






Using Machine Learning to Predict Psychosomatic Rehabilitation Success

Marcel E. K. F. Holzer¹ , Natalie M. Hogh¹ , Paul-Gerrit Velthuysen^{2,3} ,
Mirjam Körner^{2,4} , and Anja S. Göritz¹ 

¹Behavioral Health Technology, University of Augsburg, Bavaria, Germany

²Institute of Medical Psychology and Medical Sociology, University of Freiburg, Germany

³Institute of Psychology, University of Hildesheim, Germany

⁴Competence Center Interprofessionality, Bern University of Applied Sciences, Bern, Switzerland

Abstract: This study used data from 7,436 patients across seven psychosomatic rehabilitation clinics to explore predictors of treatment success and to assess the feasibility of predicting rehabilitation outcomes using machine learning. Five outcomes based on the biopsychosocial model – activity, depressive symptoms, participation, phobic fears, and somatoform complaints – were derived from the HEALTH-49 and ICF AT-50 Psych. The outcomes were dichotomized to indicate relevant change from admission (T1) to discharge (T2). Random forests using baseline scores, PAREMO-20, and SIMBO-C items, treatment year, clinic, sex, and age, outperformed all dummy classifiers, yielding accuracies between 63.0% (participation) and 75.9% (phobic fears). Feature importances revealed baseline scores as key predictors. Skepticism toward rehabilitation predicted all outcomes except phobic fears, while willingness to change predicted all outcomes except depressive symptoms. Difficulties in social interaction impacted depressive symptoms and phobic fears. Work-related challenges were key predictors for activity and participation. Age, sex, year, and clinic had no impact.

Keywords: psychosomatic rehabilitation, machine learning, biopsychosocial model, treatment success, random forest

Psychosomatic Rehabilitation

Psychosomatic rehabilitation (PSR) accounts for the second largest portion of rehabilitation services (after orthopedic rehabilitation) and has positive economic effects (Deutsche Rentenversicherung, 2024; Steffanowski et al., 2005). PSR is an interdisciplinary treatment approach that addresses both physical and psychological complaints with a multimodal approach that includes psychotherapy, occupational therapy, sport and nutritional therapy, and art therapy (Deutsche Rentenversicherung, 2018). Germany has about 200 PSR facilities that treat 190,000 patients annually with an average treatment duration of 35 days (Statistisches Bundesamt, 2023). PSR is the second largest area of inpatient care for mental illness after psychiatry (Steffanowski et al., 2005). Yearly treatment costs for PSR were estimated at one billion Euro in 2007 (Steffanowski et al., 2005) and since then have increased sharply (Deutsche Rentenversicherung, 2024). At the same time, it is estimated that every Euro invested in PSR pays for itself economically up to four times over (Nübling et al., 2020).

Although PSR is generally effective, its effectiveness might be improved (Reuter et al., 2014). Many patients experience

improved functioning and a successful return to work after PSR (Steffanowski et al., 2005). However, 20%–30% of patients are considered nonresponders, and about another 10% report a deterioration in health following treatment (Reuter et al., 2014). This not only results in financial loss but also leads to patient suffering. To enhance the effectiveness of PSR and attempt that more patients benefit from treatment, previous research has explored various factors and patient characteristics that may influence treatment success. Nonetheless, further research is needed to expand on these initial findings, clarify inconsistent results, and investigate the benefits of advanced methods such as machine learning (ML).

Related Work

Treatment Goals

In PSR research, treatment goals and outcomes vary widely, with patient-reported outcome measures (PROMs) frequently used and objective measures used less often

(Benson, 2022). Some of the most prevalent indicators of treatment success include symptom reduction (Henn et al., 2021; von Hörsten et al., 2019), general psychosocial health and well-being (Henn et al., 2021; Kleineberg-Massuthe et al., 2023; Lange et al., 2012), as well as specific mental health conditions such as depression (Henn et al., 2021; Kleineberg-Massuthe et al., 2023). Beyond psychological factors, work-related outcomes are frequently used, for example, subjective measures such as self-reported work ability (Oster et al., 2009), or objective measures such as the return-to-work rate (von Hörsten et al., 2019) and earning capacity (Papst & Köllner, 2022). Researchers employ a multitude of questionnaires and/or rating methods to operationalize these outcomes, such as the Beck Depression Inventory-2 (e.g., Kleineberg-Massuthe et al., 2023), the Global Severity Index (e.g., Lange et al., 2012), and the HEALTH-49 (e.g., Henn et al., 2021; Kaier et al., 2014). Thus, PSR failure and PSR success do not denote a uniform construct in the literature but depend on how each study operationalizes the outcomes. This heterogeneity complicates the comparison of PSR treatment results across studies.

Analysis Approaches

In PSR, the application of more advanced methods such as ML is scarce. The majority of PSR studies employed traditional statistical methods, including *t*-tests, cross tables, chi-square tests, ANOVAs, and regression analyses (Henn et al., 2021; Kessemeier et al., 2017; Sandweg et al., 2001). The only PSR study we could find that used ML is Papst and Köllner (2022): They used a gradient boosting model to predict earning capacities, identifying psychological well-being as the strongest predictor. In other areas of rehabilitation, ML approaches are more common and have demonstrated promising results, particularly in orthopedics (Tschuggnall et al., 2021) and stroke rehabilitation (Campagnini et al., 2022).

Predictors of Treatment Success

Research has identified several patient characteristics that predict treatment success, most of which can be categorized into the following groups: sociodemographics, patient health history, work-related factors, and attitudes toward PSR.

Findings on sociodemographics are heterogeneous: Age and sex do not influence PSR outcomes (Reuter et al., 2014; Steffanowski et al., 2005; von Hörsten et al., 2019). While one study (Oster et al., 2009) found education to be irrelevant to treatment success, other research suggests lower education correlates with poorer PSR outcomes

(Lange et al., 2012; Reuter et al., 2014). Some studies reported worse health outcomes for Turkish migrants compared to a German control group (Möske et al., 2008). However, when controlling for high-risk socio-medical factors such as low education and unemployment, nationality or migration background no longer influenced PSR outcomes in yet another study (de Vries & Petermann, 2012).

