Psychotherapy Research, 2025
https://doi.org/10.1080/10503307.2025.2517567

RESEARCH ARTICLE

Targeted Learning for Optimal Patient Assignment to Psychotherapy

VEERA K. MALKKI [1,2], SUOMA E. SAARNI [1,3,4], WOLFGANG LUTZ [5], & TOM H. ROSENSTÖM [1,2]

[1]*Psychiatry, Helsinki University Hospital, Helsinki, Finland;* [2]*Department of Psychology, Faculty of Medicine, University of Helsinki, Helsinki, Finland;* [3]*Faculty of Medicine and Health Technology, University of Tampere, Tampere, Finland;* [4]*Family and Social Services, Psychiatry, Well-being Services County of Päijät-Häme, Finland &* [5]*Department of Psychology, Trier University, Trier, Germany*

(Received 13 May 2024; revised 26 May 2025; accepted 30 May 2025)

Abstract
Objective: Previous studies often fell short in identifying differences in treatment effects between psychotherapeutic frameworks. Instead of focusing on the overall treatment effects, we aimed to identify the effects of individually optimal treatment choice [cf. treatment personalization].
Method: We used a causal-inference machine learning (i.e., targeted learning) framework to estimate effects from observational data obtained from the Finnish Psychotherapy Quality Registry, which includes adult patients diagnosed with various mental disorders (n = 2255). Our objective was to estimate the difference in average treatment outcomes between the optimal individualized treatment and a randomly allocated treatment (i.e., the average of all treatment options). Outcomes were changes in self-assessed symptom scores and clinician-assessed functioning. In addition, we estimated counterfactual total-population outcomes for psychodynamic, solution-focused, cognitive–behavioral, and integrative or cognitive-analytic therapies.
Results: Compared to the average treatment effects, the counterfactual optimal treatment produced 0.28–0.29 standard deviations larger benefits for all the outcomes (confidence intervals between 0.20–0.39). Assuming all patients underwent psychotherapy within a single framework, treatment effects on symptom scores were similar across frameworks, but some differences emerged for change in therapist-assessed functioning.
Conclusion: Identifying optimal treatment rules for psychotherapy frameworks is feasible and may significantly improve outcomes.

Keywords: personalized treatment; optimal individualized treatment rule

Clinical or methodological significance of this article: We used a targeted learning framework on Finnish registry data to estimate the optimal psychotherapy framework for each individual and found enhanced treatment effects overall when all patients received the treatment identified as optimal for them. Our results indicate that algorithmic tools aimed at aiding clinicians in optimal treatment selections for individual patients hold promise for improving psychotherapy services.

Objective

Precision mental health care aims to improve treatment outcomes by utilizing patients' background and process data to individualize treatment processes, drawing on ideas from precision (i.e., personalized) medicine (Barkham et al., 2021; Lutz et al., 2022). Both targeting (Cohen & DeRubeis, 2018;

Correspondence concerning this article should be addressed to Veera K. Malkki, Department of Psychology, Faculty of Medicine, University of Helsinki, PL 21 (Haartmaninkatu 3), 00014 Helsinki, Finland. Email: veera.malkki@helsinki.fi

Deisenhofer et al., 2018; Schwartz et al., 2021) and tailoring (Boswell et al., 2015; Lambert et al., 2018; Lutz et al., 2022) treatments have the potential to improve treatment outcomes (Nye et al., 2023). This article pertains to targeting in the sense of prescribing treatments optimally.

Algorithms and statistical methods have often outperformed clinical judgment in outcome prediction (Ægisdóttir et al., 2006; Grove et al., 2000). For instance, identifying patients at risk of treatment failure is more accurate when using questionnaires and decision rules than relying on clinicians' judgments (Hannan et al., 2005). Machine learning methods offer a flexible and data-driven approach to predict outcomes for individual patients and have thus gained significant traction in psychotherapy research (Aafjes-van Doorn et al., 2021; Chekroud et al., 2021), particularly in the area of personalized treatment selection (Cohen & DeRubeis, 2018). However, achieving sufficient power for multivariate prediction models requires large sample sizes —a criterion that is often challenging to fulfill (Aafjes-van Doorn et al., 2021; Lorenzo-Luaces et al., 2021). When this condition is met, personalization algorithms have improved outcomes (Deisenhofer et al., 2018; Delgadillo & Gonzalez Salas Duhne, 2020; Schwartz et al., 2021). Although most studies on personalized treatment allocation rely on statistical estimates in observational data (Salazar de Pablo et al., 2021), experiments also support personalization (Delgadillo et al., 2021; Lutz et al., 2022).

