranging from 17 to 80 years;  $M = 38.73$ ,  $SD = 11.04$ ). Length of stay was between 1 to 224 days ( $M = 55.27$ ,  $SD = 23.61$ ). The clinic mainly focused on depression and anxiety disorders, personality disorders (especially borderline personality disorders), psychosomatic diseases, eating disorders, and substance use disorders. Patients suffering from severe internal or brain-organic diseases or with acute psychosis were not admitted. The clinic offered multimodal treatment (most prominent psychodynamic and humanistic therapies) including individual and group therapy. For a more detailed overview of patient characteristics, please see Table S1.

## **Measures**

### ***Outcome Variable***

For our analyses, we defined all patients within the category “regular completion” as patients who successfully completed their therapy and all cases in which the premature termination was initiated by the patient as dropouts. By only using patients of these categories, the resulting sample sizes were  $n = 1,691$  and  $n = 12,473$  for Sample 1 and 2, respectively (for the different reasons of discharge and their prevalence, see Table S2). Thus,

dropout was used as a dichotomous outcome: Patients either completed their therapy regularly (coded as 0) or patients prematurely terminated the therapy (coded as 1).

### ***Predictor Variables***

We used a broad range of variables to predict therapy dropout: 168 for Sample 1 and 147 for Sample 2 after dummy-coding. All variables were assessed at baseline, that is, on the day of the initial interview within a short time after admission. Predictor variables included demographics (e.g., sex, age, marital status, number of children, education, employment, income), satisfaction with different areas of life (e.g., with relationships or financial situation), information on previous treatments and impairment (e.g., diagnoses, type of admission, suicidality, traumatization, number of previous stays at similar clinics), and scores from typical measures often assessed at the beginning of psychotherapies to enable progress evaluation such as the Symptom Checklist-90-R (Franke, 1995) or the Inventory of Interpersonal Problems (Alden et al., 1990). Most constructs were available for both samples albeit operationalized in a slightly different way. For a detailed overview, we present all predictor variables and descriptive statistics including information on the amount of missingness for the full datasets in Table S1.

### **Statistical Analyses**

For all models, we employed a nested cross-validation approach combining an outer and inner validation loop (e.g., Pargent et al., 2022) to avoid any information leakage between the training and the testing data. In each iteration of the outer validation loop, the full data were split into training data (80% of the full data set) and testing data (the remaining 20%). In the inner loop, the models were trained using 10-fold cross-validation with separately imputed training datasets. We relied on up-sampling to address imbalanced datasets, where persons from the minority group (i.e., dropouts) were up-sampled to match the size of completers (e.g., Jacobucci & Li, 2022). Finally, we assessed the predictive performance of these models using the independent testing data. We performed 300 iterations of the outer validation loop

and averaged the results to obtain a robust estimate of the expected prediction performance when presented with unseen testing data.

All our analyses were conducted using the R package *caret* (Kuhn, 2008) as an interface for prediction and model evaluation. We evaluated three incremental complex modeling approaches to predict therapy dropout. We started with a logistic regression model as a baseline and then compared it to logistic elastic net regressions (Enet; using the R package *glmnet*; Friedman et al., 2010), and gradient boosting machines (GBM; using the package *gbm*; Greenwell et al., 2019). Elastic net regressions are regularized regressions that balance between ridge and least absolute shrinkage and selection operator (LASSO) regressions to create parsimonious models and to maximize predictive performance (e.g., Zou & Hastie, 2005). Gradient boosting machines (GBM) are tree-based algorithms that take into account nonlinear and higher-order interaction effects into the modeling process without requiring any prior specifications of the functions between predictor variables and the outcome (James et al., 2017). Annotated syntax for all analyses including the specific tuning parameter grids of the machine learning algorithms is available in an open project repository at [https://osf.io/dr54h/?view\\_only=7f90ef2200e3402a8e9f6022633f2825](https://osf.io/dr54h/?view_only=7f90ef2200e3402a8e9f6022633f2825).

### **Model Evaluation**

To evaluate the classification performance, we calculated several metrics that were averaged across 300 iterations of the outer loop. In more detail, we calculated the balanced accuracy (average of sensitivity and specificity), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), and the Area Under Curve (AUC). For the best performing models, we also show ROC curves for each testing split and an averaged ROC curve across all 300 testing splits. Additionally, we report variable importance measures scaled from 0 to 100 using the *varImp* function. For the elastic net regressions, these variable importances represent a transformation of the respective absolute regressions' coefficients of



the tuned training model. For the GBM, they represent the relative influence of a predictor averaged across all generated trees (Friedman, 2001).

### Results

A total of 131 out of the 1,691 patients (7.8 %, Sample 1) and 640 out of 12,473 (5.1 %, Sample 2) patients, respectively, terminated their therapy prematurely and against their therapists' recommendation. For the patients who drop out, median length of treatment was 30 days (Sample 1) and 16 days (Sample 2) in comparison to 85 days (Sample 1) and 56 days (Sample 2) for patients who regularly completed their clinic stay.

In Table 1, we summarize common metrics of predictive performance across 300 iterations of logistic regressions, elastic net regressions, and GBM for both samples. Considering only the averaged balanced accuracy (ABA) within the testing datasets in Sample 1, elastic net regressions and GBM achieved identical values (ABA = .67), whereas the predictions with logistic regressions were clearly less accurate (ABA = .59). The same pattern across algorithms occurred for AUC, an alternative measure of predictive performance. Compared to Sample 1, the ABA and AUC within Sample 2 were generally higher and the difference between the modeling approaches negligible (logistic regression: ABA = .74; Enet and GBM: ABA = .75). To further illustrate the differences in predictive performance across both samples and the two machine learning algorithms, Figure 1 shows ROC curves for each of the 300 testing iterations as well as averaged ROC curves for each sample. Differences between the performance of the algorithms within samples were moderate and unsystematic. However, it is evident that the larger sample size of Sample 2 led to less variation across different training-testing data splits. In other words, the person sampling implemented in the outer validation loop becomes less important.

There are different potential explanations for the superior prediction performance in Sample 2: (a) there might be specific important or more reliable predictor variables in the data set, (b) the larger sample size, and (c) the larger overall number of dropout events. To shed

some light on this issue, we investigated the effect of sample size, number of events, and—as an additional exploratory condition—events fraction on predictive accuracy using elastic net regressions and all available predictor variables in Sample 2 (see Figure 2). There were two main findings of this supplementary analyses: First, when reducing the overall sample size (starting at the total sample size of Sample 2) and keeping the events fraction fixed at 5.13 % (*Sample Size condition*), balanced accuracies remained stable till  $N = 4,500$ , decreasing more steeply for  $N < 2,000$ . When the sample size is reduced to the one of Sample 1, the balanced accuracies of Sample 2 were still higher in comparison, but less pronounced (.70 and .67 compared to .75 and .67 for the full samples). Second, the *Dropouts condition* showed that the effect of reducing the events fraction (i.e., keeping all completers in the sample and only reducing the number of events/dropouts) was even more crucial with an overall steeper decrease and lower accuracies for very small fractions scenarios.

Irrespective of the sample or the algorithm, averaged sensitivities in the testing dataset were generally lower than specificities, meaning that the group of dropouts could be detected less accurately compared to those who completed their therapy, which is to be expected given the low percentage of overall dropouts in the samples. Positive Predictive Values (PPV) and Negative Predictive Values (NPV) were overall similar across algorithms: The relatively low PPV for all algorithms indicate that only 15-17% of all flagged patients were actual dropouts, whereas the relatively high NPV of .94-.98 indicate that nearly all patients identified as completers were correctly classified. Given the relatively low prevalence of therapy dropouts in both samples, these patterns were to be expected since PPV and NPV are very sensitive to imbalanced data (see also Belsher et al., 2019).

As a kind of sensitivity check, we examined whether our models predominantly correctly classified early dropouts, that is, patients who dropped out during the first week of treatment, indicating that we would only be able to detect patients who probably were less inclined to participate from the onset. To do so, we excluded all early dropouts and reran the

elastic net regressions using all predictor variables. ABA slightly decreased in both samples ( $\Delta ABA = .01$  in Sample 1 and  $.03$  in Sample 2), but this decrease can be explained in part by the inherent reduction in the number of overall dropouts in these analyses. Hence, we did not find any indication for differences in the predictive accuracy dependent on the time of dropout.

### **Which Variables Predict Therapy Dropout?**

Figure 2 shows the ten most important predictor variables for the machine learning models (i.e., elastic net regressions and GBM). In Sample 1, the mean score of the Helping Alliance Questionnaire (reported by the respective therapist at baseline in the initial interview) was by far the most important predictor across both modeling approaches, followed by the patients' age. Both variables had a negative impact, that is, a good helping alliance and higher age were associated with a lower probability of dropping out. Further, albeit significantly less, important variables of the elastic net models were whether the patients had any affective/mood disorder diagnosis (section F3 in ICD-10), variables concerning their current employment, and satisfaction with their financial situation. Overall, elastic net regressions and GBM had five out of the ten most important predictors in common. However, due to their relatively low and small differences in variable importances, rankings beyond the two most important variables should be taken with a grain of salt.

In Sample 2, patients' motivation towards therapy as rated by the therapist was the most important variable across algorithms. Hence, the initial evaluation after the first interview was the most important predictor of eventual therapy dropout across samples and algorithms. Similar to the results of Sample 1, higher age was associated with a lower probability of dropping out for Sample 2 as well. Apart from that, higher therapist-rated impairments in structural integration (e.g., Zimmermann et al., 2012), a higher number of cigarettes per day, having any F3 diagnosis, receiving unemployment benefits and more previous stays in similar clinics were among the most important variables for both algorithms.

Comparing both samples, it should be noted that there was no information on the level of structural integration rating or on smoking available in Sample 1, so one cannot infer that these variables were not relevant in Sample 1.

Given the outstanding importance of the initial impression regarding patients' motivation or helping alliance as assessed by the therapist—irrespective of sample or algorithm—we additionally tested the predictive performance of elastic net regressions using all predictor variables except of helping alliance (Sample 1) or patients' motivation (Sample 2). Excluding the most important predictor from the full model led to a substantial drop in overall predictive performance for both samples ( $\Delta ABA = .08$ ). Using only the helping alliance as rated by the therapists in Sample 1 in a logistic regression model resulted in model performances nearly as good as the full elastic net regression model (see Table 1). Notably, this model even outperformed the highly overfitted logistic regressions that included all variables. In Sample 2, a model using solely patients' motivation as a predictor has slightly lower ABA ( $\Delta ABA = .02$ ) and AUC ( $\Delta AUC = .05$ ) compared to the elastic net regressions employing the full predictor set and also shifted sensitivity ( $-.07$ ) and specificity ( $+.30$ ). Because sensitivity is arguably more important in identifying potential dropouts than patients who are regularly continuing their therapy, using all available predictors still has incremental value over a simpler model based on the therapists' initial assessment of patients (Sample 2). Comparing the predictor set in both samples, there is an interesting distinction: The therapist-rated motivation variable (Sample 2) seems to detect completers better than dropouts, whereas the helping alliance (Sample 1) seems to discriminate better for higher risk individuals (as indicated by the higher decrease of sensitivity than specificity if the variable is dropped).

### **Discussion**

High rates of therapy dropout are disadvantageous for individual patients and the health care system. Thus, it is worthwhile to gain more insights into which patients are at a higher risk of subsequent dropout ultimately aiming to prevent dropout. In the current study, it

was possible to correctly classify 67% (Sample 1) and 75% (Sample 2) of all patients into either therapy dropouts or completers using machine learning prediction models including only baseline indicators of large naturalistic inpatient samples. As it is common for classification tasks with highly unequal group sizes (e.g., Belsher et al., 2019; Jankowsky et al., 2023), we found that members of the majority group, that is therapy completers, could be predicted more accurately. Compared to Bennemann et al. (2022) who predicted therapy dropout in a German outpatient sample, AUC were higher in our analyses (.74/.83 vs. .66) which might be attributed to several reasons such as the different dropout ratios, predictor variables, length of therapies, or settings (inpatient vs. outpatients).

The usefulness of any statistical model should ultimately be evaluated with a cost-benefit calculation. Although treatment response is influenced by multiple, time-variant factors emphasizing the need for multi-modal and longitudinal data (e.g., Chekroud et al., 2021), models such as the present ones can easily be implemented because they rely on the baseline assessment that is carried out in German psychotherapy clinics on a routine basis. Hence, using prediction models with baseline indicators provides valuable information on the risk of dropout for an individual patient early in the therapeutical process when intervention is still possible without imposing additional burden on therapists or patients.

### **The Role of Therapists' Impressions for the Prediction of Therapy Dropout**

We found that the initial impression regarding patients' helping alliance by the respective therapist was the most important predictor of eventual therapy dropout in Sample 1, that is, a low-quality alliance was associated with a higher probability of dropout. Previous research showed similar effects: In a recent meta-analysis covering 295 studies including more than 30,000 patients, the estimated correlation between alliance and therapy dropout was  $r = .18$  (Flückiger et al., 2018). In principle, the importance of alliance could be encouraging news, because early intervention would be possible in the event of an unproductive alliance and therapist could act on their initial assessment. Repairment of so-

called alliance ruptures has been shown to have a positive effect on treatment response (e.g.,  $r = .29$  in a meta-analysis by Eubanks et al., 2018).

In Sample 2, patients' baseline motivation as rated by the therapist was the most important predictor of dropout; highly motivated patients had a lower probability of dropping out. The single-item assessment of patients' motivation presumably tap a similar construct as the Helping Alliance Questionnaire used in Sample 1 (which includes questions such as "I believe that the patient is sufficiently motivated for treatment" or "I have the impression that the patient is affectively responsive to my therapeutic interventions"), again underlining the importance of motivational factors in therapy. Low motivation and/or dissatisfaction with the treatment has also been shown to be one of the main reasons for therapy dropout in outpatients (e.g., Bados et al., 2007) and to moderate the relationship between patient-rated therapeutic alliance and treatment outcomes in a cognitive-behavioral treatment (e.g., Rivera et al., 2023). Also, therapists might be adept at assessing which patients are suited for or able to adjust to the particular treatment concept of an inpatient setting, which might guide the impression in their assessment of patients' overall motivation.

Another important variable for the prediction of dropout in both samples was patients' age, that is, older patients were less likely to drop out of therapy. Older patients were also more often diagnosed with an affective disorder and usually have a higher level of structural integration (Zimmermann et al., 2020) which were both negatively associated with therapy dropout. Hence, the effect of age could be a combination of having a better "treatable" diagnosis and demographic factors such as fewer young children to take care of, a more stable financial situation etc. that would make it more accessible to be in treatment continuously for multiple weeks. In Sample 2, the number of cigarettes per day and past alcohol abuse were also relevant predictors of therapy dropout, indicating that addictive behaviors can be at odds with successfully following the structured treatment in inpatient clinics.

**On the Requirements of Reasonable Classification Models in Clinical Psychology**

There is an ongoing debate on the use(fulness) of machine learning prediction models in clinical psychology (e.g., Wilkinson et al, 2020). With the current study, we added to the critical literature arguing that more complex models allowing for interactions and non-linear effects (such as the GBM) did not always necessarily result in higher predictive accuracy when compared to regularized regression models. More critically, initial findings on the superiority of more complex machine learning models in clinical science have serious statistical flaws such as an incorrect cross-validation procedure (Jacobucci et al., 2021; Kapoor & Narayanan, 2022). Furthermore, such models usually require reliable indicators and large sample sizes to be able to correctly identify complex patterns in the data (e.g., Jacobucci & Grimm, 2020). In the context of medical/clinical classification tasks, however, small samples sizes and a low number of events (= the outcome to be predicted) are the rule, not the exception.

Generally, the required sample size and number of events in classification tasks depend on multiple factors and specific recommendations should always be treated as limited to their respective context. For example, Giesemann et al. (2023) recommended at least 300 patients for training samples when using machine learning algorithms to predict therapy dropout. However, they used less predictor variables (7) and had a higher event fraction (30%) as in this study. Three- or even four-hundred patients within the training sample would be similar to the smallest sample size of our simulation (see Figure 2) and thus clearly not advisable for this study. A common (and often criticized) rule of thumb is to use at least ten events per predictor to avoid biased estimates (e.g., Moons et al., 2014). More recent studies suggest that using regularized regression models allows for a relaxation of this rule (e.g., Pavlou et al., 2016) and that the number of predictors, overall sample size, and events fraction are all necessary to consider (Van Smeden et al., 2019). We underline with the present study the notion that a reduction of the events fraction leads to lower predictive performance even if the overall sample size is large ( $N > 11,000$ ).

## Limitations

Several limitations pertain to the setup of the study and the validation of the prediction models. First, no information was available whether and how therapists modified their behavior dependent on their initial impression of their patients' motivation or of the helping alliance. To shed more light on this, future research should use assessments of therapists' characteristics, strategies in dealing with non-optimal helping alliance, and the dynamics of the helping alliance (e.g., Flückiger et al., 2020) over the course of the therapy for the prediction of therapy dropout.

Second, we also were only able to use therapists' ratings of the helping alliance and treatment motivation which does not capture the patients' perspective. Since it has been meta-analytically shown that patient- and therapist-ratings on alliance are only moderately correlated ( $r = .36$ ; Tryon et al., 2007), the inclusion of the patients' view would be important to get a more nuanced picture. Fortunately, this limitation might be less significant according to studies (on outpatients) evidencing that therapists' variability as opposed to patients' variability in alliance relates to treatment outcomes (e.g., Baldwin et al., 2007).

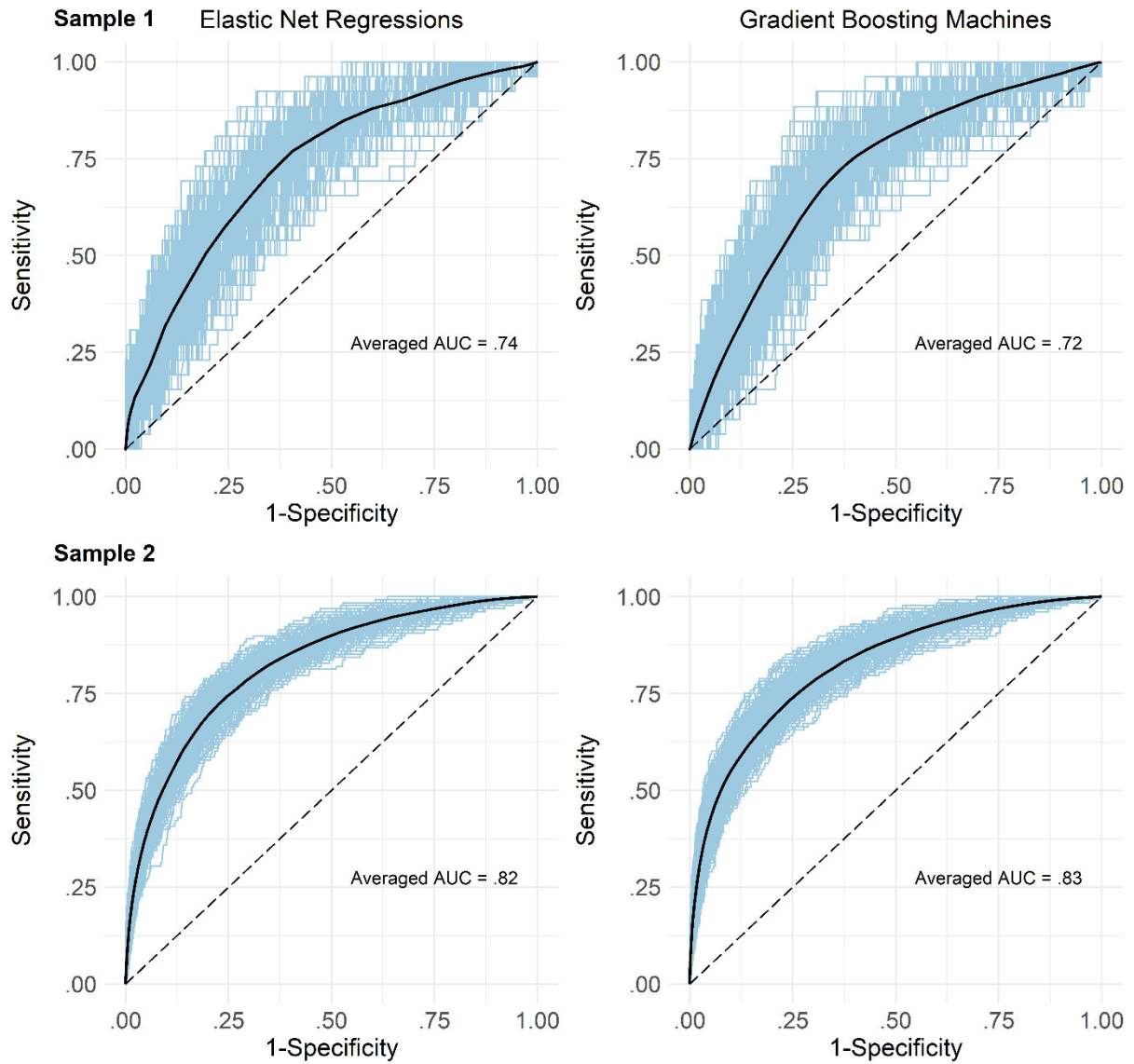
Third, the two data sets varied in many respects, among others in the composition of the patients, the implemented interventions, and the variables gathered upon admission to the clinic. We think that, on the one hand, the two naturalistic samples have high ecological validity and thus the resulting models are practically relevant, but on the other hand, some differences between the samples make a direct comparison difficult. It would therefore be highly beneficial to standardize baseline assessments across clinics—at least to a common core set of measures—to implement a more rigorous form of validation across independent samples in future studies (Dwyer et al., 2018). Concerning generalizability, it should be noted that treatment concepts and patient characteristics differ between in- and outpatients in Germany, but also internationally, so findings based on German inpatients are partly limited to this specific context. Thus, we encourage other researchers to conduct similar studies in



different contexts to study if the variables found in this study to be predictive do translate to other settings. On a methodological stance, we showed that more complex machine learning algorithms are not superior over simpler regularized algorithms in clinical datasets evaluating therapy success.