Indicators of patient health history were found to be important predictors of PSR outcomes: Health issues (Reuter et al., 2014), specifically somatoform impairments (Oster et al., 2009; Steffanowski et al., 2005), personality disorders (Reuter et al., 2014), chronification of illness (Reuter et al., 2014; Steffanowski et al., 2005), and comorbidities (Reuter et al., 2014) are associated with less favorable therapy results. In contrast, patients with affective or anxiety disorders show comparatively lower nonresponse rates (Reuter et al., 2014) and tend to benefit more from PSR relative to other diagnostic groups (Steffanowski et al., 2005).

Work-related factors predict treatment outcomes: Prolonged sick leave (de Vries & Petermann, 2012; Lange et al., 2012; Oster et al., 2009), unemployment (Petermann & Koch, 2009; Reuter et al., 2014), a desire to retire (de Vries & Petermann, 2012; Henn et al., 2021; Reuter et al., 2014; Sandweg et al., 2001), and a medically certified incapacity for work at admission (Geiser et al., 2003) worsen PSR outcomes. Conversely, pretreatment employment (de Vries & Petermann, 2012; Lange et al., 2012), a positive subjective prognosis to return to work, and higher work motivation (Kessemeier et al., 2017) are associated with better PSR outcomes.

Attitudes toward PSR are also crucial: Skepticism (de Vries & Petermann, 2012; Lange et al., 2012) and low internal motivation (de Vries & Petermann, 2012; Oster et al., 2009) predict rehabilitation failure, while realistic expectations (Oster et al., 2009) and a motivation to change (Lange et al., 2012) are associated with positive outcomes.

Study Objectives

This study uses a large data set to evaluate the feasibility and effectiveness of ML methods in predicting PSR outcomes. The employment of ML facilitates the incorporation of a vast and heterogeneous array of predictors, thereby enabling a more sophisticated comprehension of the factors that influence rehabilitation outcomes. This study's contributions are twofold:

1. Methodological advancement: The feasibility and accuracy of different types of ML-based predictions

in PSR will be evaluated. Specifically, the random forest (RF) algorithm will be employed to compare the performance results of models containing all individual items of the questionnaires in the data set with models containing only the scale scores that result from aggregating the individual items. This paves the way for in-depth research on the use of ML in PSR.

2. Theory advancement: Predictors of treatment outcomes derived from the biopsychosocial model of the International Classification of Functioning, Disability, and Health (ICF) will be identified. Subsequently, these identified predictors will be compared with prior knowledge and theory to enhance the understanding of the factors that influence rehabilitation outcomes. These insights can be used to improve PSR, which might increase the health and well-being of PSR patients and/or save money.

Methods

Study Design

The data set includes anonymized data from seven German PSR clinics. A customary course of PSR lasts five weeks. The data were collected from April 2019 until January 2021 in a pre-post design, with the two measurement time points being admission to (T1) and discharge from the PSR clinic (T2). The data were collected during routine diagnostics using self-assessment questionnaires.

Data Protection and Ethical Considerations

The patient data were collected by the clinics in accordance with GDPR Article 5 principles and provided to the authors in anonymized form. Patients were informed that their data might be used for research purposes; their privacy was upheld throughout the process. The ethics committee of the University of Freiburg approved the use of the anonymized patient data for scientific purposes through an ethics vote and granted a waiver for the informed consent (Reference Number: 22-1207-S2; 13.05.2022). The study was publicly registered in the German Clinical Trials Register (DRKS-ID DRKS00029669).

Participants

There were data from 7,436 patients, with a range of 629–1,895 patients per clinic. Around two-thirds of the

participants who provided information on their sex were women ($n = 4,831$) and one-third were men ($n = 2,548$). The rehabilitants were on average 50.4 years old ($SD = 10.0$), with the youngest being 18 and the oldest 82 years old.

Measures

We collected data on the HEALTH-49 (Rabung et al., 2009) and the ICF AT-50 Psych (Nosper, 2008) at T1 and T2, as well as the PAREMO-20 (Nübling et al., 2004) and the SIMBO-C (Streibelt et al., 2007) at T1. Sample items can be found in Appendix A in the Electronic Supplementary Material (ESM 1).

The Hamburg Modules for the Assessment of Psychosocial Health in Clinical Practice [HEALTH-49] was used to measure the psychosocial health of the patients at T1 and T2. The HEALTH-49 comprises 49 items across nine scales within six modules. Module A consists of three scales: *somatoform complaints*, *depressive symptoms*, and *phobic fears*. Modules B to E consist of one scale each: *psychological wellbeing*, *interpersonal problems*, *self-efficacy*, and *activity and participation*, respectively. Module F consists of two independent scales: *social support* and *social burden*. Modules A, C, E, and F (only *social burden*) are assessed on a five-point Likert scale, ranging from 1 (*not at all or never*) to 5 (*very much or always*). Higher scores represent poorer psychosocial health. Modules B, D, and F (only *social support*) use a Likert scale ranging from 1 (*never or not true*) to 5 (*always or very true*). Higher scores indicate a more positive health state. To ensure consistency in interpretation, items from Modules B, D, and the social burden subscale of Module F were reverse-coded: Higher values now reflect negative outcomes. Additionally, all item responses were recoded to a range from 0 to 4 instead of 1 to 5.

The ICF-Compliant Questionnaire for the Self-Assessment of Activities and Participation in Mental Disorders [ICF AT-50 Psych] reflected the degree of impairment concerning the functional level of activity and participation at T1 and T2. The ICF AT-50 Psych comprises six subscales, which are aggregated into two scales: *Activity* and *participation*. Activity consists of the three subscales *verbal competences*, *ability to meet demands*, and *fitness and well-being*. Participation consists of the three subscales *social relationships and activities*, *closeness in relationships*, and *consideration of others*. The ICF AT-50 Psych uses a five-point Likert scale, ranging from 1 (*no problem*) to 5 (*full problem*). For consistency, all items were recoded into a range from 0 to 4.