Various data-driven approaches are available to determine the most suitable treatment for individual patients. Some of the most influential in the field of psychotherapy research are the nearest-neighbors -approach, where therapeutic outcomes of each patient are predicted based on previous progress of other patients resembling most closely the patient in question (Lutz et al., 2005, 2019) and the personalized advantage index (PAI) which allows comparisons of predicted outcomes for a patient in two different treatments (DeRubeis et al., 2014). Both can be estimated using various methods, including machine-learning models. However, many methods rely on single models that are susceptible to misspecification –wrong assumptions about the data-generating mechanism– potentially leading to biased estimates (Smith et al., 2023; van der Laan & Rose, 2011, chapter 1.2, page 8–9). In contrast, more flexible semiparametric ensemble-based approaches can reduce this risk by automatically detecting useful models from a set of alternatives (van der Laan & Starmans, 2014).

The question of optimal treatment choice fundamentally revolves around understanding the causal effects of treatments, often addressed through randomized controlled trials (RCTs). However, the ecological validity of RCTs can be limited by several factors, highlighting the need for complementary approaches alongside clinical trials (Dahabreh & Bibbins-Domingo, 2024; Kazdin, 2022). The core challenge in causal inference stems from missing data, as it is impossible to observe the outcomes an individual would have experienced under an alternative treatment (i.e., counterfactual or potential outcomes, Imbens & Rubin, 2015). In RCTs, causal inference relies on the balance between treatment and control groups achieved through randomization, ensuring that the observed effects are attributable to the intervention rather than confounding factors. In observational studies, the balance between measured covariates across treatment settings can be achieved with computational methods, e.g., with propensity scores (Rosenbaum & Rubin, 1983) and inverse probability weighting (Bettega et al., 2024).

Targeted learning is a statistical framework that integrates machine learning and statistical inference, aiming to produce statistically robust, minimally biased, and interpretable estimates that can be used for causal inference in observational research (van der Laan & Rose, 2011; van der Laan & Starmans, 2014). The framework leverages the targeted maximum likelihood estimation (TMLE) technique, combining flexible, data-driven semi-parametric machine learning to adjust a target parameter for confounding (Gruber & van der Laan, 2009; Luque-Fernandez et al., 2018). This approach allows for realistic assumptions about complex data while producing interpretable estimators (e.g., including estimates of confidence intervals). Compared to other competing methods, the statistical properties of TMLE make it superior to generating causal estimates from observational data (Schuler & Rose, 2017; van der Laan & Rose, 2011; Table 6.7, page 118). The advantages of TMLE likely account for its growing adoption in treatment personalization within psychiatry (Benjet et al., 2023; Kessler et al., 2022; Zainal et al., 2024) and related fields more broadly (Smith et al., 2023).

In this study, we aim to estimate counterfactual average treatment effects (ATE) of different psychotherapy frameworks and to identify decision rules for optimal treatment selection for individual patients. This is achieved using baseline characteristics from comprehensive naturalistic psychotherapy data on over 2000 patients, a supervised machine learning approach (i.e., predictive modeling), and a targeted learning framework. As a preliminary research question, we (1) estimate the counterfactual treatment outcome for each psychotherapy framework, assuming all patients would have been allocated to that treatment class ("counterfactual treatment-specific means'). That is, we compare whether any of the included psychotherapy frameworks generally produce better results than other frameworks. We hypothesize that each framework would produce similar treatment effects on average according to Dodo's verdict (Rosenzweig, 1936).