**Figure 1**

*ROC Curves for Each of the 300 Testing Iterations Including an Averaged ROC Curve*



*Note.* For the averaged ROC curve (black bold line), ROC curve values were averaged across 50 evenly distributed cutpoints.

**Table 1**

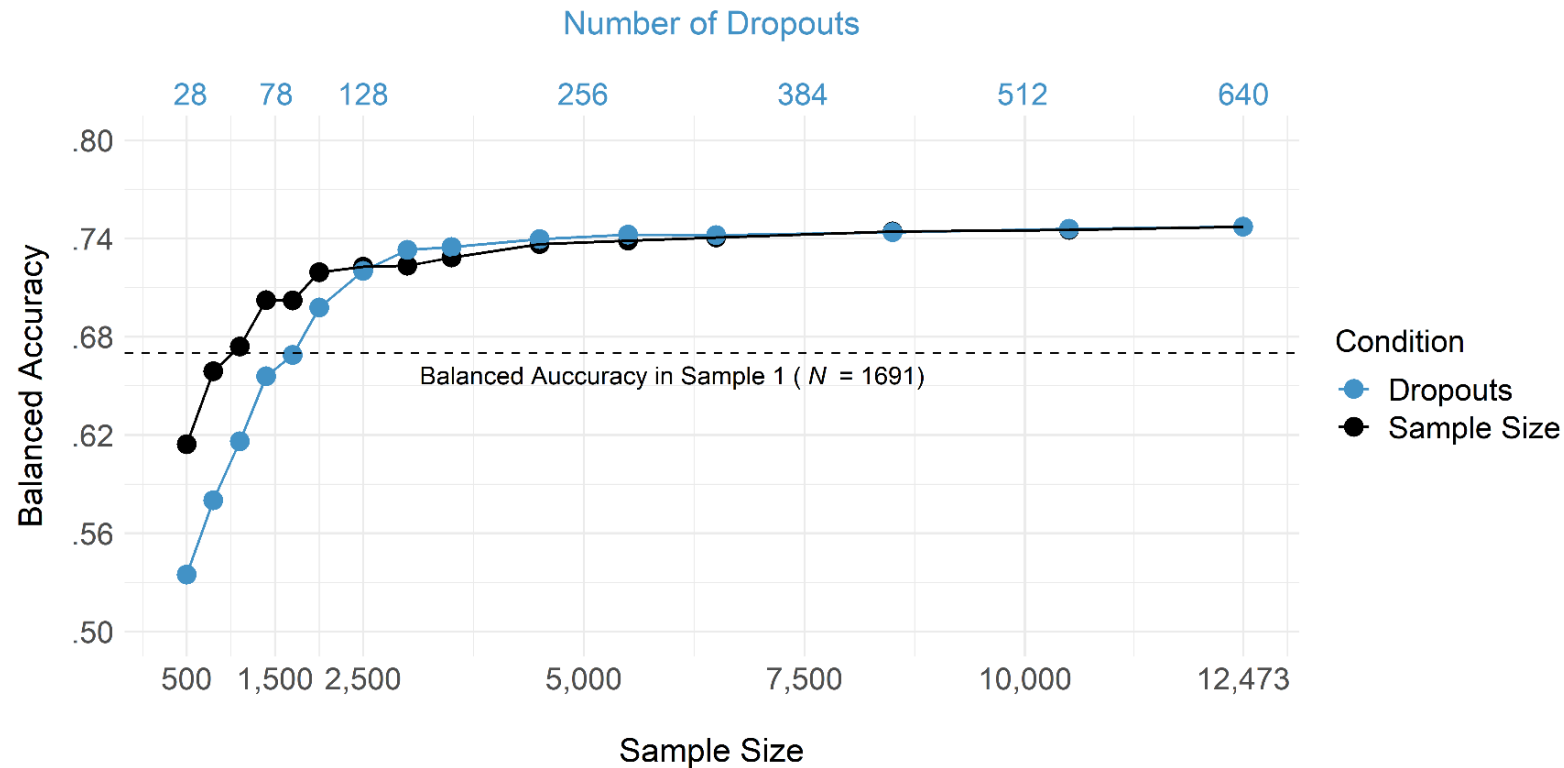
*Averaged Performance Metrics Across 300 Testing Iterations*

	Sample 1						Sample 2					
	ABA	Sensitivity	Specificity	PPV	NPV	AUC	ABA	Sensitivity	Specificity	PPV	NPV	AUC
Logistic Regression	.58 [.05]	.36 [.10]	.81 [.03]	.13 [.03]	.94 [.01]	.64 [.06]	.74 [.02]	.68 [.04]	.80 [.01]	.15 [.01]	.98 [.00]	.81 [.02]
Elastic Net Regression	.67 [.04]	.64 [.09]	.71 [.02]	.15 [.02]	.96 [.01]	.74 [.04]	.75 [.02]	.70 [.04]	.80 [.01]	.16 [.01]	.98 [.00]	.82 [.02]
Gradient Boosting	.67 [.05]	.64 [.11]	.70 [.04]	.15 [.02]	.96 [.01]	.72 [.04]	.74 [.03]	.73 [.07]	.75 [.10]	.15 [.03]	.98 [.01]	.83 [.02]
Single Predictor	.66 [.04]	.63 [.08]	.70 [.02]	.15 [.02]	.96 [.01]	.73 [.04]	.73 [.02]	.63 [.04]	.83 [.01]	.17 [.01]	.98 [.00]	.77 [.02]
All But One Predictor	.59 [.04]	.50 [.09]	.68 [.04]	.11 [.02]	.94 [.01]	.62 [.05]	.67 [.02]	.61 [.04]	.73 [.01]	.11 [.01]	.97 [.00]	.74 [.02]

*Note.* ABA = Average Balanced Accuracy; Sensitivity = Ratio of correctly classified dropouts to all dropouts; Specificity = Ratio of correctly classified completers to all completers; PPV = Positive Predictive Value (Proportion of true dropouts out of all patients who were flagged as dropouts); NPV = Negative Predictive Value (Proportion of true completers out of all patients who were flagged as completers); AUC = Area Under the Curve. Standard deviations are given in brackets. In the single predictor model, only helping alliance (Sample 1) or patients' motivation (Sample 2) as rated by the therapists are included in a logistic regression model. In the all but one predictor model, those variables were excluded of an elastic net regression model, respectively.

**Figure 2**

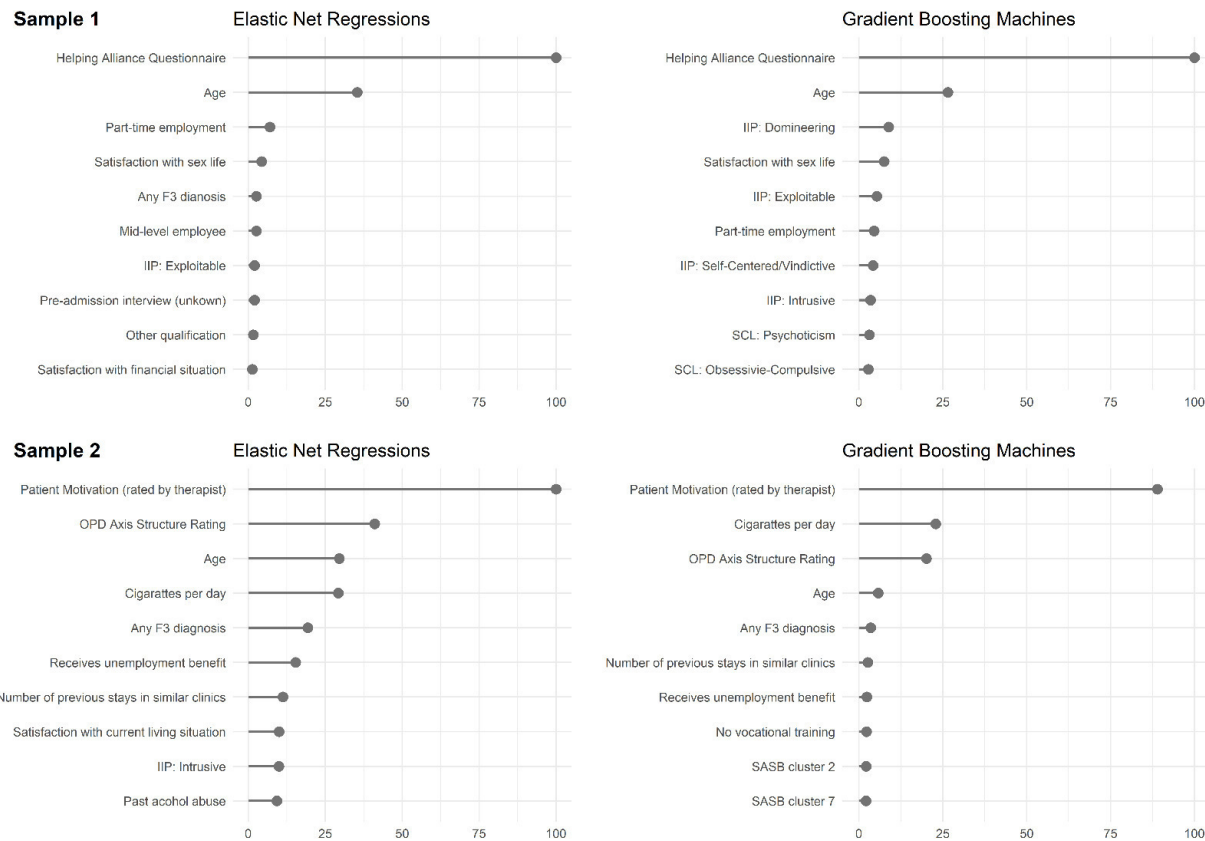
*Balanced Accuracies in Sample 2 Depending on the Sample Size and Events (Dropouts) Fraction*



*Note.* In the Sample Size condition (black line), the overall sample size was reduced stepwise (full sample, 10,500, 8,500, 6,500, 5,500, 4,500, 3,500, 2,500, 2,000, 1,700, 1,400, 1,100, 800, 500) but the events fraction was kept at the same level as in the full sample (= 5.13%). For the Dropouts condition (blue line), we always used all completers ( $N = 11,833$ ) and added 5.13% of different sample size levels as dropouts, thereby reducing only the number of dropouts to that of the Sample Size condition. Ultimately, this reduces the events fraction (5.13, 4.36, 3.55, 2.75, 2.33, 1.91, 1.50, 1.28, 1.07, .86, 0.73, 0.60, 0.47, 0.35, 0.22). Importantly, the number of dropouts is identical in both conditions at a given point on x-axis.

**Figure 3**

*Averaged Variable Importances for the ten Most Important Predictor Variables*



*Note.*  $N = 1,691$  (Sample 1) and  $N = 12,473$  (Sample 2).

### References

- Alden, L. E., Wiggins, J. S., & Pincus, A. L. (1990). Construction of circumplex scales for the Inventory of Interpersonal Problems. *Journal of Personality Assessment*, 55(3-4), 521–536. [https://doi.org/10.1207/s15327752jpa5503&4\\_10](https://doi.org/10.1207/s15327752jpa5503&4_10)
- Bados, A., Balaguer, G., & García, C. S. (2007). The efficacy of cognitive–behavioral therapy and the problem of drop-out. *Journal of Clinical Psychology*, 63(6), 585–592. <https://doi.org/10.1002/jclp.20368>
- Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology*, 75(6), 842–852. <https://doi.org/10.1037/0022-006X.75.6.842>
- 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>
- Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., Jüni, P., & Cuijpers, P. (2013). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLoS Medicine*, 10(5), e1001454. <https://doi.org/10.1371/journal.pmed.1001454>
- Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., Morgan, R. L., Evatt, D. P., Tucker, J., & Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*, 76(6), 642-651. <https://doi.org/10.1001/jamapsychiatry.2019.0174>
- 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*, 1–10. Advance online publication.

<https://doi.org/10.1192/bjp.2022.17>

Bucher, M. A., Suzuki, T., & Samuel, D. B. (2019). A meta-analytic review of personality traits and their associations with mental health treatment outcomes. *Clinical*

*Psychology Review*, 70, 51–63. <https://doi.org/10.1016/j.cpr.2019.04.002>

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>

Derogatis, L. R. (1982). *Brief Symptom Inventory (BSI)* [Database record]. APA PsycTests.

<https://doi.org/10.1037/t00789-000>

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14(1), 91–118.

<https://doi.org/10.1146/annurev-clinpsy-032816-045037>

Eubanks, C. F., Muran, J. C., & Safran, J. D. (2018). Alliance rupture repair: A meta-analysis. *Psychotherapy*, 55(4), 508–519. <https://doi.org/10.1037/pst0000185>

Fernandez, E., Salem, D., Swift, J. K., & Ramtahal, N. (2015). Meta-analysis of dropout from cognitive behavioral therapy: Magnitude, timing, and moderators. *Journal of Consulting and Clinical Psychology*, 83(6), 1108–1122.

<https://doi.org/10.1037/ccp0000044>

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, 55(4), 316–340.

<https://doi.org/10.1037/pst0000172>

Flückiger, C., Rubel, J., Del Re, A. C., Horvath, A. O., Wampold, B. E., Crits-Christoph, P., Atzil-Slonim, D., Compare, A., Falkenström, F., Ekeblad, A., Errázuriz, P., Fisher, H., Hoffart, A., Huppert, J. D., Kivity, Y., Kumar, M., Lutz, W., Muran, J. C., Strunk, D.

- R., Tasca, G. A., ... Barber, J. P. (2020). The reciprocal relationship between alliance and early treatment symptoms: A two-stage individual participant data meta-analysis. *Journal of Consulting and Clinical Psychology*, 88(9), 829–843. <https://doi.org/10.1037/ccp0000594>
- Franke, G., H. (1995). *SCL-90-R: Die Symptom-Check-Liste von Derogatis – Deutsche Version*. Beltz Test Gesellschaft, Göttingen.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/101320345>
- Gieseemann, 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*, 1–13. <https://doi.org/10.1080/10503307.2022.2161432>
- Gonzalez Salas Duhne, P., Delgadillo, J., & Lutz, W. (in press). Predicting early dropout in online versus face-to-face guided self-help: A machine learning approach. *Behaviour Research and Therapy*.
- Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers. (2019). *gbm: Generalized boosted regression models* (Version 2.1.5) [Computer software]. <https://CRAN.R-project.org/package=gbm>
- Hans, E., & Hiller, W. (2013). A meta-analysis of nonrandomized effectiveness studies on outpatient cognitive behavioral therapy for adult anxiety disorders. *Clinical Psychology Review*, 33(8), 954–964. <https://doi.org/10.1016/j.cpr.2013.07.003>
- Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, 15(3), 809–816. <https://doi.org/10.1177/1745691620902467>



- Jacobucci, R., & Li, X. (2022). Does minority case sampling improve performance with imbalanced outcomes in psychological research?. *Journal of Behavioral Data Science*, 2(1), 59–74. <https://doi.org/10.35566/jbds/v2n1/p3>
- Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of inflated prediction performance: A commentary on machine learning and suicide research. *Clinical Psychological Science*, 9(1), 129–134. <https://doi.org/10.1177/2167702620954216>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning*. Springer.
- Jankowsky, K., Steger, D., & Schroeders, U. (2023). Predicting lifetime suicide attempts in a community sample of adolescents using machine learning algorithms. *Assessment*, 0(0). <https://doi.org/10.1177/10731911231167490>
- Kamenov, K., Twomey, C., Cabello, M., Prina, A. M., & Ayuso-Mateos, J. L. (2017). The efficacy of psychotherapy, pharmacotherapy and their combination on functioning and quality of life in depression: a meta-analysis. *Psychological Medicine*, 47(3), 414–425. <https://doi.org/10.1017/S0033291716002774>
- Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science (arXiv:2207.07048). arXiv. <http://arxiv.org/abs/2207.07048>
- Karterud, S., Pedersen, G., Bjordal, E., Brabrand, J., Friis, S., Haaseth, O., Haavaldsen, G., Irion, T., Leirvåg, H., Tørum, E., & Urnes, O. (2003). Day treatment of patients with personality disorders: experiences from a Norwegian treatment research network. *Journal of Personality Disorders*, 17(3), 243–262. <https://doi.org/10.1521/pedi.17.3.243.22151>
- Kuhl, J., & Kazén, M. (2009). Persönlichkeits-Stil-und Störungs-Inventar (PSSI) [Personality Styles and Disorder Inventory (PSDI)]. Göttingen, Germany: Hogrefe.

- Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5). <https://doi.org/10.18637/jss.v028.i0>
- Leichsenring, F., Masuhr, O., Jaeger, U., Rabung, S., Dally, A., Dümpelmann, M., Fricke-Neef, C., Steinert, C., & Streeck, U. (2016). Psychoanalytic-interactional therapy versus psychodynamic therapy by experts for personality disorders: a randomized controlled efficacy-effectiveness study in cluster b personality disorders. *Psychotherapy and Psychosomatics*, 85(2), 71–80. <https://doi.org/10.1159/000441731>
- Leichsenring, F., Steinert, C., Rabung, S., & Ioannidis, J. P. A. (2022). The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: An umbrella review and meta-analytic evaluation of recent meta-analyses. *World Psychiatry*, 21(1), 133–145. <https://doi.org/10.1002/wps.20941>
- Moons, K. G., de Groot, J. A., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Medicine*, 11(10), e1001744. <https://doi.org/10.1371/journal.pmed.1001744>
- Pargent, F., Schoedel, R., Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, 6(3). <https://doi.org/10.1177/25152459231162559>
- Pavlou, M., Ambler, G., Seaman, S. R., De Iorio, M., & Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, 35(7), 1159–1177. <https://doi.org/10.1002/sim.6782>
- Rivera, A. P., Maisto, S. A., Connors, G. J., & Schlauch, R. C. (2023). Therapists' first impression of treatment motivation moderates the relationship between the client-rated therapeutic alliance and drinking outcomes during treatment. *Alcoholism: Clinical and Experimental Research*, 47(4), 806–821. <https://doi.org/10.1111/acer.15040>

- Schmidt, I. D., Forand, N. R., & Strunk, D. R. (2019). Predictors of dropout in internet-based cognitive behavioral therapy for depression. *Cognitive Therapy and Research, 43*(3), 620–630. <https://doi.org/10.1007/s10608-018-9979-5>
- Spitzer, C., Müller, S., Kerber, A., Hutsebaut, J., Brähler, E., & Zimmermann, J. (2021). Die deutsche Version der Level of Personality Functioning Scale-Brief Form 2.0 (LPFS-BF): Faktorenstruktur, konvergente Validität und Normwerte in der Allgemeinbevölkerung. *Psychotherapie Psychosomatik Medizinische Psychologie, 71*(07), 284–293. <https://doi.org/10.1055/a-1343-2396>
- Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. E., Patel, V., & Silove, D. (2014). The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology, 43*(2), 476–493. <https://doi.org/10.1093/ije/dyu038>
- Streeck, U., & Leichsenring, F. (2010). Handbuch psychoanalytisch-interaktionelle Therapie. *Psychotherapeut, 55*, 268–270.
- Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology, 80*(4), 547–559. <https://doi.org/10.1037/a0028226>
- Tryon, G. S., Blackwell, S. C., & Hammel, E. F. (2007). A meta-analytic examination of client–therapist perspectives of the working alliance. *Psychotherapy Research, 17*(6), 629–642. <https://doi.org/10.1080/10503300701320611>
- Van Smeden, M., Moons, K. G., De Groot, J. R., 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>
- Wilkinson, J., Arnold, K. F., Murray, E. J., van Smeden, M., Carr, K., Sippy, R., de Kamps, M., Beam, A., Konigorski, S., Lippert, C., Gilthorpe, M. S., & Tennant, P. W. G.