The Patient Questionnaire to Record Motivation for Rehabilitation [PAREMO-20] was assessed at T1. The

PAREMO-20 consists of six scales: *psychological distress*, *body-related restrictions*, *social support and disease gain*, *willingness to change*, *level of information regarding rehabilitation treatments*, and *skepticism*. The PAREMO-20 uses a four-point Likert scale, from 1 (*do not agree*) to 4 (*agree*). Five items were reverse coded, so that higher scores reflect greater motivation for rehabilitation. Additionally, items were recoded to range from 0 to 3.

The Screening Instrument for the Identification of a Demand for Medically-Vocationally Oriented Rehabilitation [SIMBO-C] was used to assess work-related problems at T1. The SIMBO-C consists of seven items with varying response formats. Depending on the response given for each item, a specific number of points is added to the total score. Patients can achieve a score ranging from 0 to 100.

All items were presented in German. When summarizing the items into scale values, missing values were ignored.

HEALTH-49. The *activity* and *participation* levels of the ICF were assessed with the respective scales from the ICF AT-50 Psych.

The values of these five outcomes were calculated in two steps: First, the difference between T1 and T2 was calculated for each outcome. Second, we classified each observation into one of two categories: (1) clinically relevant improvement (= “PSR success”) or (2) no clinically relevant improvement/deterioration (= “PSR failure”), based on predefined critical difference thresholds. For somatoform complaints, depressive symptoms, and phobic fears, we applied empirically validated critical difference scores based on absolute change values, as provided by the HEALTH-49 reference material (Rabung et al., 2016). In contrast, no established thresholds were available for activity and participation. To address this, we defined improvement in these areas using a relative criterion of a minimum 10% symptom reduction from T1 to T2.

Outcome Calculations and Classification

The PSR outcomes are based on the three levels of functioning according to the biopsychosocial model of the ICF (Roßbach et al., 2015): (1) *body functions and structure*, (2) *activity*, and (3) *participation* (WHO, 2002). We subdivided the *body functions and structures* level into three outcomes: (1a) *somatoform complaints*, (1b) *depressive symptoms*, and (1c) *phobic fears*. These three outcomes were derived from three corresponding scales of the

Model Types and Predictors

For each of the five outcomes, two separate binary classifiers were tested: a scale model that used all scale scores from T1 as predictors and an item model that used all individual items from T1 as predictors. Both model types incorporated additional predictor variables, including the treatment year, the clinic where the patient received treatment, sex, and age. Figure 1 provides an overview of all modeling components.

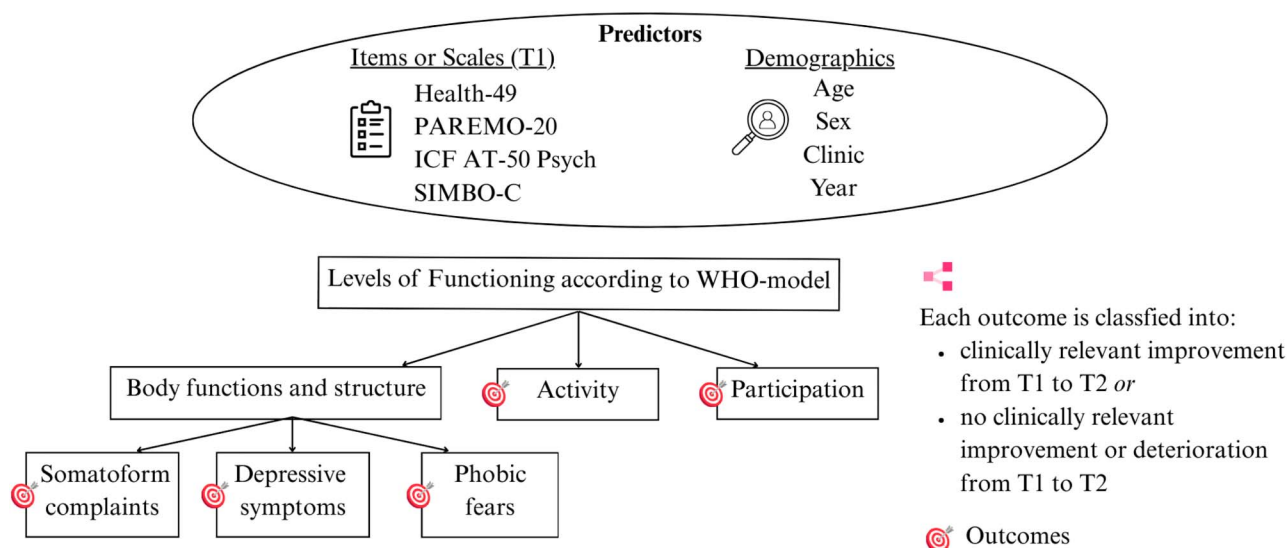


Figure 1. Overview of predictors and outcomes.

Machine Learning Pipeline

We deployed a random forest (RF) classifier (Breiman, 2001), using the *ranger* (Wright et al., 2023) and *caret* (Kuhn et al., 2023) packages. RF is an ensemble learning method that constructs multiple decision trees using bootstrapped samples and combines their predictions by majority vote for classification tasks (Breiman, 2001). RF is well-suited for high-dimensional and noisy data and can flexibly model nonlinear and nonadditive relationships, thereby capturing the complex patterns often present in real-world health data (Strobl et al., 2009). While this is an advantage, RF and the necessary hyperparameter optimization also require greater computational resources than other ML algorithms (e.g., *K*-Nearest Neighbors), which may limit its full potential. RF has already demonstrated strong performance across a variety of domains (Strobl et al., 2009), including the prediction of rehabilitation outcomes. For example, Tschuggnall et al. (2021) reported that an RF classifier outperformed several alternative ML algorithms in a rehabilitation context.

Analyses were conducted in R (R Core Team, 2024), version 4.4.0. RStudio (Posit PBC, Boston, USA) was used as the Integrated Development Environment. The computations were carried out on a Lenovo ThinkPad X1 Extreme Gen 5 with an Intel Core i7-12700H processor (14 cores, 20 threads, 2.3 GHz base clock), 32 GB RAM, running Microsoft Windows 11 Enterprise.