Our (2) main research question is: should everyone receive their optimal treatment, would they achieve better treatment effects on average, and how much better? We define optimal treatment as the psychotherapy framework that would most effectively alleviate the patient's symptoms and enhance their psychosocial functioning compared to other available frameworks. We estimate the potential benefits of optimizing treatment allocation by comparing outcomes in a scenario where everyone would have received their optimal treatment to a scenario where treatment is assigned randomly, with equal probability for each framework. Our hypothesis is that the optimal treatment shows superior outcomes compared to the randomized scenario (and compared to the real-world scenario using individual patients' estimated allocation probabilities). In this case, we (3) explore which variables most impact who benefits from what treatment.

Method

Sample

The study sample (n = 2255) consisted of adult patients referred to individual short psychotherapy [max. 20 sessions, longer treatments (n = 139) were removed from the sample] from primary or psychiatric-specialty care in Southern Finland, serving 20% of the total Finnish population. We included only patients who had completed their psychotherapy (interrupted n = 188). Participants aged 18–89 (mean = 38.2, SD = 14.6), and 79.5% were female. Most of the patients were diagnosed with mood or anxiety disorders, experienced mild or moderate symptoms, and were referred to a short psychotherapy from primary care. We included patients from different diagnostic categories, except those with already restrictive treatment recommendations (obsessive-compulsive disorder, schizophrenia).

Baseline characteristics were similar across different frameworks (see Table I and Supplementary Tables S1 and S2). With TMLE, the distributions of baseline characteristics do not need to be equal across the treatment groups, as it utilizes propensity

Table I. Characteristics of Adult Patients from the Finnish Psychotherapy Quality Registry (FPQR) by Different Psychotherapeutic Frameworks (n = 2255).

	All (n = 2255)		CBT (n = 801/ 35.5%)		Solution-focused (n = 643/ 28.5%)		Integrative and CAT (n = 525/ 23.3%)		Psychodynamic (n = 286/ 12.7%)	
Variable	Mean/ %	*SD*	Mean/ %	*SD*	Mean/ %	*SD*	Mean/ %	*SD*	Mean/ %	*SD*
Age (years)	38.2	14.6	36.6	14.3	39.3	14.1	39.1	15.1	38.5	15.1
Sex (% women)	79.5		75.7		80.4		83.8		80.4	
Site of referral (%)										
Primary-care referrals	57.3		59.4		57.5		53.7		57.3	
Specialty-care referrals	42.7		40.6		42.5		46.3		42.7	
Delay in accessing treatment (days)	101.6	99.6	104.7	95.9	94.7	97.8	107.6	104.3	97.5	104.2
Main diagnosis, grouped (ICD-10) (%)										
Depression (F32, F34, F38, F39)	33.5		29.0		36.9		34.7		36.7	
Recurrent depression (F33)	12.9		11.4		15.4		12.2		12.6	
Mixed anxiety & depression (F41.2)	11.0		10.2		10.4		12.6		11.9	
Anxiety disorders (F41.1–F41.9, F44, F45, F48)	22.8		25.3		19.8		23.8		20.6	
Panic disorder (F41.0–F41.09)	2.9		5.2		1.7		1.5		1.4	
Stress-related disorders (F43)	9.0		7.2		10.6		9.3		9.4	
Phobias (F40)	2.4		4.6		0.6		1.1		2.8	
Other	5.5		7.0		4.7		4.8		4.5	
Baseline-measures of										
SOFAS	67.5	11.6	66.8	12.3	67.3	11.8	69.0	10.4	67.6	10.5
PHQ-9	9.5	5.4	9.2	5.2	9.8	5.3	9.4	5.4	9.4	5.6
OASIS	9.2	3.7	9.3	3.7	9.2	3.8	9.0	3.6	9.0	3.7
CORE-10	15.5	6.5	15.3	6.3	16.1	6.4	15.3	6.7	15.5	6.6
AUDIT-C	1.8	1.4	1.9	1.4	1.8	1.4	1.8	1.3	1.9	1.5