(2020). Time to reality check the promises of machine learning-powered precision medicine. *The Lancet. Digital Health*, 2(12), e677–e680.

[https://doi.org/10.1016/S2589-7500\(20\)30200-4](https://doi.org/10.1016/S2589-7500(20)30200-4)

Zimmermann, J., Ehrental, J. C., Cierpka, M., Schauenburg, H., Doering, S., & Benecke, C.

(2012). Assessing the level of structural integration using operationalized psychodynamic diagnosis (OPD): implications for DSM-5. *Journal of Personality Assessment*,

94(5), 522–532. <https://doi.org/10.1080/00223891.2012.700664>

Zimmermann, J., Müller, S., Bach, B., Hutsebaut, J., Hummelen, B., & Fischer, F. (2020). A

common metric for self-reported severity of personality disorder. *Psychopathology*,

53(3-4), 168–178. <https://doi.org/10.1159/000507377>

Zimmermann, D., Rubel, J., Page, A. C., & Lutz, W. (2017). Therapist effects on and

predictors of non-consensual dropout in psychotherapy. *Clinical Psychology &*

*Psychotherapy*, 24(2), 312–321. <https://doi.org/10.1002/cpp.2022>

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal*

*of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

<https://doi.org/10.1111/j.1467-9868.2005.00503.x>

# Predicting treatment response using Machine Learning: A registered report

Kristin Jankowsky<sup>1</sup>, Krakau, L.<sup>2</sup>, Schroeders, U.<sup>1</sup>,

Zwerenz, R.<sup>2</sup>, & Beutel, M. E.<sup>2</sup>

1: University of Kassel

2: University Medical Center Mainz

Status – accepted

Jankowsky, K., Krakau, L., Schroeders, U., Zwerenz, R., & Beutel, M. E. (2023). Predicting treatment response using Machine Learning: A registered report. *British Journal of Clinical Psychology*, Advance online publication. <https://doi.org/10.1111/bjc.12452>

## REGISTERED REPORT STAGE 2

# Predicting treatment response using machine learning: A registered report

Kristin Jankowsky<sup>1</sup>  | Lina Krakau<sup>2</sup> | Ulrich Schroeders<sup>1</sup> |  
Rüdiger Zwerenz<sup>2</sup> | Manfred E. Beutel<sup>2</sup>

<sup>1</sup>Psychological Assessment, University of Kassel, Kassel, Germany

<sup>2</sup>Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Mainz, Mainz, Germany

## Correspondence

Kristin Jankowsky, University of Kassel, Hollaendische Strasse 36-38, Kassel 34127, Germany.

Email: [jankowsky@psychologie.uni-kassel.de](mailto:jankowsky@psychologie.uni-kassel.de)

## Abstract

**Objective:** Previous research on psychotherapy treatment response has mainly focused on outpatients or clinical trial data which may have low ecological validity regarding naturalistic inpatient samples. To reduce treatment failures by proactively screening for patients at risk of low treatment response, gain more knowledge about risk factors and to evaluate treatments, accurate insights about predictors of treatment response in naturalistic inpatient samples are needed.

**Methods:** We compared the performance of different machine learning algorithms in predicting treatment response, operationalized as a substantial reduction in symptom severity as expressed in the Patient Health Questionnaire Anxiety and Depression Scale. To achieve this goal, we used different sets of variables—(a) demographics, (b) physical indicators, (c) psychological indicators and (d) treatment-related variables—in a naturalistic inpatient sample ( $N = 723$ ) to specify their joint and unique contribution to treatment success.

**Results:** There was a strong link between symptom severity at baseline and post-treatment ( $R^2 = .32$ ). When using all available variables, both machine learning algorithms outperformed the linear regressions and led to an increment in predictive performance of  $R^2 = .12$ . Treatment-related variables were the most predictive, followed psychological indicators. Physical indicators and demographics were negligible.

**Conclusions:** Treatment response in naturalistic inpatient settings can be predicted to a considerable degree by using baseline indicators. Regularization via machine learning algorithms leads to higher predictive performances as

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *British Journal of Clinical Psychology* published by John Wiley & Sons Ltd on behalf of British Psychological Society.

opposed to including nonlinear and interaction effects. Heterogenous aspects of mental health have incremental predictive value and should be considered as prognostic markers when modelling treatment processes.

#### KEYWORDS

inpatients, machine learning, predictive modelling, prognostic markers, treatment response

#### Practitioner Points

- The present study shows that patients' characteristics at the start of a psychotherapy can be used to predict treatment response.
- Machine learning algorithms can help enhance predictive accuracy, however, not due to the incorporation of nonlinear or interaction effects but rather by reducing the models' overfit via regularization, stressing the need for high-quality data and reliable indicators rather than more complex models.
- Beyond baseline symptomatic, various indicators on mental health had incremental value for the prediction of treatment response and should therefore be focused on at baseline assessments (as opposed to demographics and indicators of physical health).
- Prediction models such as the one in this study could be implemented using routine baseline assessments and provide valuable information on the risk of treatment nonresponse at a time when intervention is still possible.

## BACKGROUND

Multi-modal inpatient treatment is a valid and effective treatment option for patients with severe mental disorders (Liebherz & Rabung, 2014). In Germany, psychosomatic treatment is offered both in an inpatient and day-clinic setting with psychotherapy as its main treatment modality complemented with additional somatic, psychopharmacological and specialized therapies (e.g., creative therapy). While patient samples and response rates are comparable between inpatient and day-clinic settings (Zeeck et al., 2015), not all patients respond equally well to treatment. Heterogeneous treatment responses have been well documented for outpatient treatment of depression (Kaiser et al., 2022), but several studies showed that the phenomenon also translates to other mental disorders (Altmann et al., 2020; Senger et al., 2021) and inpatient treatment settings (Hartmann et al., 2018; Zeeck et al., 2020).

To improve response rates, reduce relapse rates and avoid exposing patients to multiple treatment failures, researchers and clinicians have been interested in learning about risk factors of treatment non-response, adapting treatments to patient needs and understanding which treatment is best suited to an individual patient (Delgadoillo, 2021; Delgadoillo & Lutz, 2020; Zeeck et al., 2013). Patient, therapist and process factors have all been established to contribute to therapy outcomes (Luborsky et al., 1971; Lutz et al., 2021). Identifying reliable patient pretreatment characteristics would enable practitioners to adapt the treatment to individuals prior to starting treatment, thereby avoiding suboptimal attempts as well as saving financial, time and personnel resources. Despite considerable research effort, findings so far have been mixed with most of the prognostic markers identified making only a minor contribution in explaining treatment response (Chekroud et al., 2021). The most robust finding pertains to the impact

of symptom load (severity) at baseline (Cuijpers et al., 2022), yet studies modelling course trajectories have repeatedly found patient groups with high baseline load who either responded very well or did not change reliably (Altmann et al., 2015).

Other psychological and psychiatric variables that have shown associations with treatment response (for depression) are among others chronicity, psychosocial functioning, psychological and physical comorbidity, personality, childhood adversity and recent trauma, cognitive deficits and coping resources (Kessler et al., 2017; Maj et al., 2020). For inpatient treatment specifically, comorbid (mental) disorders, personality, chronicity and patient motivation have been found to impact response (Beutel & Bleichner, 2011; Zeeck et al., 2016, 2020). However, the number of studies empirically addressing this question in inpatient settings is sparse. Rather than a single, predominant factor leading to treatment response, multiple predictors ‘[outweigh] and [interact] with each other in so far incomprehensible ways’ (Hilbert et al., 2021, p. 53). These predictors can be assigned to different variable groups, for example: (a) sociodemographic background variables (e.g., gender, age), (b) indicators of physical health (e.g., subjective health, BMI, smoking), (c) indicators of personality and mental health (e.g., maladaptive personality traits, anxiety or depression scores) and (d) treatment variables (e.g., number of treatments within the last 12 months). These groups of predictors differ in terms of reliability, assessment method (e.g., self-report questionnaire, clinical interview) and time- and content-related proximity to the criterion (proximal vs. distal).

Research on prognostic markers has heavily relied on re-analyses of clinical trial data. These individual studies are oftentimes underpowered, limiting the identification of reliable predictors and interaction effects (Fisher et al., 2017). However, even adequately powered individual participant data meta-analyses (IPD-MA) have mostly reported symptom severity as the single best predictor of treatment response for depression (see Cuijpers et al., 2022, for an overview). One problematic aspect of randomized controlled trials (RCT) which are considered gold-standard for therapy evaluation is their limited ecological validity (Philips & Falkenström, 2021). RCTs usually have strict in- and exclusion criteria (e.g., excluding patients with multiple comorbidities, see O'Hara et al., 2017), leading to more homogenous samples compared to the population. Patients presenting, for example, acute suicidal ideation, substance abuse, or specific personality disorders are excluded from RCTs although, from a clinical perspective, these factors likely interfere with treatment compliance or complicate treatment (Krause & Behn, 2021). Moreover, these more severe clinical characteristics are key reasons why patients seek more intensive care in inpatient and day hospital wards. As this specific group of patients is being precluded from participating in RCTs and their inclusion into RCTs is often not feasible due to ethical reasons, these trials have limited capacity to inform treatment prognosis for real-world intensive care settings (Webb et al., 2020).

Recent work has pointed to the potential benefits of machine learning (ML) techniques in large-scale observational data (Aafjes-van Doorn et al., 2021; Chekroud et al., 2021; Dwyer et al., 2018). As we do not have a clear theoretical model in which ways patients' sociodemographics interact with psychological and medical variables and how these translate to treatment (non)-response, the field embraces the possibilities of ML to examine a plethora of predictors and their potentially nonlinear and higher-order interaction effects. Rather than evaluating a specific, theoretically derived moderator of treatment response in a rather simplistic understanding of dependencies (see also the concept of *Flatland Fallacy*, Jolly & Chang, 2019), the goal in ML is to use all available information to establish connections between the variables in a data-driven way and to increase predictive power (Chekroud et al., 2021; Yarkoni & Westfall, 2017).

There have been few studies to date using ML algorithms to predict treatment response in naturalistic inpatient settings including patients with diverse diagnoses. For example, Webb et al. (2020) compared 14 different ML algorithms in the prediction of post-treatment depression scores in the *Patient Health Questionnaire-9* (PHQ-9). In doing so, the authors used a range of predictor variables that were routinely assessed at admission (including demographics, clinical measures, treatment history, or physical health variables). The best-performing algorithm (elastic net regressions) explained 38% of interindividual variance in the depression scores in a holdout sample, meaning a sample that



was not used during model training. Particularly important variables for the prediction of treatment response were the patients' expectations of improvement, baseline symptom severity—that is, baseline PHQ-9 values and baseline Generalized Anxiety Disorder Scale-7 values—as well as whether patients took mood stabilizers.

## Pros and cons of machine learning in psychotherapy research and a call for closer methodological scrutiny

Apart from the possibility of including complex interaction effects to enhance predictive performance, ML algorithms have several in-built features that are promising when trying to tackle methodological challenges usually encountered in the prediction of treatment response. For example, it is possible to reduce model overfit by using algorithms that employ some form of regularization. Overfit can be defined as the difference in predictive performance of a model using training data versus independent, unseen testing data (Urban & Gates, 2021). Especially in scenarios with small sample sizes and a large number of predictors—which is a realistic setting in many studies on treatment outcomes using baseline indicators (Chekroud et al., 2021)—unregularized regression models tend to overfit, hampering the generalizability and the clinical usefulness of the predictive models.

However, the use of ML in clinical psychology has also been viewed critically (Wilkinson et al., 2020). Typical criticisms include a lack of assessment of ML's benefits relative to its costs (Kessler et al., 2020) as well as its worse interpretability compared to simpler models (Siddaway et al., 2020) which could lead to lower clinical utility as well as lower acceptance and implementation rates by clinicians (see Lutz et al., 2022, for an example of the influence of therapists' attitudes towards and rated usefulness of machine learning-based digital decision support and feedback system on its overall effectiveness). On a much more fundamental level, there is also increasing and strong evidence across several research fields including psychiatry that ML analyses are often flawed. For example, many ML models are evaluated incorrectly, biasing model validation in favour of more complex and flexible algorithms as those are better equipped to recognize specific data patterns as well as exploiting any spillover of information between training and testing data (Jacobucci et al., 2021; Kapoor & Narayanan, 2022).

In our reading of the psychometric literature, many reproducibility issues underlying ML studies can be described by two main factors: they are often overhyped and underchecked. We refer to the term overhyped in the sense that similar to a general publication bias, that is, the tendency to publish innovative significant findings with large effects (Ferguson & Brannick, 2012; Ferguson & Heene, 2012), novel ML models that are seemingly highly predictive are more likely to gain traction and to get published. Thus, the incentive to follow a new methodological fashion and to employ new ML models is strong, especially when the outcome to be explained is multifactorially influenced and has steered a lot of inconclusive previous research such as what works for whom in therapy. At the same time, consolidated knowledge of using ML as a statistical tool is not widespread outside computational science and statistics. In a comprehensive survey, Kapoor and Narayanan (2022) showed across a wide range of disciplines (including medicine and psychiatry) that many ML models in the literature were not validated correctly, which could lead to the dissemination of false discoveries or the development of unsubstantiated theories. Consequently, these overoptimistic or biased ML models do not live up to their expectations if correctly validated (see Jacobucci et al., 2021). Thus, we propose a more open debate and culture of mutual scrutiny (Vazire, 2020) to enhance transparency and avoid common pitfalls in ML (see also Cearns et al., 2019; Kapoor & Narayanan, 2022). One way to achieve this is by employing registered reports (Scheel et al., 2021) which make methodological feedback prior to (running the actual study or) conducting the analyses the norm.

## The present study

In this study, we predict treatment response defined by the post-treatment sum score of the Patient Health Questionnaire Anxiety and Depression Scale (PHQ-ADS; Kroenke et al., 2016) in an inpatient sample. The PHQ-ADS is a composite of the 9-item Patient Health Questionnaire and the 7-item Generalized Anxiety Disorder scale (GAD-7; Gräfe et al., 2004) which has been found to be a reliable (Cronbach's alpha between .88 and .92 in three different trials; Kroenke et al., 2016), valid and (sufficiently) unidimensional indicator of depressive symptoms (depression and anxiety). We deliberately decided against grouping patients (remission vs. no remission) according to specific cut-offs to avoid information loss (and grouping those who show no change together with patients whose symptoms deteriorate). We use routine outcome monitoring data from a clinic and polyclinic in Germany, including predictor variables on demographics, personality and indicators of mental health, as well as physical health, and treatment-related variables. This study has three major goals:

First, we further examine the incremental predictive performance of different ML approaches in predicting therapy response by comparing increasingly complex ML models to linear models. Simple linear regression models using either a naïve guessing approach, the PHQ-ADS scores of the baseline assessment as the sole predictor or all information available serve as benchmark models. Thus, we aim to quantify the increment of using all baseline variables beyond naïve guessing or baseline symptom severity. The linear regression model with all available variables was then compared to (a) elastic net regressions as an example of regularized linear regressions and (b) gradient boosting machines, which allows for nonlinear and higher-order interaction effects. This comparison aimed to quantify the incremental value of ML algorithms over and above traditional methods.

Second, we establish the unique and joint contribution of all predictor groups in the prediction of treatment success by systematically rerunning the best-performing algorithm with all possible combinations of groups. The predictors are grouped as follows: (a) sociodemographic variables, (b) indicators of physical health, (c) indicators of mental health and (d) treatment variables. We examine constructs that often have been missing or range-restricted in previous research (e.g., The Personality Inventory for DSM-5) due to homogeneous person sampling in RCTs. Thus, we aim to further knowledge on predictors and moderators of treatment response, potentially screening for participants at risk of treatment nonresponse. To render our prediction models more interpretative, we provide importance measures for all variables of all models.

Third, we counter the objection that machine learning research is inevitably accompanied by increased researcher's degrees of freedom, forming the basis for another reproducibility and replication crisis (Hullman et al., 2022) by registering all analytical decisions beforehand. At first glance, this approach seems to counteract the empirically driven and flexible nature of ML algorithms. However, many aspects concerning data cleaning, variable transformation, handling of missing data, etc. can be registered in ML studies the same way as in every other study. Also, the settings for data-driven hyperparameter tuning can also be defined in advance. Surprisingly, we were not able to find any previous studies on the prediction of psychotherapy outcomes using machine learning that capitalize on the benefits of a registered report. Machine learning modelling and registered reports rarely have been combined in psychological research so far (for one of the few exceptions, see Costello et al., 2021). However, a consistent conclusion of several reviews and meta-analyses on machine learning models in clinical research is that the stark differences in implementation and the often non-transparent model evaluations hinder a useful aggregation of findings (Christodoulou et al., 2019; Lee et al., 2018). We strive to provide an example of a thorough registration of the proposed analysis pipeline that still allows for the analytical flexibility of the ML algorithms (e.g., through hyperparameter tuning).

## METHOD

### Sample

We used routine outcome monitoring data from 723 patients of a clinic and polyclinic in Rhineland-Palatinate collected between 2018 and 2021. The clinic comprises three inpatient and day hospital units offering multi-modal treatment consisting of two to three individual therapy sessions per week, two weekly sessions of art therapy, up to two sessions of body-oriented therapy and up to three sessions of group therapy. The duration of treatment is typically 4–12 weeks. Averaged treatment length in our sample was at 6 weeks. While the focus of the clinic is psychodynamic, treatment integrates different schools and modalities, including educational elements regarding the pathogenesis and maintenance of the disorder, and specific modules (e.g., relaxation training or physiotherapy) tied to the individual needs of the patients. In the group settings, a new member is admitted when a patient is discharged from the hospital. Hence, the groups comprise patients at different treatment stages. This ‘slow-open’ principle offers the possibility for peer learning, where new members can benefit from the perspectives of patients who have already undergone parts of their treatment and more experienced patients can become more aware of their change processes when confronted with attitudes and scepticism of the novices. The multi-professional team consists of psychosomatic medical specialists and residents, psychologists, creative therapists, specialized nurses and social workers. The nursing staff is constantly present, aiming at ensuring stability, holding and reassurance (Beutel et al., 2008).

Using patient data for research is regulated by the German State Hospital Act and was approved by the Rhineland-Palatinate Chamber of Physicians (nr. 837.191.16 (10510)). We provide a descriptive overview of patients' characteristics on all available measures of this study in [Table S1](#) at <https://osf.io/86zng>. The patient data are not publicly available due to privacy restrictions, but we provide a correlation matrix for all variables and a synthetic version of the data in the supplemental materials at <https://osf.io/jxst4/> to render our results as transparent and reproducible as possible. For data access upon request, please contact the second author.

### Measures

In [Table S1](#), we present an overview of all available measures that are included in the prediction models. We predicted the patients' treatment response operationalized as post-treatment PHQ-ADS scale sum scores (controlled for pre-treatment PHQ-ADS sum scores). All categorical variables were dummy-coded prior to the analysis with the first category as a reference. We excluded eight patients with more than 30% missingness on all variables. We also excluded categorical predictor variables with fewer than 10 events to avoid computational problems due to low variances (i.e., two response options regarding pensions due to reduced earning capacity). For our analysis, we used the standardized individual item scores to fully capture all potential effects since it has repeatedly been shown that individual items outperform scale scores in prediction tasks (McClure et al., 2021; Seeboth & Möttus, 2018).

### Statistical analyses

All our analyses were conducted using the R package *caret* (Kuhn, 2008) as an interface for modelling, prediction and evaluation. Irrespective of the modelling algorithm, we employed a nested cross-validation approach (Bischof et al., 2012; Pargent et al., 2023) which is often recommended to strictly separate any data pre-processing and hyperparameter tuning from the final model validation. Thus, nested cross-validation avoids information leakage between the training and the testing sample that is used for model evaluation. Nested cross-validation combines an outer and inner validation loop: First, in every iteration of the *outer validation loop*, the full data are split into training data (for our study, 80%

TABLE 1 Proposed models for the prediction of treatment response.

No.	Model/algorithm	Predictor variables	Tuning parameters
0	Linear Regression	Naïve guessing model	
1	Linear Regression	PHQ-ADS at baseline	
2	Linear Regression	All available variables	
2	Linear Regression	All available variables	
3	Elastic Net Regression	All available variables	$\alpha$ : 40 evenly distributed values between 0 and 1 $\lambda$ : data-driven; sequence between minimum and maximum $\lambda$ generated from a glmnet model, assuming that a sensible range of $\lambda$ is provided using an $\alpha$ value of .50
4	Gradient Boosting Machines	All available variables	Interaction depth of 1, 2, 3, 4, 5 Minimum leaf size of 5, 10, 20, 50 Shrinkage as a sequence between .001 and .201 using steps of .02 Number of trees 50, 100, 150, 300, 500, 1000
4	Best-performing algorithm of Model 2–4 [= BPA]	All four predictor groups <sup>a</sup> (= All available variables)	Parameter set as in best-performing algorithm of Model 2–4
5–8	BPA	Only one predictor group	Parameters of BPA
9–14	BPA	Any pair of two predictor groups	Parameters of BPA
15–18	BPA	Any triple of three predictor groups	Parameters of BPA

Abbreviation: BPA, Best-performing algorithm.