Preprocessing

Rows with missing values were deleted, as the *ranger* implementation cannot handle missing values. This process was done separately for each outcome and each model type (item or scale), which resulted in data sets with different sample sizes for the subsequent ML analyses, ranging from $n = 4,322$ (activity and participation item data set) to $n = 7,152$ (depression scale data set). Some data sets showed a slight class imbalance—ranging from 55% (participation scale) to 63% (somatoform complaints scale) for the most frequent class. The distributions of the outcome classes of the individual data sets can be found in Appendix B in ESM 1. No feature scaling was applied, as the RF algorithm is scale invariant (Ozsahin et al., 2022).

Train/Test Splits

Stratified sampling was used to split the data sets into train and test sets, ensuring that the distribution of the binary outcome classes remained consistent across both subsets. The data were divided, with 70% allocated to training and 30% to testing.

Algorithm Deployment and Tuning

We deployed an RF algorithm for classification using the *ranger* implementation. To optimize performance, we conducted a grid search with 5-fold cross-validation to tune three hyperparameters using the *caret* package:

- *mtry*: The number of variables randomly selected for each node split (set differently for item models [1–35] and scale models [1–22]).
- *splitrule*: The criterion used for splitting (gini or *extraTrees*).
- *min.node.size*: The minimum number of observations required in a terminal node (ranging from 1 to 10).

The number of trees (*num.trees*) was set to 500, as initial tests showed negligible performance gains beyond this number. The ranges of the hyperparameter tuning grids were constrained by available computing power. The optimal hyperparameter configurations were determined by maximizing the accuracy score across all cross-validation folds. The resulting hyperparameter combinations are in Appendix C in ESM 1.

Algorithm Evaluation

We took a threefold approach to ensure an informative and complete assessment of algorithm performance.

Dummy Classifiers

To provide a baseline for comparison, we included a dummy classifier that always predicts the most frequent class for each observation and uses no other predictors. This is implemented as the no information rate (NIR) in the *confusionMatrix* function of the *caret* package in R. Additionally, the package provides a binomial significance test to test the null hypothesis that the accuracy of the ML algorithm is equal to or less than the NIR.

Evaluation Metrics

We mainly used three common evaluation metrics for each classifier: accuracy, specificity, and sensitivity (Sokolova & Lapalme, 2009). Accuracy represents the overall performance of the classifier by calculating the proportion of correctly classified instances of all instances. Specificity indicates how well the algorithm identifies individuals who will not benefit from PSR (true negative), while sensitivity reflects its ability to correctly predict those who will benefit (true positive rate). Additionally, we calculated the F1 score, which is the harmonic mean of sensitivity and precision. Precision represents the proportion of correctly predicted positives (true positives) of

all positive predictions made by the model (true positives plus false positives). Since accuracy was used as the main criterion for model training, our main interpretations focus on accuracy. All confusion matrices can be found in Appendix D in ESM 1.

Predictor Importance

We used permutation-based feature (predictor) importances to estimate the relative contribution of each predictor to model performance. This method assesses importance by measuring the decrease in model accuracy (or another performance metric of choice) when the values of a single predictor are randomly permuted, thereby disrupting its relationship with the outcome. A greater drop in performance indicates a more important predictor. Unlike impurity-based methods, permutation importance is model-agnostic and less biased by the number of categories or their frequency distributions (Strobl et al., 2007). However, it can still be affected by predictor intercorrelations, as permuting one correlated predictor may indirectly affect others, leading to incorrectly estimated importance values (Strobl et al., 2008; Tolosi & Lengauer, 2011). To evaluate the extent to which correlated predictors might bias importance estimates, we additionally computed pairwise Pearson correlation matrices for each of the 10 data sets.

Directionality of Predictor Effects

We calculated accumulated local effects (ALE) plots for the most important predictors of each model, as identified through permutation-based importance scores. ALE plots illustrate how changes in a single predictor affect the model's predictions on average, while accounting for the distribution and interactions of other predictors. ALE plots are unbiased in the presence of correlated predictors (Apley & Zhu, 2020). A positive slope in an ALE plot indicates that increasing the predictor tends to increase the

prediction, while a negative slope suggests a decreasing effect. ALE plots can be found in Appendices E–N in ESM 1.

Results

All RF models predicted clinically relevant PSR success significantly better than a dummy classifier. The NIR ranged from 0.56 (participation scale model) to 0.63 (somatoform complaints scale model), while the RF models had accuracies between 0.63 (participation scale model) and 0.76 (phobic fears item model). All models passed the null hypothesis significance test of accuracy \leq NIR with $p < .001$. The highest percentage point increase in accuracy over NIR was achieved by the phobic fears item model, which yielded an improvement of 15.18 percentage points. Generally, the activity models showed the smallest accuracy gains, followed by the participation and somatoform complaints models. The largest increases were achieved by the depressive symptoms and phobic fear models. The performance of all models is in Table 1.

The item models improved accuracy by about 1–2.3 percentage points more than the scale models for all outcomes except depression, where the scale model had a 1.97 percentage point advantage. The percentage point increase from NIR to accuracy varied from 3.97 (activity) to 14.23 (phobic fears) for the scale models, and from 6.25 (activity) to 15.18 (phobic fears) for the item models.

A sensitivity analysis showed that models that do not use the baseline values of the outcomes still perform significantly better than the NIR.

Feature (predictor) importance analysis revealed that work-related issues, skepticism toward rehabilitation, current physical limitations, and phobic fears were the

Table 1. Model performance

Model	NIR	Accuracy	% point increase	Sensitivity	Specificity	F1
Activity item	.6142	.6767	6.25	.8719	.3660	.7682
Activity scale	.5979	.6376	3.97	.7930	.4065	.7236
Participation item	.5671	.6551	8.80	.8150	.4456	.7280
Participation scale	.5580	.6301	7.21	.7353	.4984	.6886
Depression item	.6291	.7348	10.57	.9118	.4345	.8122
Depression scale	.6135	.7389	12.54	.8815	.5127	.8058
Phobic fear item	.6076	.7594	15.18	.8310	.7132	.7308
Phobic fear scale	.6096	.7519	14.23	.8423	.6940	.7264
Somatoform complaint item	.6261	.7163	9.02	.4701	.8633	.5536
Somatoform complaint scale	.6325	.7071	7.46	.4860	.8355	.5492

Note. NIR = no information rate.

most important predictors of change in activity. These factors also played a key role in predicting change in participation, except phobic fears. Instead, willingness to change emerged as an influential predictor for participation. For change in depression, the baseline scores at T1, activity and participation levels, general well-being, psychological strain, skepticism toward rehabilitation, and difficulties in social interactions were critical predictors. For change in phobic fears, the baseline scores at T1, depressive symptoms, willingness to change, and difficulties in social relationships and activities were most influential. Finally, in predicting change in somatoform complaints, the baseline scores at T1, current physical limitations, skepticism toward rehabilitation, willingness to change, and work-related issues were the pivotal factors.