CORE-10 = the Clinical Outcomes In Routine Evaluation −10 (scale = 0–40), PHQ-9 = the Patient Health Questionnaire (scale = 0–27), SOFAS = The Social and Occupational Functioning Assessment Scale (scale = 0–100), OASIS = The Overall Anxiety Severity and Impairment Scale (scale = 0–20), AUDIT-C = The Alcohol Use Disorders Identification Test Consumption (scale = 0–12). SD = Standard deviation.

score-based methods embedded within the double-robust estimation process to adjust for confounding (Schuler & Rose, 2017). We included patients from four different therapy frameworks: solution-focused therapy (n = 643), cognitive behavioral therapy (CBT, n = 801), integrative and cognitive-analytic therapies (n = 525), and psychodynamic therapy (n = 286). Combining cognitive analytic and integrative psychotherapies is a viable solution, given their shared theoretical foundations and training methodologies evident in this Finnish context. Furthermore, neither would constitute a large enough sample on their own.

Treatments

The patient's situation was evaluated by a doctor, who then wrote a referral to short-term psychotherapy. After this, the patient could receive personal guidance to help choose a preferred framework. The treatments were administered as part of normal healthcare services, and their content structure was planned by the responsible psychotherapists who followed national treatment guidelines. In Finland, "psychotherapist" is a protected profession, meaning that all practicing psychotherapists must be licensed by the National Supervisory Authority for Welfare and Health. The length of the treatment was determined by the patient and the therapist, up to a maximum of 20 sessions. Both patients and therapists completed separate questionnaires at the beginning and at the end of the treatment to track outcomes.

The therapists (n = 409) reported the psychotherapy framework they generally practiced, determined by the theoretical background of their formal psychotherapy training. Therefore, it is likely that some content overlap between the treatment groups exists, as therapists can apply methods. Regardless, the included psychotherapy frameworks (1-4) have some main distinctive features. (1) Integrative psychotherapists flexibly use perspectives from different therapeutic frameworks to match the treatment to the individual patient's needs (Kuusinen et al., 2020). Their focus is on building the therapeutic alliance, strengthening the patient's agency and self-perception, and creating positive interactions. In Finland, similar principles are present in cognitive-analytic therapy, which is rooted in the psychoanalytic object relations theory and stresses how the self is formed in the developmental processes embedded in social relationships (Kalska & Särkelä, 2020). According to the core idea of (2) cognitive–behavioral therapy, the patient's feelings, thoughts, actions, and physical reactions form an ongoing interactive process that gives rise to psychological experiences (Newman et al., 2021). Therapists guides the patient to re-evaluate and change this system by affecting its parts, for example, with cognitive restructuring and behavioral changes. In (3) psychodynamic psychotherapies, the unconscious processes rooted in childhood, past unresolved conflicts, and interpersonal relationships are identified, modified, and integrated into the adult personality to alleviate mental health problems (Barber et al., 2021). In contrast, the (4) solution-focused therapy emphasizes future-oriented thinking, aiming to translate the patient's current problems into treatment goals and to reinforce the existing strengths (Aarnisalo & Mattila, 2020). The therapist acts as an equal participant in the therapy process, concentrating on finding solutions for the problems the patient is experiencing. The therapists were nested within frameworks, i.e., represented only one kind of framework.

Registries

The data was sourced from the Finnish Psychotherapy Quality Registry (FPQR) under the Hospital District of Helsinki and Uusimaa (HUS), which serves as a platform for planning personalized patient care and assessing psychotherapy accessibility and effectiveness (for more details, see Saarni et al., 2022). The registry data was collected from June 12th, 2018, to September 28th, 2022. This study was conducted as a part of the research projects "Machine learned causal factors for psychotherapy" and "Availability, effectiveness, and quality of psychosocial care", and it has received authorization from the HUS Ethical Review Board (HUS/1861/2020 & HUS/3150/2020) and Findata, the national social and health data permit authority (THL/4810/14.02.00/2020 & THL/1303/14.06.00/2023).

We also combined data from two other registries. HILMO is a care register maintained by the Finnish Institute of Health and Welfare, which contains information from the treatment notification system about all previous specialty healthcare visits. From there, we derived information about previously diagnosed mental health and somatic disorders. In addition, we used the register on socioeconomic status (SES) data coordinated by Statistics Finland. This administrative SES variable complemented the self-reported background SES variable from FPQR.