<sup>a</sup>Demographics, physical health, mental health, treatment variables. For more information on the implementation of the glmnet tuning grid, please see also <https://github.com/topepo/caret/blob/master/models/files/glmnet.R>. The repeated model numbers (2 and 4) do not indicate that these models will be estimated multiple times but divide the table into different blocks/sequences of comparison.

of the full data set) and a holdout sample (the remaining 20%) as testing data. Any missing values will be imputed separately for the training and testing datasets (i.e., after the 80/20 split) using multiple imputations ( $k$ =per cent of missingness averaged across all predictor variables, but a minimum of 10) via the random forest algorithm implemented in the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2011). Within the *inner validation loop*, we trained the respective models for each of the imputed training datasets using 10-fold cross-validation. Predictive performance of these models was then calculated as the average performance across the  $k$ -test datasets. Further, we averaged these results across 100 iterations of the *outer validation loop* to provide an accurate estimate of the expected prediction performance using unseen testing data.

Table 1 arranges the different models we compared in this study into three major blocks. As a first step, we quantified the incremental predictive validity of using all variables in comparison to more simple benchmark models. In more detail, the first comparison consists of three linear regression models, (a) a naïve guessing model (or null model, Model 0), (b) a model solely using the PHQ-ADS score at baseline to predict the post-treatment PHQ-ADS scores, because initial symptom severity has been demonstrated a strong predictor of treatment response (Model 1) and (c) a linear regression model with all available predictor variables (Model 2). As a second step, we completed this regression model using all variables (Model 2) to two ML algorithms—elastic net regressions and gradient boosting machines (Model 3 and 4). Finally, in a third step, we examined the unique and combined contribution of different variable groups in the prediction of treatment response (Model 5–18). To this end, we used the algorithm and tuning parameters that showed the highest predictive performance in the aforementioned methodological comparison.

In the following, we briefly describe the key characteristics of the two ML modelling approaches. Elastic net regressions are regularized regressions that lead to parsimonious models by penalizing the regression weights of certain predictors. They compromise between ridge regressions and least absolute shrinkage and selection operator (LASSO) regressions. By using the tunable shrinkage parameter  $\lambda$  and penalty parameter  $\alpha$  (Zou & Hastie, 2005), they strike for an optimal balance between minimizing the sum of squared weights (assigning variables small, but non-zero weights) and the sum of absolute weights (leading to models with many variables given weights of zero), thereby aiming to maximize predictive performance.

Gradient boosting machine algorithms are tree-based algorithms that allow for the integration of nonlinear and higher-order interaction effects into the modelling without the need for specific assumptions on functions between predictor variables and the respective outcome (James et al., 2017). They sequentially combine multiple decision trees into an ensemble. Every new tree aims at fitting the residual error of the previous one, leading to potentially better predictive performance. Their complexity depends on hyperparameter settings (e.g., number of trees, minimum leaf size) which should be sensibly tuned to avoid overfitting due to overly complex models (McNamara et al., 2022).

## Model evaluation

We use the following indices to evaluate predictive models: explained variance ( $R^2$ ), the root mean squared error (RMSE) and the Mean Absolute Error (MAE). All indices are calculated for the training sample and also for the holdout sample across the 100 iterations of splitting the data into training and testing data. We present the results for all indices using box and jitter plots to illustrate (a) the overall predictive performance and (b) the amount of overfit for all modelling approaches. For all models, variable importances are presented using the `varImp` function in `caret`. However, we focus the discussion of important predictor variables on the model with the highest prediction performance.

To further the comprehensibility and accountability of the described analytical approach, we provide annotated R syntax of our analyses. These materials can be found at <https://osf.io/jxst4/>. The time-stamped Stage 1 of this registered report can be found at <https://osf.io/tkm2h>.

## RESULTS

A unidimensional factor explained 42% of the variance of the PHQ-ADS items at baseline. Reliability was high ( $\alpha = .92$ ). The averaged raw difference between baseline and outcome sum scores was  $-9.0$  (empirical range:  $-37; 19$ ). For 13% of the full sample ( $N = 94$  patients, using multiple imputed data), the PHQ-ADS values, that is, symptom load did not change or even increased. This statement is not about statistical or clinical significance, but to describe the wide range of individuals' treatment successes.

Table 2 provides an overview of the averaged predictive performances of all 18 models for the 100 training and testing datasets, respectively. Overall, there was a strong link between the PHQ-ADS at baseline and the post-treatment PHQ-ADS ( $R^2 = .319$ ; Model 1), which is used as the point of reference when making statements about the incremental predictive validity of predictor variable groups.

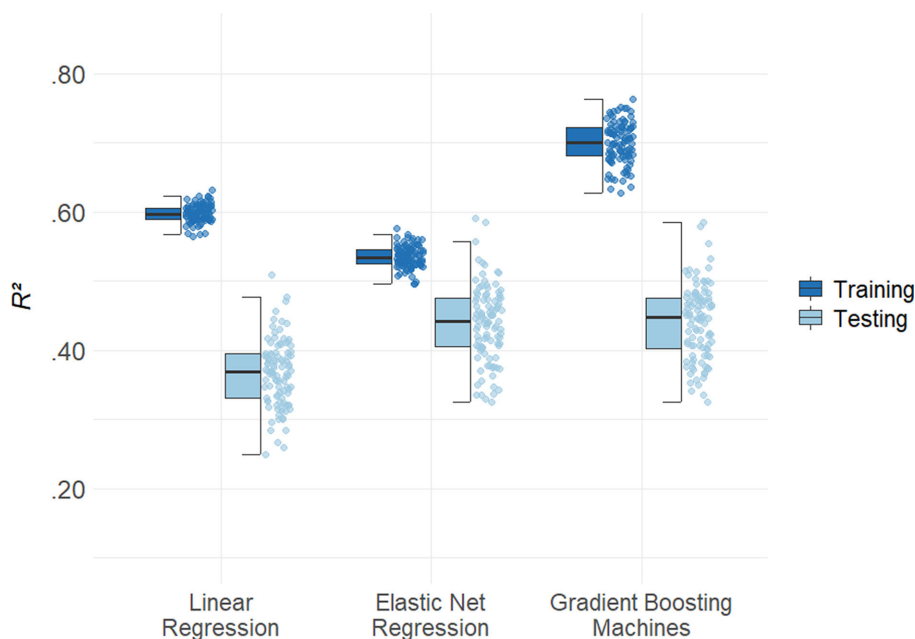
Figure 1 shows the distribution of the 100 different  $R^2$  values for the three different model families using all available variables (Model 2–4), indicating a large variance across different data splits and thus, underlining the need to use outer cross-validation to obtain reliable estimates for model performance using unseen testing data. Whereas the linear regression models on average explained an increment of 4.8% of the outcome's variance over the baseline model, both machine learning algorithms outperformed the linear regressions with an increment of 12.0% (elastic net regression)

TABLE 2 Predictive performances in training and test data for the prediction of treatment response.

No	Algorithm	Predictors	Train			Test		
			$R^2$	RMSE	MAE	$R^2$	RMSE	MAE
0	LinReg	Naïve guessing model		.464	.377		.455	.365
1	LinReg	PHQ-ADS at baseline	.329	.380	.302	.319	.376	.298
2	LinReg	All available variables	.598	.294	.233	.367	.374	.296
3	Enet	All available variables	.535	.320	.252	.439	.342	.268
4	GBM	All available variables	.700	.259	.205	.441	.342	.269
5	Enet	Demographics	.351	.375	.299	.330	.373	.296
6	Enet	Physical	.333	.340	.303	.315	.377	.299
7	Enet	Mental	.449	.347	.275	.370	.362	.286
8	Enet	Treatment	.433	.350	.275	.401	.354	.277
9	Enet	Demographics, Physical	.356	.373	.298	.328	.374	.296
10	Enet	Demographics, Mental	.452	.346	.275	.371	.362	.286
11	Enet	Demographics, Treatment	.447	.346	.273	.403	.352	.277
12	Enet	Physical, Mental	.448	.347	.276	.368	.363	.286
13	Enet	Physical, Treatment	.438	.349	.274	.396	.355	.278
14	Enet	Mental, Treatment	.532	.320	.252	.441	.341	.268
15	Enet	Demographics, Physical, Mental	.451	.347	.275	.370	.362	.286
16	Enet	Demographics, Physical, Treatment	.452	.345	.272	.402	.353	.277
17	Enet	Demographics, Mental, Treatment	.533	.320	.252	.440	.342	.268
18	Enet	Physical, Mental, Treatment	.534	.319	.252	.440	.342	.268

Abbreviations: Enet, elastic net regression; GBM, gradient boosting machines; Linreg, linear regression; MAE, Mean Absolute Error; RMSE, Root Mean Square Error.



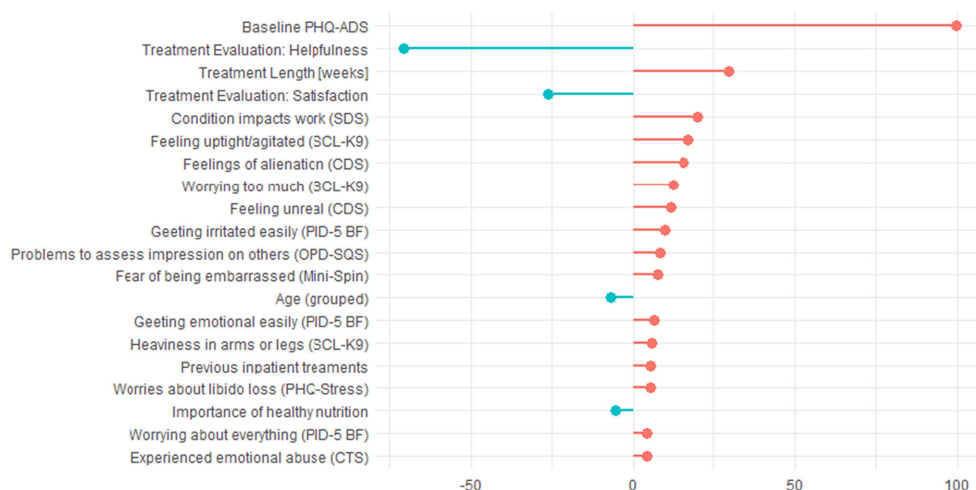


**FIGURE 1** Explained variance for all 100 iterations of the outer validation loop for the three models using all available predictor variables. *Note:* The box reflects the interquartile range (IQR), the solid line the median and the whiskers 1.5 times the IQR across 100 iterations (i.e., outer loop data splits).  $R^2$  values of the 100 models are displayed as a jittered distribution on the right.

and 12.2% (GBM) over the baseline model. Since the two machine learning models using all available predictor variables were nearly identical in predictive performances within the test data, we conducted all following variable group comparisons with both algorithms to provide interested readers with a full result set.

In [Table 2](#), we present the results using elastic net regressions since those show a considerably less amount of overfit (i.e., difference between training and testing performance) across the different models (for the results of Model 5–18 using GBM, see [Table S2](#)). The rank order of the performances of the different models with respect to variables included in the models was similar across the two algorithms so that conclusions about the importance of variable groupings were identical. The main findings were as follows: First, demographics and indicators of physical health seem negligible (see [Table S2](#)). Second, out of the models using a single predictor set (Model 5–8), the model with the treatment variables performed best, followed by indicators of mental health. Third, out of all possible combinations, the model using treatment and mental health variables (Model 14) had the highest predictive performance in the test data. Even the models with additional variables (Models 15–18 and Model 3) showed no further improvement, which can be attributed to less overfit of the more parsimonious model.

As preregistered, we provide a detailed overview of variable importances for all models in the online supplement (see <https://osf.io/gnzqs>), but present and discuss only the most predictive model within the manuscript. Since variable importance measures are based on the training data and the GBM models overfitted significantly more than the elastic net regressions (at similar performance in the testing data), we will focus on the latter for the most accurate estimations. [Figure 2](#) shows the 20 most important variables for Model 3 which overall confirm the results of the previous model comparisons: The baseline PHQ-ADS value was by far the most important variable, followed by whether the patients found the treatment helpful, treatment length and satisfaction with treatment. There was only one demographic variable (patients' age) and one indicator of physical health (healthy nutrition) among the 20 most important predictors. Importantly, the latter variable represents a self-reported indicator of health behaviour rather than an objective measure of physical health. Relevant indicators of mental health



**FIGURE 2** Twenty variables with the highest averaged variable importances for Model 3. *Note:* PHQ-ADS, Patient Health Questionnaire Anxiety and Depression Scale (Kroenke et al., 2016); SDS, Sheehan Disability Scale (Sheehan et al., 1983); SCL-K9, Symptom Checklist (Petrowski et al., 2019); CDS-2, Cambridge Depersonalization Scale 2 (Michal et al., 2010); PID-5, The Personality Inventory for DSM-5—Brief Form (Krueger et al., 2013); OPD-SQS, OPD-Structure Questionnaire Short (Ehrenthal et al., 2015); Mini-Spin, Mini-Social-Phobia-Inventory (Wiltink et al., 2017); PHQ-Stress, Patient Health Questionnaire Stress (Gräfe et al., 2004); CTS, Childhood Trauma Screener (Grabe et al., 2012).

included the self-rated level of functioning at work or school (measured by the Sheehan Disability Scale), depersonalization-derealization, items of the SCL9 (which is to be expected since they can be seen as an alternative measure of symptom load), a PID-5 item measuring whether patients are quickly annoyed by all sorts of things, aspects of social phobia (whether patients were afraid to be ashamed or feel dumb), level of structural integration (as indicated by an item of the OPD-SQS), remembered childhood emotional abuse and being worried about libido loss.

## DISCUSSION

With this registered report, we aimed to further the knowledge on predictors of treatment response in ecologically valid naturalistic inpatient samples, ultimately working towards the admittedly ambitious goal of reducing treatment failures and relapse rates by proactively screening for patients at risk. Our methodological comparisons showed that the prediction of treatment response can be enhanced by using machine learning algorithms, however, not due to the incorporation of nonlinear or interaction effects but rather by reducing the models' overfit via regularization as indicated by the equally good performance of the elastic net regressions and the GBM models using all available predictor variables. The regularized elastic net regressions had a higher predictive performance in independent testing samples than the non-regularized version. Hence, the results provide yet another argument for focusing on collecting high-quality data in large samples with reliable indicators for clinical prediction models instead of on more and more complex modelling approaches when aiming for generalizability and in turn, clinical usability.

### Important variables for the prediction of treatment response

With an  $R^2$  of .44, the overall predictive performance of the best-performing model in this study is comparable to or even slightly higher than the results of similar previous research using elastic net regressions for the prediction of treatment response. For example, Webb et al. (2020) were able to



explain 38% of variance in post-treatment depression scores. As it has been often found (Cuijpers et al., 2022), anxiety and depression symptoms at discharge (i.e., post-treatment PHQ-ADS values) were best explained by their baseline levels ( $R^2 = .32$ ) with an increment of  $\Delta R^2 = .12$  in predictive performance adding all other available variables. This increment has to be seen in light of the two-fold nature of symptom severity as a predictor of poorer prognosis and as a predictor of successful/positive change (Constantino et al., 2021). If modelled explicitly, the literature points to distinct response patterns associated with high baseline severity: Patients with high baseline severity who improve strongly are often additionally characterized by lower impairment in other domains or low risk-related behaviour (e.g., self-harm or externalizing symptoms; Uckelstam et al., 2019), underscoring the need of a multidimensional assessment of functioning. Several authors have called for assessing the complexity of mental disorders across interlinked domains of functioning to better explain the variability in their phenotypes and responses to treatment (Barton et al., 2017).

For example, patients' ratings of their condition impacting their workability emerged as an important predictor of treatment response. Occupational functioning is a relevant criterion typically rated alongside psychological and social functioning in the psychiatric global assessment of function (GAF; Aas, 2011). Interestingly, self-rated workability had a more pronounced impact compared to self-rated social functioning, and thus, might serve as a more important indicator of clinical severity. Patients reporting a higher number of previous inpatient treatments were also more likely to report higher symptom severity at discharge, with 'unsuccessful' treatments being an indicator of disorder chronicity (Fava et al., 1996; Taylor et al., 2012). High levels of depersonalization and derealization (DP/DR), which describe the phenomenon of feeling detached and alienated from the self and the environment, were also among the most predictive variables. DP/DR can be classified as a disorder but is also coded as a symptom of the dissociative subtype of posttraumatic stress disorder, the dissociative features of dissociative identity disorder, cannabis intoxication, borderline personality disorder and anxiety disorders according to DSM-5. DP/DR often takes a chronic course (Baker et al., 2003) and individuals with co-occurring DP/DR are also at higher risk for chronic courses of comorbid mental health disorders (Mula et al., 2007; Schlax et al., 2020). DP/DR is therefore understood as an indicator of disorder severity (Baker et al., 2003; Michal et al., 2011; Simeon et al., 2003) and has previously been associated with less favourable therapy courses across different mental health disorders (Bae et al., 2016; Kleindienst et al., 2016; Lyssenko et al., 2018). Our study shows that DP/DR is predictive of anxiety and depression severity at discharge in a sample of mixed psychosomatic inpatients. This is an important finding as DP/DR is often underdiagnosed and not likely to be part of the assessment in routine outcome monitoring (Michal & Beutel, 2009). Unfortunately, the literature on evidence-based treatment of DP/DR is still scarce (Wang et al., 2023).

The SCL-K9 is an alternative measure of symptom severity. In previous research, the general severity index (GSI) of the SCL-K-9 correlated highly with measures of anxiety and depression (Petrowski et al., 2019; Prinz et al., 2008). Within the context of the present investigation, items tapping into depression (worry), anxiety (tightness) and somatic symptoms (heaviness) were important predictor variables, pointing to the interrelatedness of somatic, anxious and depressive symptom experiences—also called the SAD triad (Löwe et al., 2008). Though typically captured as a symptom, worry is also understood as a trait component associated with proneness to experience negative emotions (Weiss & Deary, 2020). A tendency to worry 'about everything', was also among the predictive PID-5 items. Similar symptoms related to domains of cognitive-affective schemes and therefore overlapping with personality (e.g., shame, worry, irritability) represented another group of important predictor variables. Excessive worrying is a form of repetitive negative thinking (RNT). It is typically defined as uncontrollable negative thoughts regarding close or distant future events, while worries focused on the past are labelled rumination (Nolen-Hoeksema et al., 2008). RNT is thought to be implicated in the development and maintenance of anxiety and depressive disorders and pretreatment levels of RNT have previously been associated with worse treatment outcomes across different mental disorders (Bredemeier et al., 2020; Kertz et al., 2015; Sarter et al., 2021). Beyond worry, other relevant PID-5 items were high irritability and emotional instability. The three most relevant items from the PID-5 all stem from the negative

affectivity domain which is associated with depressivity (Gonçalves et al., 2022) but also closely overlaps with the construct of borderline personality disorder in empirical research (Gutiérrez et al., 2023). Relatedly, one item assessing personality functioning, namely the capacity to have a correct impression on how others might perceive oneself emerged as predictive. Adequate self-other functioning is the hallmark of personality disorders closely relating to mentalizing capacities (Wendt et al., 2023).

Only one demographic variable was important: Higher age was associated with better treatment response. We are unaware of inpatient studies with a similar finding. Some studies have pointed to comparable benefits across the age range (Cuijpers et al., 2018; Haigh et al., 2018). In our sample there was a preponderance of young patients. Population-based studies (Beutel et al., 2016) have shown that procrastination is particularly frequent in young people. This characteristic is likely to counteract symptom improvement but has so far been neglected in outcome studies of inpatient psychotherapy.

The group of variables that were most strongly associated with post-treatment PHQ-ADS values dealt with the treatment itself. This finding might seem rather straightforward because the self-report variables about the treatment are a direct, albeit subjective patients' evaluation of the overall treatment process and also the most temporal proximal predictors to the post-treatment outcome measure. Interestingly, whether patients found the treatment helpful was more predictive than mere satisfaction with the treatment, possibly underlining that wording matters in assessments of such subjective patients' evaluations (see also Ammerman et al., 2021 for wording effects in the context of self-harm). Also, this might hint to the fact that even momentarily dissatisfying (e.g., physically or mentally demanding) treatments do not necessarily lead to an impression of unhelpfulness. Treatment length and the number of previously undergone treatments were predictive of outcome in the sense that longer treatment and a higher number of treatment attempts were associated with more severe symptoms at discharge. Previous research indicates that the relationship between treatment duration and outcome is complex. On average, improvements of relationship patterns and personality functioning need more time than symptom improvement (Haase et al., 2008). On the contrary, the rate and magnitude of further change decline with an increasing duration of inpatient therapy (Liebherz & Rabung, 2014). Particularly complex cases unresponsive to previous outpatient or shorter inpatient treatments may require lengthy inpatient treatments followed by day hospital treatments which achieve overall comparably small benefits. In the case of the present study, it might be that highly complex cases received longer stays but were less likely to show vast change rates within their stay.