The pairwise Pearson correlation matrices for each of the 10 data sets (five outcomes \times two predictor sets) revealed consistent median absolute correlations, averaging 0.26 ($SD = 0.04$). This suggests a low to moderate level of intercorrelation among predictors on average.

Discussion

The goals of this study were to (1) assess the feasibility and accuracy of ML-based predictions in PSR research, and (2) enhance PSR theory and practice by identifying key predictors of meaningful changes in outcomes and engage with these findings in the context of existing research to provide deeper insights.

Feasibility and Evaluation of Machine Learning-Based Predictions in Psychosomatic Rehabilitation

Our study is one of the first to explore the feasibility and accuracy of ML-based prediction in PSR research by applying the RF algorithm to routine diagnostic data. We successfully predicted clinically relevant changes in activity, participation, depressive symptoms, phobic fears, and somatoform complaints following standard PSR treatment. All models outperformed dummy classifiers, with accuracy improving from 3.97 percentage points in the activity scale model to 15.18 percentage points in the phobic fear item model. These promising results demonstrate the feasibility of ML-based predictions in PSR and open the door for more in-depth exploration of ML's potential in PSR research.

The majority of the models demonstrated higher sensitivity compared to specificity, indicating that they were

more effective at identifying patients who would benefit from PSR. In contrast, the participation models exhibited higher specificity, making them better suited for identifying patients who would not benefit from PSR. Notably, the phobic fear models achieved high levels of both sensitivity and specificity, indicating a balanced ability to predict both positive and negative treatment outcomes. It is important to note that the trade-off between sensitivity and specificity is dependent on the outcome class distribution and the selected probability thresholds and can thus be adjusted according to the clinical or research objective. Moreover, while the models were trained and tested on a large data set, the observed performance patterns may still be influenced by characteristics unique to this sample, underscoring the need for further validation. These findings highlight the importance of clearly defining model objectives and evaluation criteria during the development of ML applications in clinical settings.

Another interesting finding was that all item models, except for the depressive symptom model, performed better than their corresponding scale model, which relied on scale scores (i.e., aggregated items) as predictors. This underlines the potential benefit of considering individual items rather than scale scores, as is the conventional approach in PSR research. By going deeper onto the item level, the most important predictors of specific outcomes could be identified, paving the way for more refined and personalized PSR treatment. For example, the scale on difficulties in social interactions was one of the most important predictors of the phobic fear scale model. However, the corresponding item model used only two of the seven items of the scale and yet achieved better performance. This showcases that focusing on individual items could enhance efficiency and reduce the burden on patients by minimizing the need for extensive questionnaires.

Key Predictors of Treatment Success

The RF algorithm's ability to handle high-dimensional data and evaluate the importance of individual predictors allowed us to analyze the wide range of variables typically collected in PSR. However, it is worth noting that permutation-based importance scores, as used in the analyses at hand, can be influenced by highly correlated predictors. While we found that median absolute correlations across the data sets were low to moderate, the large number of predictors in the item models inevitably included many subsets of moderately to highly correlated variables. Therefore, the following interpretation of predictor importance results should be approached with caution.

Baseline items at T1 were identified as major predictors for several outcomes, particularly for all depressive symptoms, phobic fears, and somatoform complaints models. This finding aligns with what is commonly observed in clinical practice and is supported by other PSR studies, where the severity of symptoms and the health status at admission are recognized as key indicators of expected improvement (e.g., de Vries & Petermann, 2012; Lange et al., 2012; Reuter et al., 2014). However, baseline values for activity and participation played only an average role in predicting change within those domains in the scale models and no role at all in the respective item models, suggesting that other factors be more influential in driving change in activity and participation.

Patients' skepticism toward rehabilitation and willingness to change were important predictors, which aligns with existing literature (e.g., de Vries & Petermann, 2012; Lange et al., 2012; Oster et al., 2009). At a nuanced level, skepticism was revealed to be a strong predictor of change in activity and participation, moderately important for somatoform complaints and depressive symptoms, and negligible for phobic fears. Willingness to change had no significant role in predicting change in depressive symptoms but was moderately influential for predicting change in activity and participation, and a minor predictor for change in the phobic fear models and in the somatoform complaint scale model. It is noteworthy that the RF algorithm deemed all items on patients' skepticism to be important, but only one of three items on willingness to change.

Several studies have highlighted the importance of work-related factors in predicting PSR outcomes, and the results at hand underline this importance (e.g., de Vries & Petermann, 2012; Geiser et al., 2003; Henn et al., 2021; Petermann & Koch, 2009): Sick leave (current and last 12 months), current employment situation, incapacity to work, and subjective prognosis for future employment and retirement emerged as key predictors of change in activity and participation. These constructs also played a minor to moderately influential role in predicting change in depressive symptoms, phobic fears, and somatoform complaints.

In addition to baseline scores, attitudes, and work-related aspects, psychosocial factors were identified as significant contributors to PSR outcomes. Difficulties in social interaction emerged as moderately influential predictors of change in depressive symptoms and phobic fears. This finding aligns with psychotherapy research, which suggests that limited improvement in interpersonal issues during therapy is associated with poorer outcomes (e.g., Haase et al., 2008). Interestingly, social support had a modest impact on the predictions of all scale models but played no role in the item models. This indicates that the

items related to difficulties in social interaction may sufficiently capture this aspect of health.

Sociodemographics, such as age, sex, treatment year, and clinic did not impact PSR outcomes, corroborating the literature. This finding suggests that a focus be placed on individual health status, psychological factors, and work-related factors when predicting PSR outcomes.