Predictors and Outcomes

We quantified the treatment outcomes as a change in psychological distress measured by patient-reported CORE-10 (Clinical Outcomes In Routine

Evaluation; Barkham et al., 2013), depression symptom severity measured by patient-reported PHQ-9 (Patient Health Questionnaire; Kroenke et al., 2001), and psychosocial functioning measured by therapist-rated SOFAS (Social and Occupational Functioning Assessment Scale; Goldman et al., 1992). Previous research has demonstrated the satisfctory psychometric properties of CORE-10 (Barkham et al., 2013; Rosenström et al., 2022), PHQ-9 (Kroenke et al., 2001; Stochl et al., 2020), and SOFAS (Hilsenroth et al., 2000; Jovanović et al., 2008; Smith et al., 2011).

We included 21 predictors in our models to predict counterfactual treatment effects and optimal treatment assignment. When dummy-coding the nominal variables, 50 predictors were obtained. Predictive baseline variables were selected according to general recommendations of the TMLE literature (e.g., Phillips et al., 2023). That is, we employed a broad set of predictors together with algorithms capable of pruning them, complemented by some degree of *a priori* pruning based on expert judgment. For example, we used sum scores rather than the highly correlated items, but in the case of social support, separated the construct from general functioning based on its known role in predicting psychotherapy outcomes (Roehrle & Strouse, 2008; Zimmermann et al., 2021). All the authors verified the relevance of all the baseline covariates.

Demographics included in the models were age, gender, work status, and SES. Predictors reflecting clinical status of the patients were diagnosis, comorbidity, site of referral, delay of the treatment onset, patient-reported use of psychotropic medication, information of how the patient chose the therapist, and measures of baseline symptom severity. Symptom severity was assessed with sum scores of CORE-10, PHQ-9, SOFAS, AUDIT-C (The Alcohol Use Disorders Identification Test Consumption; Bush et al., 1998), and OASIS (The Overall Anxiety Severity and Impairment Scale; Norman et al., 2006). OASIS is a measure for anxiety symptoms and anxiety-related functioning, and AUDIT-C consists of three questions measuring the amount and frequency of alcohol consumption.

Predictors assessing societal factors included information on living alone and on experience of social support (single items from CORE-OM, Q1: "I have felt terribly alone and isolated.", scale = 0–4, 0 = "Not at all", 4 = "Most or all the time"; Q3: "I have felt I have someone to turn to for support when needed." (R); Q19: "I have felt warmth or affection for someone." (R); Q26: "I have thought I have no friends."). (See Table I and Supplementary Table S1 for more information on the predictors.)

Data Analytic Plan

Targeted learning framework. Targeted learning is a statistical framework that can be employed to estimate the population-mean effect of treatments that maximize outcomes for each individual patient (Coyle et al., 2020). Consequently, it may serve to formulate Individualized Treatment Rules (ITR, van der Laan & Luedtke, 2015) and test individualized treatment recommendations. TMLE (Gruber & van der Laan, 2009) is the core of targeted learning; an algorithm combining the estimation of the treatment assignment mechanism (e.g., propensity scores) and the outcome model while targeting the estimation process towards the specific parameter of interest, such as the ATE, using efficient influence curve to control variance (Hines et al., 2022). TMLE estimates are double robust and efficient for the true (target) parameter of interest, thus possessing several desirable statistical properties (van der Laan & Rose, 2011; van der Laan & Starmans, 2014).

Targeted learning employs a stacking method called super learner, which can combine multiple machine learning and statistical models into a single predictive algorithm. The super learner uses cross-validation to avoid overfitting and generates a weighted average, or an ensemble algorithm, which is asymptotically guaranteed to perform better than any single algorithm included (Naimi & Balzer, 2018; Polley & van der Laan, 2010). In the estimation process of TMLE, two separate super learners provide estimates for the treatment assignment mechanism (treatment conditioned on covariates) and the outcome mechanism (outcome conditioned on treatment and covariates). The inverse probabilities of receiving a particular treatment are used as weights in the outcome estimation, simulating a situation that closely resembles randomization and thus supports causal interpretation.