## Strengths and limitations

This study represents one of the first registered reports in the field of psychotherapy research using machine learning algorithms for predictive modelling. Open science practices are by no means a sign of high-quality research by themselves, but rather a prerequisite (Bakker et al., 2020). By using two-stage approaches of publication, the risk for a posteriori modification of the research rationale or the analyses based on the study results can be reduced or at least be made transparent. Thinking of a larger context, this could also be highly relevant to address any presumptions about or effects of researcher allegiance biases in psychotherapy research. Also, transparent methods and openly available data are necessary to enable meaningful aggregations of research findings across studies in the form of reviews or meta-analyses. We acknowledge that sharing raw data is not always possible in clinical science due to privacy concerns (as it was the case in the current study). However, the provision of a synthetic version of the data seems a useful compromise between reproducibility of the analyses as well as reuse of data on the one hand, and data protection and privacy concerns on the other hand. To provide a good practice example, we also created a synthetic dataset using the convenient R package *synthpop* (Nowok et al., 2016).

In this study, we used a large naturalistic inpatient sample with mixed diagnoses and modelled treatment response using easily available baseline indicators which can be equally seen as a strength and limitation. On the one hand, we aim for validity for inpatient treatments without exclusion of

patients, for example, those with recent suicide attempts or with multiple comorbid disorders, on the other hand, our results may not be immediately applied to the treatment of certain specific disorders. Additionally, using only baseline indicators can be seen at odds with findings on treatment response being affected by multiple, time-variant factors underlining the need for multi-modal and intensive longitudinal data (Chekroud et al., 2021). Nonetheless, prediction models such as the one in this study could be more easily implemented than models requiring further data acquisition since baseline assessments are carried out in German psychotherapy clinics on a routine basis and they provide valuable information on the risk of treatment nonresponse at a time when intervention is still possible.

One limiting factor that our study shares with similar research is a specific, fixed set of available predictor variables. In any case, there are further variables (e.g., patients' motivation to participate in the therapy at baseline; Jankowsky et al., 2023) that might incrementally explain treatment response. The present results have to be seen against this background: For example, we found that demographics had a negligible role in our models. However, there have been studies providing the first evidence for tailored treatments for minority patients, for example, queer patients (Bochicchio et al., 2022). Related information was not systematically assessed in our sample. Generally, the awareness of the topics of diversity and inclusion has increased in Germany only in recent years (Kluge et al., 2020) and studies investigating the experiences and needs of minority groups in the mental health care system are needed.

Due to our study design, we cannot make statements about patients' long-term treatment responses since we predicted post-treatment scores that were assessed directly at discharge. Previous research has shown that symptom severity at follow-ups can differ strongly from these assessments (Steinert et al., 2014). Thus, a worthwhile endeavour for further research would be to examine to what extent predictive models using outcomes at discharge still hold when tested at a later point in time. If this were not the case, one could argue that long-term response is the clinically more relevant outcome and should be used to train prediction models, thereby additionally providing more information on who relapses and why, which could then inform clinical decisions about relapse prevention.

## CONCLUSION

In this registered report, we demonstrated that it is possible and worthwhile to combine rigorous open science practices with the analytical flexibility of complex machine learning algorithms for the prediction of treatment response. It was possible to predict treatment response to a considerable degree, taking advantage of regularization approaches inherent to the algorithms that were used. Our results again underline the large association between baseline and post-treatment symptoms; however, they also show the importance of a multidimensional assessment of functioning and identify possible prognostic markers. Our results highlight the importance of negative affectivity and self-other regulatory capacities related to depression and anxiety symptoms but also of symptoms such as depersonalization and derealization that have not been focused on in previous research.

## AUTHOR CONTRIBUTIONS

**Kristin Jankowsky:** Conceptualization; formal analysis; methodology; visualization; writing – original draft; writing – review and editing. **Lina Krakau:** Conceptualization; data curation; writing – original draft; writing – review and editing. **Ulrich Schroeders:** Methodology; supervision; validation; writing – review and editing. **Rüdiger Zwerenz:** Data curation; project administration; resources; writing – review and editing. **Manfred E. Beutel:** Data curation; project administration; resources; writing – review and editing.

## ACKNOWLEDGEMENTS

Open Access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

## DATA AVAILABILITY STATEMENT

The patient data are not publicly available due to privacy restrictions. For data access upon request, please contact Lina Krakau at [lina.krakau@unimedizin-mainz.de](mailto:lina.krakau@unimedizin-mainz.de). We present all predictor variables, sample descriptives as well as a correlation matrix and a synthetic dataset at <https://osf.io/jxst4/>.

## ORCID

Kristin Jankowsky  <https://orcid.org/0000-0002-4847-0760>

## REFERENCES

- 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>
- Aas, I. M. (2011). Guidelines for rating global assessment of functioning (GAF). *Annals of General Psychiatry, 10*(1), 2. <https://doi.org/10.1186/1744-859X-10-2>
- Altmann, U., Gawlytta, R., Hoyer, J., Leichsenring, F., Leibing, E., Beutel, M., Willutzki, U., Herpertz, S., & Strauss, B. (2020). Typical symptom change patterns and their predictors in patients with social anxiety disorder: A latent class analysis. *Journal of Anxiety Disorders, 71*, 102200. <https://doi.org/10.1016/j.janxdis.2020.102200>
- Altmann, U., Steyer, R., Kramer, D., Steffanowski, A., Wittmann, W. W., von Heymann, F., Auch-Dorsch, E., Bruckmayer, E., Pfaffinger, I., Fembacher, A., & Strauß, B. (2015). Verlaufsmuster depressiver Störungen bei ambulanten psychotherapeutischen Behandlungen und deren Vorhersage [typical patterns of depressive disorders during outpatient psychotherapy and their prediction]. *Zeitschrift für Psychosomatische Medizin und Psychotherapie, 61*(2), 156–172. <https://doi.org/10.13109/zptm.2015.61.2.156>
- Ammerman, B. A., Burke, T. A., Jacobucci, R., & McClure, K. (2021). How we ask matters: The impact of question wording in single-item measurement of suicidal thoughts and behaviors. *Preventive Medicine, 152*, Article 106472. <https://doi.org/10.1016/j.ypmed.2021.106472>
- Bae, H., Kim, D., & Park, Y. C. (2016). Dissociation predicts treatment response in eye-movement desensitization and reprocessing for posttraumatic stress disorder. *Journal of Trauma & Dissociation, 17*(1), 112–130. <https://doi.org/10.1080/15299732.2015.1037039>
- Baker, D., Hunter, E., Lawrence, E., Medford, N., Patel, M., Senior, C., Sierra, M., Lambert, M. V., Phillips, M. L., & David, A. S. (2003). Depersonalisation disorder: Clinical features of 204 cases. *The British Journal of Psychiatry: the Journal of Mental Science, 182*, 428–433.
- Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLoS Biology, 18*(12), e3000937. <https://doi.org/10.1371/journal.pbio.3000937>
- Barton, S., Armstrong, P., Wicks, L., Freeman, E., & Meyer, T. D. (2017). Treating complex depression with cognitive behavioural therapy. *The Cognitive Behaviour Therapist, 10*, e17. <https://doi.org/10.1017/S1754470X17000149>
- Beutel, M. E., & Bleichner, F. (2011). Inpatient psychosomatic treatment of anxiety disorders: Comorbidities, predictors, and outcomes. *International Journal of Clinical and Health Psychology, 11*(3), 443–457.
- Beutel, M. E., Klein, E. M., Aufenanger, S., Brähler, E., Dreier, M., Müller, K. W., Quiring, O., Reinecke, L., Schmutzer, G., Stark, B., & Wölfling, K. (2016). Procrastination, distress and life satisfaction across the age range – A German representative community study. *PLoS One, 11*(2), e0148054. <https://doi.org/10.1371/journal.pone.0148054>
- Beutel, M. E., Michal, M., & Subic-Wrana, C. (2008). Psychoanalytically-oriented inpatient psychotherapy of somatoform disorders. *The Journal of the American Academy of Psychoanalysis and Dynamic Psychiatry, 36*(1), 125–142. <https://doi.org/10.1521/jaap.2008.36.1.125>
- Bischi, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation, 20*(2), 249–275. [https://doi.org/10.1162/EVCO\\_a\\_00069](https://doi.org/10.1162/EVCO_a_00069)
- Bohicchio, L., Reeder, K., Ivanoff, A., Pope, H., & Stefancic, A. (2022). Psychotherapeutic interventions for LGBTQ + youth: A systematic review. *Journal of LGBT Youth, 19*(2), 152–179. <https://doi.org/10.1080/19361653.2020.1766393>
- Bredemeier, K., Lieblich, S., & Foa, E. B. (2020). Pretreatment levels of rumination predict cognitive-behavioral therapy outcomes in a transdiagnostic sample of adults with anxiety-related disorders. *Journal of Anxiety Disorders, 75*, 102277. <https://doi.org/10.1016/j.janxdis.2020.102277>
- Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry, 9*(1), 271. <https://doi.org/10.1038/s41398-019-0607-2>
- Chekrou, 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>

- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>
- Constantino, M. J., Boswell, J. F., & Coyne, A. E. (2021). Patient, therapist, and relational factors. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Handbook of psychotherapy and behavior change* (7th ed., pp. 225–262). Wiley.
- Costello, C., Srivastava, S., Rejaie, R., & Zalewski, M. (2021). Predicting mental health from followed accounts on twitter. *Collabra: Psychology*, *7*(1), Article 18731. <https://doi.org/10.1525/collabra.18731>
- Cuijpers, P., Ciharova, M., Quero, S., Miguel, C., Driessen, E., Harrer, M., Purgato, M., Ebert, D., & Karyotaki, E. (2022). The contribution of “individual participant data” meta-analyses of psychotherapies for depression to the development of personalized treatments: A systematic review. *Journal of Personalized Medicine*, *12*(1), 93. <https://doi.org/10.3390/jpm12010093>
- Cuijpers, P., Karyotaki, E., Reijnders, M., & Huibers, M. J. H. (2018). Who benefits from psychotherapies for adult depression? A meta-analytic update of the evidence. *Cognitive Behaviour Therapy*, *47*(2), 91–106. <https://doi.org/10.1080/16506073.2017.1420098>
- Delgadillo, J. (2021). Machine learning: A primer for psychotherapy researchers. *Psychotherapy Research*, *31*(1), 1–4. <https://doi.org/10.1080/10503307.2020.1859638>
- Delgadillo, J., & Lutz, W. (2020). A development pathway towards precision mental health care. *JAMA Psychiatry*, *77*(9), 889–890. <https://doi.org/10.1001/jamapsychiatry.2020.1048>
- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. <https://doi.org/10.1146/annurev-clinpsy-032816-045037>
- Ehrental, J. C., Dinger, U., Schauenburg, H., Horsch, L., Dahlbender, R. W., & Gierk, B. (2015). Entwicklung einer Zwölf-item-version des OPD-Strukturfragebogens (OPD-SFK) [development of a 12-item version of the OPD-structure questionnaire (OPD-SQS)]. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, *61*(3), 262–274. <https://doi.org/10.13109/zptm.2015.61.3.262>
- Fava, M., Alpert, J. E., Borus, J. F., Nierenberg, A. A., Pava, J. A., & Rosenbaum, J. F. (1996). Patterns of personality disorder comorbidity in early-onset versus late-onset major depression. *American Journal of Psychiatry*, *153*(10), 1308–1312. <https://doi.org/10.1176/ajp.153.10.1308>
- Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. <https://doi.org/10.1037/a0024445>
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. <https://doi.org/10.1177/1745691612459059>
- Fisher, D. J., Carpenter, J. R., Morris, T. P., Freeman, S. C., & Tierney, J. F. (2017). Meta-analytical methods to identify who benefits most from treatments: Daft, deluded, or deft approach? *BMJ*, *356*, Article j573. <https://doi.org/10.1136/bmj.j573>
- Gonçalves, B., Pires, R., Henriques-Calado, J., & Sousa Ferreira, A. (2022). Evaluation of the PID-5 depressivity personality dimensions and depressive symptomatology in a community sample. *Anales de Psicología*, *38*(3), 409–418. <https://doi.org/10.6018/analesps.486921>
- Grabe, H., Schulz, A., Schmidt, C., Appel, K., Driessen, M., Wingenfeld, K., Barnow, S., Spitzer, C., John, U., Berger, K., Wersching, H., & Freyberger, H. (2012). Ein Screeninginstrument für Missbrauch und Vernachlässigung in der Kindheit: Der Childhood Trauma Screener (CTS). *Psychiatrische Praxis*, *39*(3), 109–115. <https://doi.org/10.1055/s-0031-1298984>
- Gräfe, K., Zipfel, S., Herzog, W., & Löwe, B. (2004). Screening psychischer Störungen mit dem “Gesundheitsfragebogen für Patienten (PHQ-D)”. *Diagnostica*, *50*(4), 171–181. <https://doi.org/10.1026/0012-1924.50.4.171>
- Gutiérrez, F., Aluja, A., Ruiz Rodríguez, J., Peri, J. M., Gárriz, M., García, L. F., Sorrel, M. A., Sureda, B., Vall, G., Ferrer, M., & Calvo, N. (2023). Borderline, where are you? A psychometric approach to the personality domains in the international classification of diseases, 11th revision (ICD-11). *Personality Disorders, Theory, Research, and Treatment*, *14*(3), 355–359. <https://doi.org/10.1037/per0000592>
- Haase, M., Frommer, J., Franke, G. H., Hoffmann, T., Schulze-Muetzel, J., Jäger, S., Grabe, H. J., Spitzer, C., & Schmitz, N. (2008). From symptom relief to interpersonal change: Treatment outcome and effectiveness in inpatient psychotherapy. *Psychotherapy Research*, *18*(5), 615–624. <https://doi.org/10.1080/10503300802192158>
- Haigh, E. A. P., Bogucki, O. E., Sigmon, S. T., & Blazer, D. G. (2018). Depression among older adults: A 20-year update on five common myths and misconceptions. *The American Journal of Geriatric Psychiatry*, *26*(1), 107–122. <https://doi.org/10.1016/j.jagp.2017.06.011>
- Hartmann, A., von Wietersheim, J., Weiss, H., & Zeeck, A. (2018). Patterns of symptom change in major depression: Classification and clustering of long term courses. *Psychiatry Research*, *267*, 480–489. <https://doi.org/10.1016/j.psychres.2018.03.086>
- Hilbert, K., Jacobi, T., Kunas, S. L., Elsner, B., Reuter, B., Lueken, U., & Kathmann, N. (2021). Identifying CBT non-response among OCD outpatients: A machine-learning approach. *Psychotherapy Research*, *31*(1), 52–62. <https://doi.org/10.1080/10503307.2020.1839140>
- 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. <https://doi.org/10.1145/3514094.3534196>



- Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of inflated prediction performance: A commentary on machine learning and suicide research. *Clinical Psychological Science, 9*(1), 129–134. <https://doi.org/10.1177/2167702620954216>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning*. Springer.
- Jankowsky, K., Zimmermann, J., Jaeger, U., Mestel, R., & Schroeders, U. (2023). First impressions count: Therapists' impression on patients' motivation and helping alliance predicts psychotherapy dropout. <https://doi.org/10.31234/osf.io/nhs6c>
- Jolly, E., & Chang, L. J. (2019). The flatland fallacy: Moving beyond low-dimensional thinking. *Topics in Cognitive Science, 11*(2), 433–454. <https://doi.org/10.1111/tops.12404>
- Kaiser, T., Volkmann, C., Volkmann, A., Karyotaki, E., Cuijpers, P., & Brakemeier, E.-L. (2022). Heterogeneity of treatment effects in trials on psychotherapy of depression. *Clinical Psychology: Science and Practice, 29*, 294–303. <https://doi.org/10.1037/cps0000079>
- Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science (arXiv:2207.07048). arXiv. <http://arxiv.org/abs/2207.07048>
- Kertz, S. J., Koran, J., Stevens, K. T., & Björgvinsson, T. (2015). Repetitive negative thinking predicts depression and anxiety symptom improvement during brief cognitive behavioral therapy. *Behaviour Research and Therapy, 68*, 54–63. <https://doi.org/10.1016/j.brat.2015.03.006>
- Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2020). Suicide prediction models: A critical review of recent research with recommendations for the way forward. *Molecular Psychiatry, 25*(1), 168–179. <https://doi.org/10.1038/s41380-019-0531-0>
- Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., de Jonge, P., Nierenberg, A. A., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences, 26*(1), 22–36. <https://doi.org/10.1017/S2045796016000020>
- Kleindienst, N., Priebe, K., Görg, N., Dyer, A., Steil, R., Lyssenko, L., Winter, D., Schmahl, C., & Bohus, M. (2016). State dissociation moderates response to dialectical behavior therapy for posttraumatic stress disorder in women with and without borderline personality disorder. *European Journal of Psychotraumatology, 7*(1), 30375. <https://doi.org/10.3402/ejpt.v7.30375>
- Kluge, U., Aichberger, M. C., Heinz, E., Udeogu-Gözalán, C., & Abdel-Fatah, D. (2020). Rassismus und psychische Gesundheit. *Der Nervenarzt, 91*(11), 1017–1024. <https://doi.org/10.1007/s00115-020-00990-1>
- Krause, M., & Behn, A. (2021). Depression and personality dysfunction: Towards the understanding of complex depression. In G. de la Parra, P. Dagnino, & A. Behn (Eds.), *Depression and personality dysfunction* (pp. 1–13). Springer International Publishing. [https://doi.org/10.1007/978-3-030-70699-9\\_1](https://doi.org/10.1007/978-3-030-70699-9_1)
- Kroenke, K., Wu, J., Yu, Z., Bair, M. J., Kean, J., Stump, T., & Monahan, P. O. (2016). Patient health questionnaire anxiety and depression scale: Initial validation in three clinical trials. *Psychosomatic Medicine, 78*(6), 716–727. <https://doi.org/10.1097/PSY.0000000000000322>
- Krueger, R., Derringer, J., Markon, K., Watson, D., & Skodol, A. (2013). *The personality inventory for DSM-5—Brief form (PID-5-BF) adult*. American Psychiatric Association.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Lee, Y., Ragugett, 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>
- Liebherz, S., & Rabung, S. (2014). Do patients' symptoms and interpersonal problems improve in psychotherapeutic hospital treatment in Germany? – A systematic review and meta-analysis. *PLoS One, 9*(8), e105329. <https://doi.org/10.1371/journal.pone.0105329>
- Löwe, B., Spitzer, R. L., Williams, J. B. W., Mussell, M., Schellberg, D., & Kroenke, K. (2008). Depression, anxiety and somatization in primary care: Syndrome overlap and functional impairment. *General Hospital Psychiatry, 30*(3), 191–199. <https://doi.org/10.1016/j.genhosppsych.2008.01.001>
- Luborsky, L., Auerbach, A. H., Chandler, M., Cohen, J., & Bachrach, H. M. (1971). Factors influencing the outcome of psychotherapy: A review of quantitative research. *Psychological Bulletin, 75*(3), 145–185. <https://doi.org/10.1037/h0030480>
- Lutz, W., de Jong, K., Rubel, J. A., & Delgadillo, J. (2021). Measuring, predicting, and tracking change in psychotherapy. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Handbook of psychotherapy and behavior change* (7. Aufl., S. 89–134). Wiley.
- 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>
- Lyssenko, L., Schmahl, C., Bockhacker, L., Vonderlin, R., Bohus, M., & Kleindienst, N. (2018). Dissociation in psychiatric disorders: A meta-analysis of studies using the dissociative experiences scale. *American Journal of Psychiatry, 175*(1), 37–46. <https://doi.org/10.1176/appi.ajp.2017.17010025>