The feature (predictor) importance analysis revealed unique relationships between specific predictors and outcomes that have not been described in the literature so far: For example, phobic fears was one of the most important predictors for change in activity. This makes sense, as pronounced phobic fears are a major obstacle to coping with the demands of everyday life. Furthermore, phobic fears and somatoform complaints together had small to moderate importance in predicting change in participation. Another emerging finding was that the items on depressive symptoms predicted change in phobic fears while the reverse was not the case. As the literature suggests high comorbidity between mood and anxiety disorders (Saha et al., 2020), one would have expected a bidirectional relationship. However, factors other than phobic fears seem to play a larger role in predicting depressive symptoms.

Limitations

Several limitations should be acknowledged when interpreting the findings of this study. First, the available computing power (i.e., a state-of-the-art personal computer) during hyperparameter tuning restricted the grid range. Thus, the number of variables randomly selected for each node split (mtry) was 35 at maximum, which is a fourth of all 139 variables. An exhaustive optimization using all available variables might have revealed even better performance of the item models.

Second, our analyses may have been influenced by bias in predictor importance estimates due to intercorrelations among predictors. Although we observed low to moderate median intercorrelations, the large number of predictors in the item models likely included many highly correlated variables. More advanced approaches, such as conditional permutation importance, could address this issue more robustly (e.g., Strobl et al., 2008). However, implementing such methods was not feasible given our big data sets, large numbers of predictors, and limited computational resources. Future studies with greater computing capacity should consider these techniques to obtain more reliable importance estimates in the presence of correlated predictors.

Moreover, in want of empirically validated difference scores in the ICF AT-50 Psych, we applied a somewhat

arbitrary 10% cutoff for critical differences between T1 and T2 to assess change in activity and participation. This approach might have led to less clearly defined outcome classes in terms of clinical relevance, thus potentially contributing to a poorer performance of the ML models for these two outcomes.

Finally, the analysis was limited to the constructs typically assessed in routine diagnostics, meaning other potentially influential factors outside the scope of standard assessments were not included. However, variables not included in the data set at hand may play an important role in predicting PSR outcomes.

Future Directions

Studies should explore the potential of combining scales with individual items to enhance model performance, especially when most items within a scale are relevant. For instance, in our study, nearly all baseline items for depressive symptoms and phobic fears at T1 were key predictors and could have been aggregated to be entered as a scale score. A hybrid approach like this could reduce computational demands while boosting predictive accuracy.

To optimize model performance, future research should consider incorporating additional factors, such as details of patient history (e.g., prior treatments, chronic conditions, sleep issues), ICD or DSM diagnoses, and more comprehensive sociodemographic data (e.g., education, marital status, income, migration background). Including these predictors in large-scale, prospective studies could uncover factors that increase model accuracy.

Moreover, in-depth model comparisons are needed to better understand how specific predictors influence sensitivity and specificity, beyond the effects of outcome class distribution and probability threshold selection. While both class imbalance and threshold settings are known to affect these metrics, systematically omitting or modifying individual predictors and re-evaluating model performance may help pinpoint which variables most critically contribute to improving sensitivity and specificity. Such analyses can provide more nuanced insights into model behavior and guide the design of clinically meaningful prediction tools.

Looking ahead, the future of ML in PSR research and practice is promising. As computational power grows and “big” device-collected data become available, ML will likely become an even more powerful tool for refining PSR treatment and improving patient outcomes. This study showcases one of the first successful applications of ML to PSR data and provides insights that may drive future research in this area.

Electronic Supplementary Material

The following electronic supplementary material is available at <https://doi.org/10.1027/2151-2604/a000610>

ESM 1. Study particulars:

Appendix A. Overview of constructs.

Appendix B. Outcome classes of the individual data sets.

Appendix C. Hyperparameter combinations for the algorithms.

Appendix D. Confusion matrices.

Appendices E–N. Accumulated local effects plots for the most important predictors.

References

- Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4), 1059–1086. <https://doi.org/10.1111/rssb.12377>
- Benson, T. (2022). Why PROMs and PREMs matter? In T. Benson (Ed.), *Patient-reported outcomes and experience: Measuring what we want from PROMs and PREMs* (pp. 3–12). Springer. <https://doi.org/10.1007/978-3-030-97071-0>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Campagnini, S., Arienti, C., Patrini, M., Liuzzi, P., Mannini, A., & Carrozza, M. C. (2022). Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: A systematic review. *Journal of NeuroEngineering and Rehabilitation*, 19(1), Article 54. <https://doi.org/10.1186/s12984-022-01032-4>
- de Vries, U., & Petermann, F. (2012). Psychosomatische Rehabilitation: Konzepte und Ergebnisse [Psychosomatic rehabilitation: Concepts and results]. *Physikalische Medizin, Rehabilitationsmedizin, Kur- und ortmedizin*, 22(6), 316–322. <https://doi.org/10.1055/s-0032-1327581>
- Deutsche Rentenversicherung. (2018). *Anforderungsprofil für eine stationäre Einrichtung zur medizinischen Rehabilitation von Erwachsenen mit psychosomatischen und psychischen Störungen* [Requirements profile for an inpatient facility for the medical rehabilitation of adults with psychosomatic and mental disorders]. https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Experten/infos_reha_einrichtungen/med_reha_erwachsene_psycho.html
- Deutsche Rentenversicherung. (2024). *Reha-Bericht 2024* [Rehabilitation report 2024]. https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Statistiken-und-Berichte/Berichte/rehabericht_2024.html
- Geiser, F., Bassler, M., Bents, H., Carls, W., Joraschky, P., Kriebel, R., Michelitsch, B., Ullrich, J., & Liedtke, R. (2003). Zusammenhang der Arbeitsunfähigkeit vor Therapiebeginn mit Störungsgrad und Therapieerfolg bei stationären Angstpatienten [Correlation of the inability to work before the start of therapy with disorder severity and therapy success in inpatients with anxiety]. *Psychotherapie, Psychosomatik, Medizinische Psychologie*, 53(3–4), 185–190. <https://doi.org/10.1055/s-2003-38007>
- 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: Journal of the Society for Psychotherapy Research*, 18(5), 615–624. <https://doi.org/10.1080/10503300802192158>