Individual learners in the super learner models were linear regression, lasso-, ridge-, and elastic net regressions, random forest, extreme gradient boosting, multivariate adaptive regression splines, and support vector machines, each of which we had a few versions with different tuning parameters. The included algorithms utilize different computational techniques with various qualities and strengths. Due to the data requirements in our analysis, we removed all the missing values from our sample (Phillips et al., 2023). (See Supplementary materials p.2 for data preprocessing and pp. 15–18 for super learner models.)

Estimating counterfactual treatment effects. We used TMLE to estimate *(i)* the counterfactual treatment effects (i.e., counterfactual outcomes for each patient in every treatment arm) and *(ii)* the treatment effect under optimal treatment choice. The counterfactual treatment-specific means for each framework may reveal selection effects relative to raw framework means. The ATE for optimal treatment choice was estimated as the mean difference in outcomes between a scenario where everyone received their optimal treatment and one where treatment was assigned randomly ($E[Y_{\text{optimal treatment}}] - E[Y_{\text{random treatment}}]$). The optimal individualized treatment was defined separately for each outcome variable as the treatment predicted to produce the largest improvement in the outcome in question. (See Supplements pp. 8–11 on targeted learning and our data analysis process.)

The random choice of a treatment was made with uniform probability across the treatment frameworks (agnostic allocation), but we conducted a sensitivity analysis in which the random treatment was selected based on the patient-specific (estimated) probability of receiving that particular treatment (i.e., here the treatments corresponded in mean to what each patient had in fact experienced). In addition, we estimated the contribution of each predictor to the predictions of individualized treatment models using SHAP values [Shapley Additive Explanations, (Lundberg & Lee, 2017)]. SHAP values measure each variable's influence on the model's predictions, with larger values indicating a stronger positive impact. The analyses were done with R-Studio version 4.2.2 (R Core Team, 2022) and targeted learning software ecosystem tlverse (van der Laan et al., 2023; van der Laan et al., 2023). (See Supplementary Table S10 and Table S20 for additional information on used R packages.)

Results

The improvement in CORE-10 score from baseline to the end of the treatment was on average 4.20 points ($t(2080) = 30.31$, $p = < .001$, Cohen's $d = 0.66$, 95% CI = 0.62–0.72), 2.91 points in PHQ-9 ($t(2083) = 27.55$, $p = < .001$, Cohen's $d = 0.60$, 95% CI = 0.56–0.65), and 9.40 points in SOFAS ($t(2186) = -44.39$, $p = < .001$, Cohen's $d = 0.95$, 95% CI = 0.90–1.00). The Pearson's correlations for observed change scores between the SOFAS and CORE-10 were 0.26 (95% CI = 0.22–0.30), 0.23 between the SOFAS and PHQ-9 (95% CI = 0.18–0.27), and 0.69 between the CORE-10 and PHQ-9 (95% CI = 0.66–0.71). (See Supplementary Table S3 for framework-specific effect sizes and symptom measures.)

Super learner models predicting treatment outcomes accounted for 17% of the observed variance in the change of CORE-10 scores, 12% of the outcome variance in PHQ-9, and 27% of the outcome variance in SOFAS, as measured by the cross-validated coefficient of determination (R^2). The accuracy of the cross-validated super learner models predicting treatment allocation was 0.85 (95% CI = 0.83–0.87) without CORE-10 as a predictor (this super learner model was later combined with the super learner model predicting CORE-10 as treatment outcome), 0.77 without PHQ-9 as a predictor (95% CI = 0.75–0.79), and 0.71 without SOFAS as a predictor (95% CI = 0.69–0.73). The numbers greatly exceeded the chance performance obtained by always guessing the majority group (accuracy of 0.36). The performance of all super learner models was considered adequate for the context (Supplements pp.18–52). The variable importance plots for the super learners predicting treatment outcome and treatment allocation are shown in the Supplements (Figures S2–S7).