- Maj, M., Stein, D. J., Parker, G., Zimmerman, M., Fava, G. A., De Hert, M., Demyttenaere, K., McIntyre, R. S., Widiger, T., & Wittchen, H. (2020). The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry, 19*(3), 269–293. <https://doi.org/10.1002/wps.20771>
- McClure, K., Jacobucci, R., & Ammerman, B. A. (2021). Are items more than indicators? An examination of psychometric homogeneity, item-specific effects, and consequences for structural equation models. <https://doi.org/10.31234/osf.io/n4mxv>
- McNamara, M. E., Zisser, M., Beevers, C. G., & Shumake, J. (2022). Not just “big” data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions. *Behaviour Research and Therapy, 153*, 104086. <https://doi.org/10.1016/j.brat.2022.104086>
- Michal, M., & Beutel, M. E. (2009). Weiterbildung CME: Depersonalisation/Derealisation – Krankheitsbild, Diagnostik und Therapie. *Zeitschrift für Psychosomatische Medizin und Psychotherapie, 55*(2), 113–140. <https://doi.org/10.13109/zptm.2009.55.2.113>
- Michal, M., Wiltink, J., Till, Y., Wild, P. S., Blettner, M., & Beutel, M. E. (2011). Distinctiveness and overlap of depersonalization with anxiety and depression in a community sample: Results from the Gutenberg heart study. *Psychiatry Research, 188*(2), 264–268. <https://doi.org/10.1016/j.psychres.2010.11.004>
- Michal, M., Zwerenz, R., Tschan, R., Edinger, J., Lichy, M., Knebel, A., Tuin, I., & Beutel, M. (2010). Screening nach depersonalisation-Derealisation mittels zweier items der Cambridge depersonalisation scale. *Psychotherapie · Psychosomatik · Medizinische Psychologie, 60*(5), 175–179. <https://doi.org/10.1055/s-0029-1224098>
- Mula, M., Pini, S., & Cassano, G. B. (2007). The neurobiology and clinical significance of depersonalization in mood and anxiety disorders: A critical reappraisal. *Journal of Affective Disorders, 99*(1–3), 91–99. <https://doi.org/10.1016/j.jad.2006.08.025>
- Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science, 3*(5), 400–424. <https://doi.org/10.1111/j.1745-6924.2008.00088.x>
- Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software, 74*(11), 1–26. <https://doi.org/10.18637/jss.v074.i11>
- O'Hara, R., Beaudreau, S. A., Gould, C. E., Froehlich, W., & Kraemer, H. C. (2017). Handling clinical comorbidity in randomized clinical trials in psychiatry. *Journal of Psychiatric Research, 86*, 26–33. <https://doi.org/10.1016/j.jpsychires.2016.11.006>
- Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science, 6*(3). <https://doi.org/10.1177/25152459231162559>
- Petrowski, K., Schmalbach, B., Kliem, S., Hinz, A., & Brähler, E. (2019). Symptom-checklist-K-9: Norm values and factorial structure in a representative German sample. *PLoS One, 14*(4), e0213490. <https://doi.org/10.1371/journal.pone.0213490>
- Philips, B., & Falkenström, F. (2021). What research evidence is valid for psychotherapy research? *Frontiers in Psychiatry, 11*, 625380. <https://doi.org/10.3389/fpsy.2020.625380>
- Prinz, U., Nutzinger, D., Schulz, H., Petermann, F., Braukhaus, C., & Andreas, S. (2008). Die Symptom-Checkliste-90-R und ihre Kurzversionen: Psychometrische Analysen bei Patienten mit psychischen Erkrankungen. *Physikalische Medizin, Rehabilitationsmedizin, Kurortmedizin, 18*(6), 337–343. <https://doi.org/10.1055/s-0028-1093323>
- Sarter, L., Heider, J., Kirchner, L., Schenkel, S., Witthöft, M., Rief, W., & Kleinstäuber, M. (2021). Cognitive and emotional variables predicting treatment outcome of cognitive behavior therapies for patients with medically unexplained symptoms: A meta-analysis. *Journal of Psychosomatic Research, 146*, 110486. <https://doi.org/10.1016/j.jpsychores.2021.110486>
- Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science, 4*(2), 1–12. <https://doi.org/10.1177/25152459211007467>
- Schlag, J., Wiltink, J., Beutel, M. E., Münzel, T., Pfeiffer, N., Wild, P., Blettner, M., Ghaemi Kerahrodi, J., & Michal, M. (2020). Symptoms of depersonalization/derealization are independent risk factors for the development or persistence of psychological distress in the general population: Results from the Gutenberg health study. *Journal of Affective Disorders, 273*, 41–47. <https://doi.org/10.1016/j.jad.2020.04.018>
- Seeboth, A., & Möttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality, 32*(3), 186–201. <https://doi.org/10.1002/per.2147>
- Senger, K., Rubel, J. A., Kleinstäuber, M., Schröder, A., Köck, K., Lambert, M. J., Lutz, W., & Heider, J. (2021). Symptom change trajectories in patients with persistent somatic symptoms and their association to long-term treatment outcome. *Psychotherapy Research, 1–16*, 624–639. <https://doi.org/10.1080/10503307.2021.1993376>
- Sheehan, D. V., Giddens, M. J. M., & Lisensi, P. P. (1983). Sheehan disability scale (SDS). *International Clinical Psychopharmacology, 11*, 89–95.
- Siddaway, A. P., Quinlivan, L., Kapur, N., O'Connor, R. C., & de Beurs, D. (2020). Cautions, concerns, and future directions for using machine learning in relation to mental health problems and clinical and forensic risks: A brief comment on “model complexity improves the prediction of nonsuicidal self-injury” (Fox et al., 2019). *Journal of Consulting and Clinical Psychology, 88*(4), 384–387. <https://doi.org/10.1037/ccp0000485>

- Simeon, D., Knutelska, M., Nelson, D., & Guralnik, O. (2003). Feeling unreal: A depersonalization disorder update of 117 cases. *The Journal of Clinical Psychiatry, 64*(9), 990–997. <https://doi.org/10.4088/JCP.v64n0903>
- Steinert, C., Hofmann, M., Kruse, J., & Leichsenring, F. (2014). Relapse rates after psychotherapy for depression – Stable long-term effects? A meta-analysis. *Journal of Affective Disorders, 168*, 107–118. <https://doi.org/10.1016/j.jad.2014.06.043>
- Taylor, D., Carlyle, J., McPherson, S., Rost, F., Thomas, R., & Fonagy, P. (2012). Tavistock adult depression study (TADS): A randomised controlled trial of psychoanalytic psychotherapy for treatment-resistant/treatment-refractory forms of depression. *BMC Psychiatry, 12*(1), 60. <https://doi.org/10.1186/1471-244X-12-60>
- Uckelstam, C.-J., Philips, B., Holmqvist, R., & Falkenström, F. (2019). Prediction of treatment outcome in psychotherapy by patient initial symptom distress profiles. *Journal of Counseling Psychology, 66*(6), 736–746. <https://doi.org/10.1037/cou0000345>
- Urban, C. J., & Gates, K. M. (2021). Deep learning: A primer for psychologists. *Psychological Methods, 26*(6), 743–773. <https://doi.org/10.1037/met0000374>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>
- Vazire, S. (2020). A toast to the error detectors. *Nature, 577*(7788), 9. <https://doi.org/10.1038/d41586-019-03909-2>
- Wang, S., Zheng, S., Zhang, X., Ma, R., Feng, S., Song, M., Zhu, H., & Jia, H. (2023). The treatment of depersonalization-derealization disorder: A systematic review. *Journal of Trauma & Dissociation, 1–24*. <https://doi.org/10.1080/15299732.2023.2231920>
- Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology, 88*(1), 25–38. <https://doi.org/10.1037/ccp0000451>
- Weiss, A., & Deary, I. J. (2020). A new look at neuroticism: Should we worry so much about worrying? *Current Directions in Psychological Science, 29*(1), 92–101. <https://doi.org/10.1177/0963721419887184>
- Wendt, L. P., Müller, S., & Zimmermann, J. (2023). Development and validation of the certainty about mental states questionnaire (CAMSQ): A self-report measure of mentalizing oneself and others. *Assessment, 30*(3), 651–674. <https://doi.org/10.1177/10731911211061280>
- Wilkinson, J., Arnold, K. F., Murray, E. J., van Smeden, M., Carr, K., Sippy, R., de Kamps, M., Beam, A., Konigorski, S., Lippert, C., Gilthorpe, M. S., & Tennant, P. W. G. (2020). Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health, 2*(12), e677–e680. [https://doi.org/10.1016/S2589-7500\(20\)30200-4](https://doi.org/10.1016/S2589-7500(20)30200-4)
- Wiltink, J., Kliem, S., Michal, M., Subic-Wrana, C., Reiner, I., Beutel, M. E., Brähler, E., & Zwerenz, R. (2017). Mini – social phobia inventory (mini-SPIN): Psychometric properties and population based norms of the German version. *BMC Psychiatry, 17*(1), 377. <https://doi.org/10.1186/s12888-017-1545-2>
- 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>
- Zeeck, A., von Wietersheim, J., Weiss, H., Beutel, M., & Hartmann, A. (2013). The INDDEP study: Inpatient and day hospital treatment for depression – Symptom course and predictors of change. *BMC Psychiatry, 13*(1), 100. <https://doi.org/10.1186/1471-244X-13-100>
- Zeeck, A., von Wietersheim, J., Weiß, H., Eduard Scheidt, C., Völker, A., Helesic, A., Eckhardt-Henn, A., Beutel, M., Endorf, K., Knoblauch, J., Rochlitz, P., & Hartmann, A. (2015). Symptom course in inpatient and day clinic treatment of depression: Results from the INDDEP-study. *Journal of Affective Disorders, 187*, 35–44. <https://doi.org/10.1016/j.jad.2015.07.025>
- Zeeck, A., von Wietersheim, J., Weiss, H., Hermann, S., Endorf, K., Lau, I., & Hartmann, A. (2020). Self-criticism and personality functioning predict patterns of symptom change in major depressive disorder. *Frontiers in Psychiatry, 11*, 147. <https://doi.org/10.3389/fpsy.2020.00147>
- Zeeck, A., von Wietersheim, J., Weiss, H., Scheidt, C. E., Völker, A., Helesic, A., Eckhardt-Henn, A., Beutel, M., Endorf, K., Treiber, F., Rochlitz, P., & Hartmann, A. (2016). Prognostic and prescriptive predictors of improvement in a naturalistic study on inpatient and day hospital treatment of depression. *Journal of Affective Disorders, 197*, 205–214. <https://doi.org/10.1016/j.jad.2016.03.039>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 67*(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Jankowsky, K., Krakau, L., Schroeders, U., Zwerenz, R., & Beutel, M. E. (2023). Predicting treatment response using machine learning: A registered report. *British Journal of Clinical Psychology, 00*, 1–19. <https://doi.org/10.1111/bjc.12452>



# Epilogue

In the following, I will summarize the main results of the four manuscripts included in this thesis. In each manuscript, we used ML algorithms to predict future behavior, but also focused on specific methodological issues that are currently debated in the field of predictive modeling: In manuscript 1, we proposed a temporal model validation approach to address potential information leakage in temporal dependent data. In manuscript 2, we addressed the ongoing debate about whether increasingly complex ML algorithms should be used to predict suicidal behavior, and under what circumstances the models can be translated into clinical practice. Manuscript 3 discussed the difficulty of predicting rare events and in manuscript 4, we argued for and highlighted the potential of registered reports to improve the transparency and reproducibility of ML-based research. In the subsequent parts of the epilogue, I will discuss some aspects of predictive modeling more broadly, namely model generalizability, feature selection, and timing in ML-based models. Finally, I will provide an outlook on improvement opportunities for ML modeling in psychology.

### **Manuscript 1: Validation and generalizability of machine learning prediction models on attrition in longitudinal studies.**

In the first manuscript, we compared the accuracy of logistic regressions and GBM for the prediction of attrition in two longitudinal panels (MIDUS and pairfam) using baseline demographic, health, and personality indicators. To do so, we used two different model validation strategies which we termed “80/20” and “temporal 80/20”. The first corresponds to the very common way of splitting the data into training data (80%) and testing data (20%) but disregards the actual temporal aspect of the attrition modeling (i.e., the outcome for training and testing was taken from the same measurement wave). With the temporal approach, however, we validated the trained models using outcome data from the still-active participants of the 20% testing data in a subsequent measurement wave. In both validation schemes, there is a strict separation of training and testing data, or in other words, there is no information

leakage that biases the estimated model accuracies. It could be argued that the temporal validation is a more realistic and informative scenario, as it takes into account potential changes in the population over time (i.e., participants who permanently dropped out at a previous measurement wave). In addition, panel researchers usually are interested in predicting the unknown participation status in a future wave in order to target still-active at-risk participants and optimize retention strategies. Our results showed that attrition was predicted rather inaccurately regardless of panel study or algorithm (e.g., balanced accuracies of .61 or lower) and further decreased for models that were validated in a temporal framework. With respect to the previous findings on the prediction of study attrition which we discussed in the prologue of this thesis, this study underlines two important aspects: First, the issue of longitudinal study attrition is unlikely to be solved by the use of increasingly complex algorithms since their previously promising reported accuracies were largely due to information leakage between training and testing data. Second, researchers need to be aware that for predictive models to be useful for actually predicting future behavior, they have to be generalizable across measurement occasions (similar to the concept of stationarity in time series, i.e., the data generating process not changing across time). Unfortunately, this assumption is unlikely to be true for longitudinal attrition, as the population for which the predictive model has to hold is systematically altered with every subsequent measurement occasion, which may cause shifts in effects and variable importances.

**Manuscript 2: Predicting lifetime suicide attempts in a community sample of adolescents using machine learning algorithms.**

The second manuscript dealt with the prediction of self-reported lifetime suicide attempts in a large community sample of 17-year-old adolescents, comparing two ML algorithms (elastic net regressions and GBM) with a baseline logistic regression and combining a heterogeneous set of predictor variables (including, for example, indicators of

physical and mental health, demographics, personality, victimization, offenses, and emotions) across different timeframes (either using predictors from a prior measurement wave or from the same wave the outcome was taken from). Our results can be summarized in terms of three comparisons: First, both ML algorithms outperformed the highly overfitting logistic regression models irrespective of timeframe or predictor set (e.g., with an ABA of .85 for both ML algorithms vs. .73 for logistic regressions when using data from the same measurement wave). Second, predictions were less accurate for models using data from the measurement wave 3 years prior (ABA = .76 vs. .85 for the GBM models). And third, in contrast to the group of studies I presented in the prologue of this thesis that employed optimism bootstrapping as a model validation strategy, allowing for interaction or nonlinear effects did not improve predictive performance (as indicated by the equally accurate predictions of elastic net regressions and GBM). Overall, previous self-harm, mental health problems, and indicators of loneliness and low self-esteem were among the most important predictor variables. Shifts in variable importances across adolescence (i.e., with respect to variables related to sexuality, drug use, delinquency or victimization) suggested the importance of tailoring suicide prediction models to the specific developmental phase. With this study, on the one hand, we explored potential predictors of future suicide attempts in a young community sample, with the aim of helping to identify variables that could be included in a first stage of early screening to prevent the onset of suicidal behaviors in children and adolescents. On the other hand, our results also address the already ongoing debate on the use of increasingly complex ML algorithms to predict suicidal behaviors (e.g., Fox et al., 2019 and the commentary by Siddaway et al., 2020). This debate is fueled by reasonable arguments for both points of view: for example, some researchers call for more flexible and complex models to match the presumed complexity of suicidal thoughts and behaviors since prediction efforts using linear models and a limited number of predictors have been meta-analytically

shown to be disappointing so far (e.g., Franklin, 2017). Counter to this, a more skeptical concern is that ML models are more difficult to interpret and therefore less likely to be accepted or implemented by clinicians. Our results showing that proximal predictors lead to higher accuracy than distal ones suggest an additional aspect worth considering, namely that most of the studies modeling suicidal behavior do not measure relevant indicators on the fitting timescales.

**Manuscript 3: First impressions count: Therapists' impression on patients' motivation and helping alliance predicts psychotherapy dropout.**

In the third manuscript, we compared the accuracy of logistic regressions, elastic net regressions and GBM for the prediction of therapy dropout in two German inpatient psychotherapy clinics ( $N = 1,691$  in Sample 1 and  $N = 12,473$  in Sample 2) using baseline indicators (e.g., demographics or variables on previous treatments and symptom severity) collected on the first day of the patients' stay at the respective clinic. As in the previous manuscript, both ML algorithms generally outperformed the overfitting logistic regressions (to a greater extent in the smaller sample) and achieved comparable predictive accuracies (AUC of .74 and .82 using elastic net regressions vs. AUC of .72 and .83 using GBM for Samples 1 and 2, respectively). Key predictors of therapy dropout, and thus potential points of intervention to reduce dropout, were the therapists' initial assessment of patients' motivation and the therapeutic alliance. The results of the third manuscript included in this dissertation again underlined the usefulness of model regularization to improve the generalizability of predictive performances (i.e., to reduce overfitting), but also showed that more complex modeling is not necessarily a remedy for every prediction task. In additional analyses, we focused on sample size considerations for rare event classification tasks. We used the larger of the two samples to systematically reduce the total sample size (and thereby number of events) and the event fraction (i.e., the number of events relative to the total sample size) and found

that the higher predictive performance in Sample 2 was partly due to the larger sample size (reducing the sample size of Sample 2 to that of Sample 1 reduced the differences in predictive performance between the two samples). Importantly, reducing the event fraction had an even more pronounced effect on predictive accuracy than reducing the sample size (with a fixed event fraction). Our results are consistent with recent findings (e.g., Van Smeden et al., 2019) on the importance of considering more aspects than just the ratio of predictor variables to the number of events when building classification models. In other words, a highly unequal group size significantly increases the demands on the data and overfitting in classifications could be better reduced by additional sampling from the smaller group than by increasing the sample size in general.

#### **Manuscript 4: Predicting treatment response using machine learning: A Registered Report.**

In manuscript 4 of this thesis, we used different modeling approaches (unregularized and elastic net regression analyses and GBM) to predict treatment response as operationalized by the Patient Health Questionnaire Anxiety and Depression Scale in a naturalistic inpatient sample. For this, we used four sets of predictor variables (demographics, variables on physical and mental health, and treatment-related variables). Again, the two ML modeling approaches overfitted less than unregularized regressions and achieved higher, albeit similar, predictive performance. Beyond the substantial association between baseline and post-treatment symptom severity ( $R^2 = .319$ ), using all available predictors resulted in an increment of  $\Delta R^2 = .12$ , largely driven by treatment- and mental health related variables. One could argue that this increment is rather small compared to the effect of baseline symptom severity and that the increment could be inflated by including variables conceptually very similar to the Patient Health Questionnaire Anxiety and Depression Scale and variables assessed post-treatment into the model. However, our results also showed the importance of variables that are rarely

included in routine outcome monitoring, such as those assessing depersonalization and derealization. The fourth manuscript of this dissertation differs from the previous three in that it represents one of the first registered reports in psychological research (and especially psychotherapy research) using ML algorithms for predictive modeling. This entailed the prior specification of all analysis steps and the preparation of annotated R syntax, including details on ML modeling such as hyperparameter grids and model validation. We argue that common challenges to the reproducibility of complex algorithms can be tackled by full transparency about modeling decisions and by constructive methodological feedback from peers prior to dissemination of results and conclusions. By using registered reports and their inherent two-step approach to publication, researchers are less likely to exploit degrees of freedom (which become more numerous as models become more complex) to render results more surprising or predictive, as a potentially strong incentive to do so (i.e., the decision on the publication of a manuscript) becomes result-independent (i.e., based solely on the research question and proposed analyses).

### **There is no Such Thing as a Validated Prediction Model**

It is often stated that cross-validation allows researchers to quantify the generalizability of their predictive model (e.g., Rosenbusch et al., 2021; Song et al., 2021). However, even more sophisticated approaches, such as a nested resampling approach (Bischof et al., 2012; Pargent et al., 2023) with a large number of outer loop iterations, which we used in all empirical studies in this thesis, can only estimate a model's performance on unseen data from the same population. Hence, researchers implicitly assume either representativeness of their samples or non-heterogeneity and stationarity of data generating processes across settings and time when reporting predictive performance based on validation samples generated by random data splitting. Van Calster et al. (2023) argue in their article (the title of which inspired the subheading above) that the validation of prediction models is a process that

can never truly be completed. Rather than developing additional prediction models, researchers should focus their attention on more thorough validations (and updates thereof) of already promising models that account for heterogeneity in performance across contexts. Although the authors discuss clinical prediction models and specific examples may not be fully transferable to psychological research (e.g., the introduction of new surgical procedures that alter associations between variables), the general argument holds. Model performance can potentially be affected by time, setting, culture, measures, or methods, to name a few. In the following, I will elaborate on these different aspects, discuss approaches to testing and presenting generalizability of predictive models in psychological research and provide an additional empirical example examining model generalizability across cultures and methods.

### *Time*

The effect of time on model performance is (at least) threefold: First, it can be understood in a longitudinal intraindividual sense as shown for example in our study on the prediction of lifetime suicide attempts in adolescence, that is, as a shift in performance and variable importances across the lifespan. Second, there may also be generational shifts in interindividual associations resulting in models that are not timeless. National cohort studies with matched measures of core variables are already being used for intergenerational research (e.g., King et al., 2023; Parsons et al., 2021) and could also be helpful in quantifying the likely magnitude of such shifts. Third, changes that systematically affect the composition of a population over time—as in our study of attrition in longitudinal panels—may also fall into this category since a predictive model that was established at baseline of a longitudinal study decreased in performance at a later measurement occasion.