- Henn, J., Kessemeier, F., Kobelt-Pöncke, A., Bassler, M., Schmidt, J., & Nübling, R. (2021). Psychosomatische Rehabilitation bei Patientinnen und Patienten mit Rentenüberlegungen: Reha-Erfolg und therapeutische Beziehung [Rehabilitation outcome and therapeutic alliance of inpatients of psychosomatic rehabilitation with pension request]. *Psychotherapie Psychosomatik Medizinische Psychologie*, 71(8), 311–319. <https://doi.org/10.1055/a-1303-4861>
- Kaier, K., Knecht, J., Nalbach, L., & Körner, M. (2024). The impact of the Covid-19 pandemic on the effectiveness of psychosomatic rehabilitation in Germany. *BMC Health Services Research*, 24(1), Article 719. <https://doi.org/10.1186/s12913-024-11170-1>
- Kessemeier, F., Stöckler, C., Petermann, F., Bassler, M., Pfeiffer, W., & Kobelt, A. (2018). Die Bedeutung von Arbeitsmotivation für den Reha-Erfolg [The importance of work motivation for successful rehabilitation]. *Die Rehabilitation*, 57(4), 256–264. <https://doi.org/10.1055/s-0043-118196>
- Kleineberg-Massuthe, H., Papst, L., Bassler, M., & Köllner, V. (2023). Milieu-specific differences in symptom severity and treatment outcome in psychosomatic rehabilitation in Germany. *Frontiers in Psychiatry*, 14, Article 1198146. <https://doi.org/10.3389/fpsyt.2023.1198146>
- Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2023). caret: Classification and regression training (Version 6.0.94) [Software]. <https://cran.r-project.org/web/packages/caret/>
- Lange, M., Franke, W., & Petermann, F. (2012). Wer profitiert nicht von der psychosomatischen Rehabilitation? [Who does not profit from psychosomatic rehabilitation?]. *Die Rehabilitation*, 51(6), 392–397. <https://doi.org/10.1055/s-0032-1304612>
- Möske, M.-O., Schneider, J., Koch, U., & Schulz, H. (2008). Beeinflusst der türkische Migrationshintergrund das Behandlungsergebnis? [Does the Turkish migration background influence the treatment outcome?]. *Psychotherapie, Psychosomatik, medizinische Psychologie*, 58(3-4), 176–182. <https://doi.org/10.1055/s-2008-1067352>
- Nosper, M. (2008). ICF AT-50 Psych. Entwicklung eines ICF-konformen Fragebogens für die Selbstbeurteilung von Aktivitäten und Teilhabe bei psychischen Störungen [ICF AT-50 Psych. Development of an ICF-compliant questionnaire for the self-assessment of activities and participation in patients with mental disorders]. *DRV-Schriften*, 77, 127–128.
- Nübling, R., Kriz, D., Herwig, J., Wirtz, M. A., Fuchs, S., Hafen, K., & Bengel, J. (2004). *Patientenfragebogen zur Erfassung der REHA-Motivation (PAREMO-20) Kurzmanual* [Patient questionnaire to record motivation for rehabilitation (PAREMO-20) short manual]. ResearchGate. https://www.researchgate.net/publication/238703037_Patientenfragebogen_zur_Erfassung_der_Reha-Motivation_PAREMO-20_KURZMANUAL
- Nübling, R., Schmidt, J., Bassler, M., & Schulz, H. (2020). Evaluation psychosomatischer Rehabilitation [Evaluation of psychosomatic rehabilitation]. In V. Köllner & M. Bassler (Eds.), *Praxishandbuch psychosomatische Medizin in der Rehabilitation* (pp. 425–437). Elsevier. <https://doi.org/10.1016/B978-3-437-22611-3.00013-4>
- Oster, J., Müller, G., & Wietersheim, J. (2009). “Wer profitiert?” – Patientenmerkmale als Erfolgsprädiktoren in der psychosomatischen Rehabilitation [“Who benefits?” – Patient characteristics as predictors of success in psychosomatic rehabilitation]. *Die Rehabilitation*, 48(2), 95–102. <https://doi.org/10.1055/s-0029-1214411>
- Ozsahin, D. B., Mustapha, M. T., Mubarak, A. S., Ameen, Z. S., & Uzun, B. (2022). Impact of feature scaling on machine learning models for the diagnosis of diabetes. *Proceedings of the International Conference on Artificial Intelligence in Everything* (pp. 87–94). IEEE. <https://doi.org/10.1109/AIE57029.2022.00024>
- Papst, L., & Köllner, V. (2022). Using machine learning to investigate earning capacity in patients undergoing psychosomatic rehabilitation – A retrospective health data analysis. *Frontiers in Psychiatry*, 13, Article 1039914. <https://doi.org/10.3389/fpsyt.2022.1039914>
- Petermann, F., & Koch, U. (2009). Psychosomatic rehabilitation: Quo vadis?. *Rehabilitation*, 48(5), 257–262. <https://doi.org/10.1055/s-0029-1239550>
- R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rabung, S., Harfst, T., Kowski, S., Koch, U., Wittchen, H. U., & Schulz, H. (2009). Psychometrische Überprüfung einer verkürzten Version der “Hamburger Module zur Erfassung allgemeiner Aspekte psychosozialer Gesundheit für die therapeutische Praxis” (HEALTH-49) [Psychometric analysis of a short form of the “Hamburg Modules for the Assessment of Psychosocial Health” (HEALTH-49)]. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, 55(2), 162–179. <https://doi.org/10.13109/zptm.2009.55.2.162>
- Rabung, S., Harfst, T., Koch, U., & Schulz, H. (2016). “Hamburger Module zur Erfassung allgemeiner Aspekte psychosozialer Gesundheit für die therapeutische Praxis (HEALTH)” – Referenzdaten zur verkürzten 49-Item-Version “HEALTH-49” [“Hamburg Modules for the Assessment of Psychosocial Health in Clinical Practice (HEALTH)” – Reference data for the shortened 49-item form “HEALTH-49”] [PowerPoint slides]. <http://hamburger-module.de/download/health-49-normen.pdf>
- Reuter, L., Bengel, J., & Scheidt, C. E. (2014). Therapie-Non-Response in der psychosomatischen Krankenhausbehandlung und Rehabilitation – Eine systematische Übersicht [Therapy non-response in psychosomatic hospital treatment and rehabilitation – A systematic review]. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, 60(2), 121–145. <https://doi.org/10.13109/zptm.2014.60.2.121>
- Roßbach, G., Weinbrenner, S., Brüggemann, S., Martin, S., & Rose, A. (2015). Die Bedeutung psychischer Erkrankungen aus der Perspektive der Deutschen Rentenversicherung [The significance of mental disorders from the perspective of the German Pension Insurance]. *RVaktuell*, 5/6, 114–124. https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Zeitschriften/RVaktuell/2015/Artikel/heft_5-6.pdf
- Saha, S., Lim, C. C. W., Cannon, D. L., Burton, L., Bremner, M., Cosgrove, P., Huo, Y., & J McGrath, J. (2021). Co-morbidity between mood and anxiety disorders: A systematic review and meta-analysis. *Depression and Anxiety*, 38(3), 286–306. <https://doi.org/10.1002/da.23113>
- Sandweg, R., Bernardy, K., & Riedel, H. (2001). Prädiktoren des Behandlungserfolges in der stationären psychosomatischen Rehabilitation muskuloskelettaler Erkrankungen [Predictors of treatment success in inpatient psychosomatic rehabilitation for musculoskeletal disorders]. *Psychotherapie, Psychosomatik, Medizinische Psychologie*, 51(9-10), 394–402. <https://doi.org/10.1055/s-2001-16902>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Statistisches Bundesamt. (2023). *Statistischer Bericht – Grunddaten der Vorsorge- oder Rehabilitationseinrichtungen* [Statistical report – Basic data on preventive care or rehabilitation facilities]. <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Vorsorgeeinrichtungen-Rehabilitationseinrichtungen/Publikationen/Downloads-Vorsorge-oder-Reha/statistischer-bericht-grunddaten-vorsorge-reha-2120612237005.html>
- Steffanowski, A., Löschmann, C., Schmidt, J., Wittmann, W. W., & Nübling, R. (2005). Meta-Analyse der Effekte stationärer psychosomatischer Rehabilitation: MESTA-Studie [Meta-analysis of the effects of inpatient psychosomatic rehabilitation: MESTA study]. *Zeitschrift für Psychotherapie Psychosomatik*