### *Settings*

Data in psychological research can be organized hierarchically, that is observations (e.g., students or patients) may be nested within groups (schools, clinics, or therapists),



making it necessary to ensure that models hold across these specific (group) settings. How this is done depends largely on the availability of data and the intended use of the prediction models. If data with many groups are already available and the goal is to obtain a model that generalizes across them, it is recommended to use specific forms of blocked cross-validation (e.g., Dragicevic & Casalicchio, 2020) that train predictive models to capture the overall common associations between variables. In other cases, a predictive model that captures group-specific effects (e.g., within a single clinic) may be more desirable. However, such models would need to be externally validated using data from different settings (e.g., by employing leave-one-cluster-out cross-validation approaches; Takada et al., 2021) to assess their suitability for eventual further dissemination.

### ***Cultures***

Strictly speaking, different cultures could also be subsumed under the previous section and the suggested methods for testing generalizability across groups apply here as well. However, cultural effects on measurement and prediction models are global in the truest sense of the word, and psychological research has a history of relative ignorance of this fact, so the topic merits separate mention. Recently, Stewart et al. (2023) examined the cross-cultural generalizability (across English, Russian and Mandarin speakers) of models using the Big Five personality domains, facets and nuances (measured with a newly developed 90-item measurement instrument) to predict various life outcomes (e.g., health, committed crimes, life satisfaction, education, income, social functioning). They found that items out-predicted facets and domains in all cultures, but to varying degrees (e.g., nuances achieved an average  $R^2$  of .16 vs.  $R^2 = .10$  for the English and Russian-speaking participants, respectively). Model performance decreased significantly when validated across cultures: For example, nuances-based models trained on Russian and Chinese samples were half as predictive ( $R^2 = .08$ ) for the English sample as the model trained and tested on the English sample. Thus, differences in

predictive performance due to cultural differences appear comparable to the intracultural incremental validity of nuances vs. domains (the latter of which achieved an  $R^2 = .07$  for the English sample). Similarly, Hofmann et al. (2023) showed that the performances of models trained on a U.S. sample using personality nuances (IPIP-NEO items) to predict gender varied largely between an AUC = .81 (e.g., when tested on a Canadian sample) and an AUC = .59 (when tested on a Filipino sample). As it is often the case with cross-cultural studies, it is not entirely clear whether these effects can be attributed solely to the different cultures or also to translation or comprehension issues (see Bader et al., 2021 for an approach to disentangling these different sources of bias), but the results nonetheless provide insights into cross-sample generalizability, or rather a lack thereof.

### ***Measurement***

The common use of different measures for ostensibly similar concepts in psychological research can be an additional hurdle to the generalizability of predictive models. An example from personality science is a study by Hang et al. (2021), who used nuances of different personality measures to predict participants' age, resulting in correlations between predicted and actual age of  $r = .65$  (IPIP-NEO; Johnson, 2014),  $.69$  (NEO-PI-R; Costa & McCrae, 2008),  $.50$  (HEXACO; Ashton & Lee, 2007), and  $.24$  (BFI-2; Soto & John, 2017). These results can be easily explained; for example, the IPIP-NEO is an openly available instrument that is largely based on the NEO-PI-R (hence, the similarity), and both contain four times as many individual items as the BFI-2, which probably limits the predictive validity of the latter's nuances. However, if one wants to make universal statements about the performance of the Big Five in predicting age, these results seem somewhat sobering and underline the importance of item sampling. In some cases, even the addition of a single word to a single-item measure can have a substantial effect on its approval rates and, presumably, on predictive models that use this indicator as an outcome: Nearly 40% of a U.S. adult online

sample including 613 participants endorsed lifetime suicide ideation when asked "Did you think about committing suicide?", whereas adding the word "seriously" (as in "Have you seriously thought about committing suicide") led to a lower endorsement of only 26% (Ammermann et al., 2021).

### ***Methods***

As we have done in the four manuscripts included in this thesis, it is common in studies using ML algorithms to compare different increasingly complex modeling approaches. In general, generalizability across algorithms cannot be easily assumed for models that vary in their flexibility. Even in scenarios where researchers a) compare tree-based algorithms such as random forests and GBM, and b) both algorithms achieve basically the same predictive accuracy, and c) they use a measure of variable importance that is based on the influence of a predictor averaged over all generated trees, the implications of the models may still differ. This phenomenon has been called Rashomon effect (a multiplicity of good models; Breiman, 2001), of which I present an empirical example below. For the example of random forests vs. GBM, an explanation can be found in their modeling behavior (see also the Prologue): Random forests aggregate across several independent smaller trees, each using only a subset of the predictor variables and the data. This way, in some trees, "unimportant" variables are randomly selected for a first split of the tree, leading to importance measures  $> 0$  for these variables. In contrast, GBM stack trees and model residuals of a previous weaker learner with the subsequent tree which reduces the probability of unimportant variables being selected in further iterations.

### ***Testing for Generalizability Across Cultures and Methods: An Empirical Example***

Strict tests of generalizability across any of the sources of model heterogeneity discussed above would, in principle, require that all other sources of performance heterogeneity be held constant. This condition is often difficult to meet in real datasets. For

example, for the prediction of psychotherapy dropout, we found differences in predictive performance and variable importances between the two clinics we studied, but since the data we used were not collected specifically for the comparisons we made, they differed in sample size, available predictor variables, timing of data collection, etc. As we discussed in the manuscript, baseline assessments would need to be aligned across multiple clinics to allow for direct tests of model generalizability. Another recent example for a test of generalizability across clinical samples can be found in Chekroud et al. (2024) who predicted patient remission in six trials of antipsychotic medication for schizophrenia. Predictive performance was better when models were tested within trials (balanced accuracies ranging from .56 to .67) than when using leave-one-trial-out-validation (balanced accuracies ranging from .50 to .58). Predictor variables were largely held constant across trials, however, the sample size and number of events varied substantially ( $N$  ranging from 99 to 481, number of events ranging from 24 to 153), which may account for any differences in model performance due to different degrees of model overfit. In the following, I present an empirical example in which I tested predictive models across cultures and methods while holding measurement, time, and setting constant.

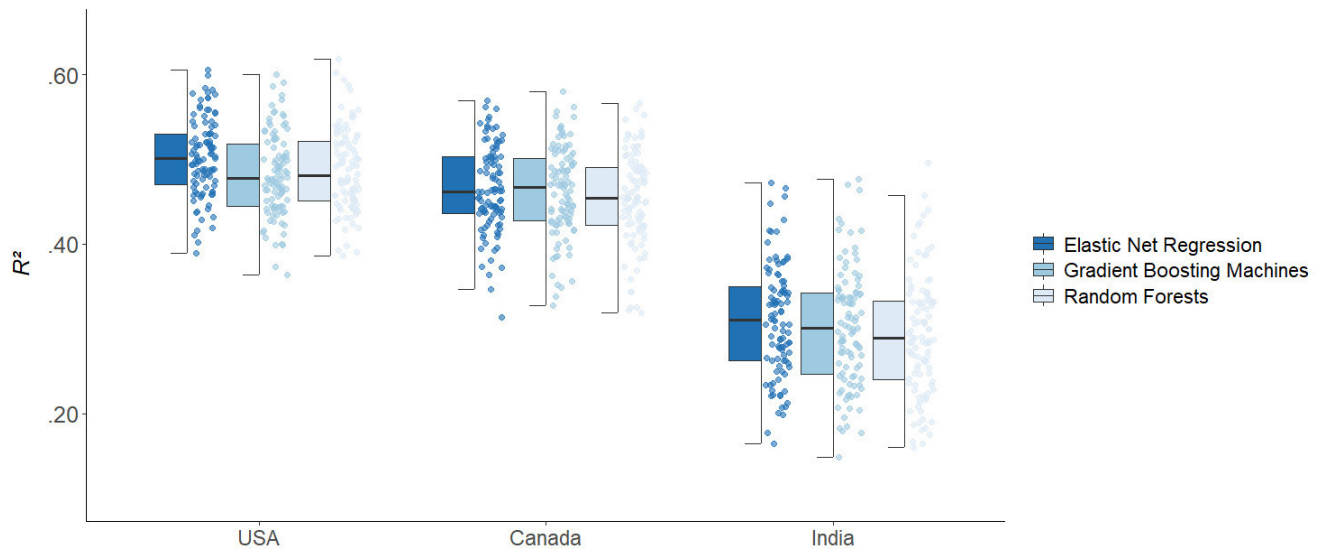
Using gender- and age-matched samples of 966 adults from five countries each, I examined the validity of character strengths measured with the VIA-Inventory of Strength-Positive (VIA-IS-P; McGrath, 2019) for predicting 13 indicators of well-being, mental and physical health (e.g., purpose in life, sleep quality, anxiety, or healthy eating) across modeling approaches and cultural (dis)similarity. I compared three ML algorithms, namely elastic net regressions, GBM, and random forests. Predictive models were trained and tested within and across countries using nested resampling and 100 iterations of the outer validation loop. Based on Hofstede's cultural dimensions (Hofstede et al., 2010), I assumed that there would be more differences between culturally dissimilar countries (U.S. vs. India and Mexico) than

between more similar countries (U.S. vs. Australia and Canada). Figure 5 shows exemplary results for one of the outcomes across three countries using a model trained on the U.S. sample: Purpose in life could be predicted to a considerable extent by character strengths, with up to  $R^2 = .50$  when tested within the U.S. sample. Irrespective of country, there was little difference in predictive performance between the three ML algorithms. The largest difference was found not in a comparison of methods, but in the evaluation of models across countries, especially when generalizing the model to a culturally dissimilar country (e.g.,  $\Delta R^2 = .19$  between elastic net regressions tested in the U.S. and India samples).

Comparing the overall model performance across methods and cultures addresses only one aspect of generalizability; another question would be whether the same variables are relevant for seemingly generalizable predictions. To examine whether there were any shifts in variable importances, I computed Spearman's rank correlations of averaged variable importance measures across methods and cultures. For the model trained with the U.S. sample, rank correlations were highest between GBM and random forests ( $r_s = .75$ ), lower between elastic net regressions and GBM ( $r_s = .64$ ), and lowest between elastic net regressions and random forests ( $r_s = .31$ ). In contrast to the negligible differences in predictive performance, the latter correlation in particular implies a high potential for different conclusions about variable importances, depending on which model a researcher examines more closely. With respect to the different predictive performances in culturally dissimilar countries, it would also be relevant to know whether this result is due to generally smaller effects (e.g., in India than in the U.S.) or predominantly driven by shifts in variable importance ranks. For example, rank correlation between elastic net regression models trained in the U.S. sample and Canada was lowest ( $r_s = .10$ ), followed by U.S. and India ( $r_s = .16$ ) and Canada and India ( $r_s = .24$ ). Again, the effect of culture on variable ranks proved to be more pronounced than that of different methods.

**Figure 5**

*Model Test Performances for the Prediction of Purpose in Life Based on 96 Character Strength Nuances*



*Note.* The box reflects the interquartile range (IQR), the solid line the median, and the whiskers 1.5-times the IQR across 100 iterations (i.e., outer loop data splits).  $R^2$  values of the 100 models are displayed as jittered distribution on the right.

In summary, these results show that prediction models cannot easily be assumed to be generalizable across cultures, but also that nearly identical predictive performances can lead to different model interpretations (as in Rashomon effects) depending on the specific algorithm. Thus, the take-home message of this part of my epilogue is that model generalizability is far from a given and even reproducibility across similar algorithms becomes more complicated with increasing model complexity, which is why it is essential to communicate predictive models within their boundary conditions (see, e.g., Simons et al., 2017 for a proposal to add a paragraph discussing generalizability constraints to empirical studies).

### **Asking Better and Fewer Questions...**

In predictive modeling, researchers have to strike a balance between including all available information in the model (considering that unreliable or irrelevant variables may

reduce predictive performance) and preselecting variables based on theoretical assumptions or in a data-driven manner (at the risk of either excluding relevant variables or increasing overfitting). In psychological research, with its relatively small sample sizes (Fraley et al., 2022), the inherent regularization of ML algorithms easily reaches its limits and cannot always compensate a high number of noisy and / or unreliable predictor variables (e.g., Jacobucci & Grimm, 2020). This raises the question of whether more sparse and accessible models might be generally preferable to complex ML algorithms, especially in application contexts where the most accurate and user-friendly decision rules would be optimal. In this respect, it may be good news that there is still a lot of room for improvement in psychological research in terms of more drastic feature selection; in recent studies, often only a handful out of a large number of variables turned out to be essential for predicting the respective outcome. For example, it seems hardly surprising that out of 240 NEO-PI-R items, “overeats favorite foods” had the highest association with participants’ body mass index (self- and other-reported and across samples; e.g., Möttus et al., 2017) and largely drove the higher predictive performance of models including personality nuances compared to facets and factors for this outcome.

Another example can be found in Reiter and Schoedel (2023) who compared logistic regressions, elastic net regressions, and random forests for predicting compliance in an experience sampling setting using more than 400 person, behavior, and context variables. They found that elastic net regressions slightly outperformed the other two methods with an averaged AUC = .72 and past participation behavior (operationalized for example by the previous mean answer rate, previous mean answer latency, and compliance at last beep) was by far the most important predictor category (predictive performance dropped to AUC = .59 when this category was excluded from the model). In comparison, excluding participants’ baseline characteristics such as demographics or personality traits had virtually no effect on

predictive performance (drop in AUC = .001). Although they apply to a different study design, these findings are consistent with our results of predicting attrition in that it may not be possible to accurately infer some sort of “attrition personality” using only baseline participant characteristics, and they suggest that the best way to predict future participation may be to examine only the extent of previous regular participation in longitudinal studies.

For the prediction of suicidal behavior, transparent and simple decision rules have been proposed, such as classifying any person who has ever engaged in self-harm as being at risk for future self-harm, including suicide attempts (e.g., Van Vuuren et al., 2021). Given that past self-harm has been meta-analytically shown to be the most important predictor of future self-harm and suicidal behavior (Beckman et al., 2018; Geulayov et al., 2019), such single-indicator decision rules appear to be a valid baseline against which more complex models should be evaluated. In manuscript 2, we found that the use of such decision rules resulted in similar averaged balanced accuracy for predicting adolescent lifetime suicide attempts as the models that included all available information (.74 vs. .76), but averaged sensitivity was higher for the more complex model (.69 vs. .59). On the one hand, for the context of suicide attempts, an argument can be made for the importance of sensitivity, since accurate detection of at-risk cases is obviously highly relevant. On the other hand, a more complex and potentially less robust model based on household panel data might be difficult to translate into clinical practice, and thus this higher sensitivity would be merely theoretical.

In a similar vein, therapists’ ratings of patients’ initial motivation and helping alliance emerged as by far the most relevant variables for predicting therapy dropout in the third project included in this thesis. Models using only these variables in the respective samples performed only slightly worse than those including all available (> 100 for each sample) predictor variables (a decrease in averaged balanced accuracy of .01 for Sample 1 and .02 for Sample 2). From a cost-benefit perspective, the question arises as to whether it is helpful to



include all possible information in elaborate models for a preliminary screening, or whether it would be sufficient to ask the respective therapists about their impressions of the patient after the first meeting and then, to guide the course of therapy accordingly, if necessary.

Alternatively, one could also build on these findings and assess therapists' first impressions much more comprehensively to examine which specific aspects are predictive of therapy dropout since, for example, a single-item measure of patient motivation is likely to be answered by aggregating a wide variety of potentially relevant cues.

In general (and this is also true for the four manuscripts of this dissertation), many ML models in psychological research to date are based on data that were not primarily collected for specific prediction purposes, but were generated in other contexts or research projects. Often, all available variables are simply included in prediction models on a "more is better" basis, implicitly assuming that expert knowledge is less relevant and that ML models will be able to separate the signal from irrelevant noise. Therefore, there is certainly a lot of untapped potential in feature selection and improved measurement of tailored predictor variables to enhance performances of prediction models across different psychological sub-disciplines.

### **... at the Right Time?**

If one is interested not only in *whether* someone is at risk, but also in *when* exactly (and hence, when the respective behavior is most likely to be exhibited), the prediction model must also account for the temporal instability of the outcome and its associations. A perfect example to illustrate this problem of mismatched timescales is the prediction of suicidal behavior. For the prediction of lifetime suicide attempts in adolescents, we used panel data taken from a measurement wave three years before the self-reported outcome was assessed. With such a model, it is only possible to obtain a very global risk score on the basis of which adolescents could be further contacted, but this score does not have any information about

when adolescents were at increased risk. The latter, however, is of interest if the goal is to prevent actual suicides.

Widely adapted and empirically supported theoretical models of suicide underline the notion that the causes of suicide are multifaceted and that the process leading to the actual behavior involves different factors and stages. For example, the integrated motivational–volitional model of suicidal behavior (O’Connor & Kirtley, 2018; Wetherall et al., 2018) describes suicide as a behavior that occurs when subsequent facilitating phases are experienced, thus adding a temporal perspective to emerging suicidal plans and behaviors. In a pre-motivational phase, background factors and triggering events potentially foster the onset of suicide ideation within the subsequent motivational phase. Generally, predictive models using panel data and measurement waves years apart would merely be able to capture this first proposed phase. In a second phase, feelings of defeat and humiliation are thought to lead to feelings of entrapment which in turn lead to suicide ideation and intent. In the volitional phase—given moderating factors such as impulsivity and access to means—suicide intention turns into suicidal behavior. To ultimately help prevent suicide attempts, a predictive model would have to accurately model concrete behaviors within such a third, volitional phase.

Regardless of the specific theoretical model of suicidal behaviors, empirical evidence supports the general premise that suicide ideation and intent are highly fluctuating states. For example, Coppersmith et al. (2023) examined the timescale of suicidal thinking in 105 adults using a 42-day real-time monitoring study. They found that a) elevated states of suicidal thinking lasted less than 3 hours on average, b) current suicidal intent (“How strong is your intent to kill yourself right now?”) predicted future suicidal intent only up to 3 hours whereas current suicidal desire (“How strong is your desire to kill yourself right now?”) was predictive of future suicidal desire for up to 20 hours and c) the estimation of the states' stability depended on the frequency of assessment (every 10 minutes over the course of an hour vs.

hours apart over the course of a day) as in stability was estimated to be lower when using data from the more frequent assessment. Taken together, these numbers show that it is quite possible to miss narrow windows of elevated risk, even in daily diary studies or with similar longitudinal designs using assessments more than 3 hours apart. However, the answer to this dilemma is not to monitor everyone continuously for increased suicidal intent. In addition to being highly intrusive and impractical, we cannot yet rule out the possibility that repeatedly asking about topics on suicide may have negative effects. Although Coppersmith et al. (2022) were able to show that repeated assessment of suicidal thinking did not increase the severity of suicidal thinking, their findings were based on an online sample of 101 adults who already had current suicidal ideation and it is unclear whether they are generalizable to community, adolescent, or even higher-risk clinical samples. Therefore, for ethical reasons alone, closer monitoring should be an opt-in approach.

From a purely methodological point of view, predictive models would benefit from matching the frequency of assessments to the timescales of relevant predictors, and not only in the context of suicidal behavior. In psychotherapy, for example, models that use weekly monitoring can provide early warning signs of treatment nonresponse or dropout (e.g., Lutz et al., 2018) and can be integrated into the therapeutic process as feedback for therapists (e.g., Barkham et al., 2023), thus providing valuable information beyond baseline assessments.

### **Scientific Utopia for Machine Learning in Psychological Research**

Nosek and Bar-Anan (2012), Nosek et al. (2012), and Uhlmann et al. (2019) published a series of articles (titled Scientific Utopia I, II, and III) with ideas and strategies for improving scientific practices and rendering psychological science more accurate, open, and collaborative. To end this epilogue on a positive note, in the following, I will adopt this mindset for an outlook regarding ML modeling in psychological research. There are essentially three pillars where there is the most obvious room for improvement: a) the

methodological aspects of model training and model evaluation, b) conceptual aspects including model interpretation, and c) transparency and open science practices.