- Medizinische Psychologie*, 55, Article S_051. <https://doi.org/10.1055/s-2005-863397>
- Streibelt, M., Gerwin, H., Hansmeier, T., Thren, K., & Müller-Fahrnow, W. (2007). SIMBO: A screening instrument for identification of work-related disabilities – Analyses of construct and prognostic validity. *Die Rehabilitation*, 46(5), 266–275. <https://doi.org/10.1055/s-2007-970583>
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9, Article 307. <https://doi.org/10.1186/1471-2105-9-307>
- Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources, and a solution. *BMC Bioinformatics*, 8, Article 25. <https://doi.org/10.1186/1471-2105-8-25>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, 14(4), 323–348. <https://doi.org/10.1037/a0016973>
- Tolosi, L., & Lengauer, T. (2011). Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics*, 27(14), 1986–1994. <https://doi.org/10.1093/bioinformatics/btr300>
- Tschuggnall, M., Grote, M., Pirchl, M., Holzner, B., Rumpold, G., & Fischer, M. (2021). Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures. *Informatics in Medicine Unlocked*, 24, Article 100589. <https://doi.org/10.1016/j.imu.2021.100589>
- von Hörsten, N., Schulz, W., Gissendanner, S. S., & Schmid-Ott, G. (2019). Geschlechterunterschiede im Verlauf und Erfolg psychosomatischer Rehabilitation [Sex differences in the course and success of psychosomatic rehabilitation]. *Zeitschrift für Physikalische Medizin, Rehabilitationsmedizin, Kurortmedizin*, 29(4), 190–198. <https://doi.org/10.1055/a-0852-3471>
- World Health Organization. (2002). *Towards a common language for functioning, disability and health: ICF*. <https://www.who.int/classifications/icf/icfbeginnersguide.pdf>
- Wright, M. N., Wager, S., & Probst, P. (2023). *Ranger: A fast implementation of random forests* (Version 0.16.0) [Software]. <https://cran.r-project.org/web/packages/ranger/>

History

Received April 14, 2025

Revision received September 4, 2025

Accepted September 16, 2025

Published online November 19, 2025

Conflict of Interest

The authors declare that there is no conflict of interest.

Publication Ethics

The patient data were collected by the clinics in accordance with GDPR Article 5 principles and provided to the authors in anonymized form. Patients were informed that their data might be used for research purposes; their privacy was upheld throughout the process. The ethics committee of the University of Freiburg provided its approval for the utilization of anonymized patient data for scientific purposes through an ethics vote.

Authorship


Marcel E. K. F. Holzer: conceptualization, data curation, formal analysis, methodology, software, visualization, writing – original draft, writing – review & editing; Natalie M. Hogh: conceptualization, formal analysis, methodology, software, visualization, writing – original draft, writing – review & editing; Paul-Gerrit Velthuysen: conceptualization, methodology, writing – review & editing; Mirjam Körner: conceptualization, investigation, project administration, writing – review & editing; Anja S. Göritz: conceptualization, methodology, supervision, writing – review & editing.

Funding

Open access publication of this article was supported by the publication fund of the University of Augsburg, Germany.

ORCID


Marcel E. K. F. Holzer

 <https://orcid.org/0009-0004-8202-6374>

Natalie M. Hogh

 <https://orcid.org/0009-0001-1486-4040>

Paul-Gerrit Velthuysen

 <https://orcid.org/0009-0001-3653-4333>

Mirjam Körner

 <https://orcid.org/0000-0002-7719-278X>

Anja S. Göritz

 <https://orcid.org/0000-0002-4638-0489>

Marcel E. K. F. Holzer

Behavioral Health Technology

University of Augsburg

Alter Postweg 101, Room 7022

86159 Augsburg

Germany

marcel.holzer@uni-a.de