The magnitude of errors in ML model validation across different scientific disciplines (e.g., neuropsychiatry, nutrition research, or genetics) in at least 294 published research papers has already prompted the term “reproducibility crisis” (Kapoor & Narayanan, 2023). There has not yet been a similar review of psychological research, but it is safe to assume that it is not exempt from pitfalls and issues in modeling and model validation (see, for example, the findings I described in the prologue regarding the prediction of suicidal behavior; Jacobucci et al., 2021). The problem responsible for most of these biased, overly optimistic, and misleading predictive performances, especially when complex algorithms are used, is information leakage between training and testing data. Leakage can take on many forms but is usually due to distribution- or target-dependent preprocessing (e.g., handling missingness or unequal group sizes) prior to splitting the data into training and testing data (e.g., Vanderwiele et al., 2021) or failure to account for hierarchical temporal dependencies in the data during model validation. In my view, these overly optimistic predictive performances should not necessarily be seen as some form of results-hacking, but rather reflect the fact that ML algorithms are relatively new tools in many scientific disciplines and that knowledge about possible pitfalls to watch out for has not been sufficiently disseminated. Thus, it seems worthwhile to create checklists and recommendations (e.g., Pargent et al., 2023; Rosenbusch et al., 2021) tailored to psychological research, similar to tools available in other contexts such as the Checklist for Artificial Intelligence in Medical Imaging (CLAIM; Mongan et al., 2020). These should be open, educational, highly accessible, and, at best, based on expert consensus in order to be widely accepted and used. Scientific journals could use these tools to screen submissions for common training and validation pitfalls or—even more desirably—in combination with adopting initiatives such as the newly added statistics, transparency, and

rigor (STAR) editors in Psychological Science, who perform transparency and reproducibility checks on submitted manuscripts (Vazire, 2023).

In addition to sound model training and validation, much more thought should be given to whether ML algorithms are even indicated or useful for specific research projects. Predictive modeling coupled with ML algorithms may seem like a trendy topic at the moment, however, researchers should ask themselves the following questions before abandoning (often somewhat dismissively referred to as) "traditional methods": Is my research question a prediction task at all, or am I interested in the (unregularized) size of a single effect? Is it plausible to assume that there are any nonlinear or interaction effects in my data? If so, do I have enough data to reliably detect such effects? Will this prediction model ever be used for real-world applications, or could I better use my data to further validate existing models? How will I interpret my results, and by what measures? What (causal) claims do I want to make and do my modeling choices fit to these goals? Especially the last two questions concern potentially problematic inferences when models control for a large number of conveniently sampled variables, some of which may be collider variables (e.g., Rohrer 2018). Even when researchers do not explicitly plan for causal inferences, semantics (such as using the words "risk factors" for any variable associated with an outcome) often imply causality and lead to methodological confusion<sup>2</sup>. In terms of model interpretability in general, ML based models are often viewed as black boxes, which is a convenient excuse not to further think about the algorithms' functionality. For psychological research, Henninger et al. (2023) provide a comprehensive review of interpretability for ML algorithms that should be consulted by psychologists who are unsure which measure might fit their models and research goals. For more complex algorithms such as neural networks, there are additional strategies available for

---

<sup>2</sup> Huitfeldt (2016) describes this problem in an entertaining way with a short article titled "Is caviar a risk factor for being a millionaire?".

accessing what is happening at specific layers (e.g., Montavon et al., 2018). From a model fairness or alignment perspective, it is not enough to merely maximize predictive performance, an understanding of why the prediction works as well as it does is also required. Otherwise, researchers are effectively handing over responsibility for the potential consequences of their findings to the algorithms and simple errors in data preparation or coding could lead to absurd policy recommendations (e.g., Caruana et al., 2015).

Predictive modeling using ML algorithms inherently entail many researchers' degrees of freedom (more so, the more complex a model becomes). Data preparation, preprocessing, model training, validation, and interpretation are each steps that involve multiple decisions that can affect the predictive performance and the conclusions drawn from the results. Providing a reproducible analytical syntax is the easiest way out of this dilemma, since it seems unrealistic that sufficient information (e.g., each mutation of predictor variables, the exact hyperparameter grid of each algorithm used or how preprocessing steps are included in which specific cross-validation scheme) can be provided in often word-limited method sections. Currently, it seems that open science practices and data-driven predictive modeling are rarely combined, in part because of the misconception that it is not worthwhile or feasible to specify data-driven exploratory analyses in advance. This idea may have been reinforced by the lack of preregistration templates explicitly tailored to predictive models; the most commonly used templates (e.g., Bowman et al., 2020) require hypotheses, do not mention internal or external (cross-)validation and list exploratory analyses as optional. As part of a DFG-funded personality computing network, my colleagues and I are currently working on an ML-tailored preregistration template and an accompanying tutorial to fill this gap. Preregistration of analyses is also beneficial for data-driven analyses, as it forces researchers to explicitly state answers to questions such as the ones I listed above. The combination of registered reports (e.g., Chambers & Tzavella, 2022) and ML-based predictive modeling

would even go a step further in that it would allow researchers to receive feedback at a conceptual and methodological level. As we argued in manuscript 4, many reproducibility issues or misleading presentation of findings could be addressed with a thorough review by ML experts prior to analysis and dissemination of results. Registered reports also remove the incentive to produce overly interesting or novel findings, which should in turn should reduce selective reporting. For example, analogous to extreme  $p$ -hacking, without any prior specification of the model validation scheme to be used, a researcher could start off with a nested resampling scheme (see the prologue for a detailed explanation), select the best performing (or a set of high performing) outer resampling iterations and present the resulting subset of analyses as originally planned.

In conclusion, there are many factors that can have a positive impact on the quality of future studies using ML algorithms in psychological research. Nosek (2019) proposed a model explaining strategies for cultural change (in terms of scientific practices), consisting of five progressive stages a) infrastructure (make it possible), b) user interface (make it easy), c) communities (make it normative), d) incentives (make it rewarding), and e) policy (make it required). Regarding fully transparent and reproducible ML-based science in psychological research, the field is probably somewhere between the first two stages, so it does not yet seem fruitful to impose many requirements on researchers as a first (probably overwhelming) step. However, ML researchers could rather start by providing more educational materials, checklists, recommendations, or templates and try to involve the scientific community in their development to facilitate normative change.

## References

- Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*(2), 150–166. <https://doi.org/10.1177/1088868306294907>
- Ammerman, B. A., Burke, T. A., Jacobucci, R., & McClure, K. (2021). How we ask matters: The impact of question wording in single-item measurement of suicidal thoughts and behaviors. *Preventive Medicine, 152*, Article 106472. <https://doi.org/10.1016/j.ypmed.2021.106472>
- Bader, M., Jobst, L. J., Zettler, I., Hilbig, B. E., & Moshagen, M. (2021). Disentangling the effects of culture and language on measurement noninvariance in cross-cultural research: The culture, comprehension, and translation bias (CCT) procedure. *Psychological Assessment, 33*(5), 375–384. <https://doi.org/10.1037/pas0000989>
- Barkham, M., De Jong, K., Delgadillo, J., & Lutz, W. (2023). Routine outcome monitoring (ROM) and feedback: Research review and recommendations. *Psychotherapy Research, 33*(7), 841–855. <https://doi.org/10.1080/10503307.2023.2181114>
- Beckman, K., Mittendorfer-Rutz, E., Waern, M., Larsson, H., Runeson, B., & Dahlin, M. (2018). Method of self-harm in adolescents and young adults and risk of subsequent suicide. *Journal of Child Psychology and Psychiatry, 59*(9), 948-956. <https://doi.org/10.1111/jcpp.12883>
- Bischi, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation, 20*(2), 249–275. [https://doi.org/10.1162/EVCO\\_a\\_00069](https://doi.org/10.1162/EVCO_a_00069)
- Bowman, S. D., DeHaven, A. C., Errington, T. M., Hardwicke, T. E., Mellor, D. T., Nosek, B. A., & Soderberg, C. K. (2020, January 22). OSF Prereg Template. <https://doi.org/10.31222/osf.io/epgjd>



- Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3). <https://doi.org/10.1214/ss/1009213726>
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for HealthCare: predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, 1721–1730. <https://doi.org/10.1145/2783258.2788613>
- Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, 6, 29–42. <https://doi.org/10.1038/s41562-021-01193-7>
- 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*, 383(6679), 164–167. <https://doi.org/10.1126/science.adg8538>
- Coppersmith, D. D. L., Fortgang, R. G., Kleiman, E. M., Millner, A. J., Yeager, A. L., Mair, P., & Nock, M. K. (2022). Effect of frequent assessment of suicidal thinking on its incidence and severity: high-resolution real-time monitoring study. *The British Journal of Psychiatry*, 220(1), 41–43. <https://doi.org/10.1192/bjp.2021.97>
- Coppersmith, D. D. L., Ryan, O., Fortgang, R. G., Millner, A. J., Kleiman, E. M., & Nock, M. K. (2023). Mapping the timescale of suicidal thinking. *Proceedings of the National Academy of Sciences of the United States of America*, 120(17), e2215434120. <https://doi.org/10.1073/pnas.2215434120>
- Costa, P. T., Jr., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment, Vol. 2. Personality measurement and testing* (pp. 179–198). Sage Publications, Inc. <https://doi.org/10.4135/9781849200479.n9>

- Dragicevic, M., & Casalicchio, G. (2020). Resampling—stratified, blocked and predefined. Mlr-Org. Retrieved January 10, 2024, from <https://mlr-org.com/gallery/basic/2020-03-30-stratification-blocking/>
- Fox, K. R., Huang, X., Linthicum, K. P., Wang, S. B., Franklin, J. C., & Ribeiro, J. D. (2019). Model complexity improves the prediction of nonsuicidal self-injury. *Journal of Consulting and Clinical Psychology, 87*(8), 684–692. <https://doi.org/10.1037/ccp0000421>
- Fraley, R. C., Chong, J. Y., Baacke, K. A., Greco, A. J., Guan, H., & Vazire, S. (2022). Journal N-Pact Factors from 2011 to 2019: Evaluating the quality of social/personality journals with respect to sample size and statistical power. *Advances in Methods and Practices in Psychological Science, 5*(4). <https://doi.org/10.1177/25152459221120217>
- Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin, 143*(2), 187-232. <https://doi.org/10.1037/bul0000084>
- Geulayov, G., Casey, D., Bale, L., Brand, F., Clements, C., Farooq, B., Kapur, N., Ness, J., Waters, K., Tsiachristas, A., & Hawton, K. (2019). Suicide following presentation to hospital for non-fatal self-harm in the multicentre study of self-harm: a long-term follow-up study. *The Lancet Psychiatry, 6*(12), 1021-1030. [https://doi.org/10.1016/S2215-0366\(19\)30402-X](https://doi.org/10.1016/S2215-0366(19)30402-X)
- Henninger, M., Debelak, R., Rothacher, Y., & Strobl, C. (2023). Interpretable machine learning for psychological research: Opportunities and pitfalls. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000560>

Hofmann, R., Rozgonjuk, D., Soto, C. J., Ostendorf, F., & Möttus, R. (2023, November 16).

There are a million ways to be a woman and a million ways to be a man: Gender differences across personality nuances and nations.

<https://doi.org/10.31234/osf.io/cedwk>

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind*. New York: McGraw-Hill

Huitfeldt A. (2016). Is caviar a risk factor for being a millionaire?. *BMJ (Clinical research ed.)*, 355, i6536. <https://doi.org/10.1136/bmj.i6536>

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, 15(3), 809–816. <https://doi.org/10.1177/1745691620902467>

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51, 78–89. <https://doi.org/10.1016/j.jrp.2014.05.003>

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804.

<https://doi.org/10.1016/j.patter.2023.100804S>

King, D., Gronholm, P. C., Knapp, M., Hoffmann, M. S., Bonin, E. M., Brimblecombe, N., Kadel, R., Maughan, B., O'Shea, N., Richards, M., Hoomans, T., & Evans-Lacko, S. (2023). Effects of mental health status during adolescence on primary care costs in adulthood across three British cohorts. *Social Psychiatry and Psychiatric Epidemiology*. Advance online publication. <https://doi.org/10.1007/s00127-023-02507-y>

- Lutz, W., Schwartz, B., Hofmann, S. G., et al. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study. *Scientific Reports*, 8, 7819. <https://doi.org/10.1038/s41598-018-25953-0>
- McGrath, R. E. (2019). Technical report: The VIA assessment suite for adults: development and initial evaluation revised edition. Cincinnati, OH: VIA Institute on Character [www.viacharacter.org](http://www.viacharacter.org)
- Mongan, J., Moy, L., & Kahn, C. E., Jr (2020). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, 2(2), e200029. <https://doi.org/10.1148/ryai.2020200029>
- Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
- Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, 112(3), 474–490. <https://doi.org/10.1037/pspp0000100>
- Nosek. (2019, June 11). Strategy for culture change. <https://www.cos.io/blog/strategy-for-culture-change>
- Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, 23(3), 217–243. <https://doi.org/10.1080/1047840X.2012.692215>
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6), 615–631. <https://doi.org/10.1177/1745691612459058>

- O'Connor, R. C., & Kirtley, O. J. (2018). The integrated motivational–volitional model of suicidal behaviour. *Philosophical Transactions of the Royal Society B*, 373, 20170268. <http://doi.org/10.1098/rstb.2017.0268>
- Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, 6(3). <https://doi.org/10.1177/25152459231162559>
- Parsons, S., Sullivan, A., Fitzsimons, E., & Ploubidis, G. (2021). The role of parental and child physical and mental health on behavioral and emotional adjustment in mid-childhood: a comparison of two generations of British children born 30 years apart. *Longitudinal and Life Course Studies*, 12(4), 517-550. <https://doi.org/10.1332/175795921X16115949616122>
- Reiter, T., Schoedel, R. (2023). Never miss a beep: Using mobile sensing to investigate (non)compliance in experience sampling studies. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-023-02252-9>
- Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27-42. <https://doi.org/10.1177/2515245917745629>
- Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated R code. *Social and Personality Psychology Compass*. <https://doi.org/10.1111/spc3.12579>

- Siddaway, A. P., Quinlivan, L., Kapur, N., O'Connor, R. C., & de Beurs, D. (2020). Cautions, concerns, and future directions for using machine learning in relation to mental health problems and clinical and forensic risks: A brief comment on “model complexity improves the prediction of nonsuicidal self-injury” (Fox et al., 2019). *Journal of Consulting and Clinical Psychology, 88*(4), 384–387. <https://doi.org/10.1037/ccp0000485>
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Song, Q. C., Tang, C., & Wee, S. (2021). Making sense of model generalizability: A tutorial on cross-validation in R and Shiny. *Advances in Methods and Practices in Psychological Science, 4*(1). <https://doi.org/10.1177/2515245920947067>
- Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. <https://doi.org/10.1037/pspp0000096>
- Stewart, R. D., Díaz, A., Hou, X., LIU, X., Vainik, U., Johnson, W., & Mõttus, R. (2023, July 7). The ways of the world? Cross-sample replicability of personality trait-life outcome associations. <https://doi.org/10.31234/osf.io/6c592>
- Takada, T., Nijman, S., Denaxas, S., Snell, K. I. E., Uijl, A., Nguyen, T. L., Asselbergs, F. W., & Debray, T. P. A. (2021). Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. *Journal of Clinical Epidemiology, 137*, 83–91. <https://doi.org/10.1016/j.jclinepi.2021.03.025>

- Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., McCarthy, R. J., Riegelman, A., Silberzahn, R., & Nosek, B. A. (2019). Scientific Utopia III: Crowdsourcing Science. *Perspectives on Psychological Science*, *14*(5), 711-733. <https://doi.org/10.1177/1745691619850561>
- Van Calster, B., Steyerberg, E. W., Wynants, L., & van Smeden, M. (2023). There is no such thing as a validated prediction model. *BMC medicine*, *21*(1), 70. <https://doi.org/10.1186/s12916-023-02779-w>
- Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J., Van Hoecke, S., & Demeester, T. (2021). Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine*, *111*, 101987. <https://doi.org/10.1016/j.artmed.2020.101987>
- Van Vuuren C. L., van Mens K., de Beurs D., Lokkerbol J., van der Wal M. F., Cuijpers P., Chinapaw M. J. M. (2021). Comparing machine learning to a rule-based approach for predicting suicidal behavior among adolescents: Results from a longitudinal population-based survey. *Journal of Affective Disorders*, *295*, 1415–1420. <https://doi.org/10.1016/j.jad.2021.09.018>
- Vazire, S. (2023). The next chapter for Psychological Science. *Psychological Science*, *0*(0). <https://doi.org/10.1177/09567976231221558>
- Wetherall, K., Robb, K. A., & O'Connor, R. C. (2019). An examination of social comparison and suicide ideation through the lens of the integrated motivational–volitional model of suicidal behavior. *Suicide and Life-Threatening Behavior*, *49*(1), 167–182. <https://doi.org/10.1111/sltb.12434>

**Anlage 1. Erklärung gemäß § 8 der Allgemeinen Bestimmungen für Promotionen der  
Universität Kassel vom 14.07.2021.**

1. Bei der eingereichten Dissertation zu dem Thema „Promises and Pitfalls of Machine Learning Modeling in Psychological Research“ handelt es sich um meine eigenständig erbrachte Leistung.
2. Anderer als der von mir angegebenen Quellen und Hilfsmittel habe ich mich nicht bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen veröffentlichten oder unveröffentlichten Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Dissertation oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die abgegebenen digitalen Versionen stimmen mit den abgegebenen schriftlichen Versionen überein.
5. Ich habe mich keiner unzulässigen Hilfe Dritter bedient und insbesondere die Hilfe einer kommerziellen Promotionsberatung nicht in Anspruch genommen.
6. Im Fall einer kumulativen Dissertation: Die Mitwirkung von Koautoren habe ich durch eine von diesen unterschriebene Erklärung dokumentiert. Eine Übersicht, in der die einzelnen Beiträge nach Ko-Autoren und deren Anteil aufgeführt sind, füge ich anbei.
7. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

---

Datum

---

Unterschrift



**Anlage 2. Erklärung über den Eigenanteil an den veröffentlichten oder zur  
Veröffentlichung vorgesehenen wissenschaftlichen Schriften innerhalb meiner  
Dissertationsschrift**

Universität Kassel, Fachbereich Humanwissenschaften Erklärung zur kumulativen  
Dissertationen im Promotionsfach Psychologie  
Ergänzung zu § 5a Abs. 4 Satz 1 der Allgemeinen Bestimmungen für Promotionen an der  
Universität Kassel vom 13. Juni 2011

**Antragsstellerin:**

Kristin Jankowsky, Institut für Psychologie, Universität Kassel  
Titel der Arbeit: Promises and Pitfalls of Machine Learning Modeling in Psychological  
Research

**Nummerierte Aufstellung der eingereichten Schriften**

1. Jankowsky, K. & Schroeders, U. (2022). Validation and generalizability of machine learning prediction models on attrition in longitudinal studies. *International Journal of Behavioral Development*, 46(2), 169–176.  
<https://doi.org/10.1177/01650254221075034>
2. Jankowsky, K., Steger, D., & Schroeders, U. (2023). Predicting lifetime suicide attempts in a community sample of adolescents using machine learning algorithms. *Assessment*, Advance online publication. <https://doi.org/10.1177/10731911231167490>
3. Jankowsky, K., Krakau, L., Schroeders, U., Zwerenz, R., & Beutel, M. E. (2023). Predicting treatment response using Machine Learning: A registered report. *British Journal of Clinical Psychology*, Advance online publication.  
<https://doi.org/10.1111/bjc.12452>
4. Jankowsky, K., Zimmermann, J., Jaeger, U., Mestel, R., & Schroeders, U. (2023, September 27). First impressions count: Therapists' impression on patients' motivation and helping alliance predicts psychotherapy dropout. Retrieved from [psyarxiv.com/nhs6c](https://psyarxiv.com/nhs6c)

### **Anlage 3. Dokumentation der genutzten Daten**

Bei den hier präsentierten Studien handelt es sich um Reanalysen, für die keine zusätzlichen Daten erhoben wurden. Ich verwalte keine der innerhalb der Dissertation genutzten Datensätze. In Manuskript 1 wurden die frei zugänglichen Daten des Midlife in the United States Panels (MIDUS) and des Panel Analysis of Intimate Relationships and Family Dynamics (pairfam) genutzt. MIDUS- Daten sind unter diesem Link kostenfrei verfügbar: <https://www.midus.wisc.edu/data/index.php>. Pairfam-Daten können hier <https://www.pairfam.de/de/daten/datenzugang/> von Forscher:innen beantragt werden. In Manuskript 2 werden die Daten der Millennium Cohort Study genutzt, die ebenfalls für Forscher:innen frei verfügbar sind (mehr Informationen: <https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/>). Die Daten für Manuskripte 3 und 4 werden jeweils von den klinisch arbeitenden Coautor:innen verwaltet und können bei diesen angefragt werden. Für den Prologue habe zusätzlich Daten des GESIS panels (<https://www.gesis.org/gesis-panel/gesis-panel-home>) analysiert, die ebenfalls nach Abschluss eines Nutzungsvertrags für Forscher:innen frei verfügbar sind. Analysesyntaxen und weitere Materialien sind jeweils in OSF-Projekten vorhanden und in den jeweiligen Manuskripten klar verlinkt. Syntaxen für die Analysen und Grafiken des Mantels sind unter <https://osf.io/yhbq6/> zu finden.