Counterfactual Treatment-Specific Means

None of the counterfactual treatment-specific means (had all patients received the same form of psychotherapy) differed from the observed stratum-specific average change in outcomes for CORE-10 and PHQ-9, meaning that allocating all patients to any one therapeutic framework would not have produced better treatment effects than observed in reality (Table II). However, when treatment outcome was defined as an improvement in therapist-reported social and occupational functioning, some differences did emerge. Measured as a SOFAS change, we observed better treatment outcomes in a scenario where all patients had received solution-focused therapy (TMLE est. = 0.099, CI = 0.030–0.167) and slightly worse outcomes if all patients had attended psychodynamic therapy (TMLE est. = −0.127, CI = −0.222– −0.033), compared to the observed distribution of treatments (0 = mean change in outcome across the whole sample). Overall, the observed treatment-specific means in different frameworks were close to the counterfactual ones.

Optimal Individualized Treatment Effects – Individually Best Treatments vs Random Treatment Allocation

The targeted estimate of the average treatment effect under the optimal individualized treatment was 0.29 for CORE-10 (untargeted estimate = 0.007, TMLE estimate = 0.294, 95% CI = 0.203–0.385), meaning that the treatment results, measured as a change in

Table II. Observed and Counterfactual Treatment Specific Means.

	CORE-10		PHQ-9		SOFAS	
Psychotherapy framework	Obs. Mean change	TMLE estimate (95% CI)	Obs. Mean change	TMLE estimate (95% CI)	Obs. Mean change	TMLE estimate (95% CI)
Solution-focused psychotherapy	0.060	0.033 (−0.041–0.107)	0.071	0.045 (−0.029–0.119)	0.101	0.099 (0.030–0.167)
Cognitive behavioral therapy	0.000	0.018 (−0.051–0.087)	−0.014	0.006 (−0.062–0.074)	−0.050	−0.046 (−0.109–0.016)
Integrative & CAT	−0.043	−0.048 (−0.131–0.036)	−0.035	−0.017 (−0.101–0.067)	−0.035	−0.029 (−0.100–0.042)
Psychodynamic psychotherapy	−0.068	−0.085 (−0.185–0.015)	−0.079	−0.102 (−0.208–0.004)	−0.121	−0.127 (−0.222– −0.033)

The counterfactual treatment-specific means (TMLE or targeted maximum likelihood estimate) refers to the mean change in the symptom scores if all patients had received the indicated treatment. Obs. Mean change is the observed change in the outcomes in standardized (z-score) units, among patients who had received a certain type of psychotherapy. Our analytical procedures required separate datasets, but samples for CORE-10 (the Clinical Outcomes In Routine Evaluation −10), PHQ-9 (the Patient Health Questionnaire), and SOFAS (the Social and Occupational Functioning Assessment Scale) showed here are essentially overlapping. n(CORE-10) = 2067, n(PHQ-9) = 2063, n(SOFAS) = 2124. CAT = Cognitive analytic therapy.

CORE-10 score, would be 0.29 standard deviations better if all patients would receive their optimal treatment, compared to the random assignment. Similarly, optimal treatment yielded estimates of better treatment outcomes compared to uniform treatment probability, as measured by changes in PHQ-9 (untargeted estimate = 0.002, TMLE estimate = 0.288, 95% CI = 0.197–0.379) and SOFAS (untargeted estimate 0.006, TMLE estimate = 0.279, 95% CI = 0.198–0.361). The numbers were highly similar when comparing the optimal treatment to patient-specific probable treatment (i.e., to treatments the sample received; see Supplementary materials, Table S4).

The numbers of patients attending different frameworks in reality, and when the allocation would have been done according to the optimal treatment rule, are shown in Table III. For all outcome measures, the optimal treatment rule would have allocated more people to the solution-focused therapy and fewer to the psychodynamic therapy, compared to the observed allocation. Yet, the optimal rule populated each treatment with some patients.

The individual treatment recommendations varied depending on which outcome measures (CORE-10, PHQ-9, SOFAS) were used in the prediction task. All our models predicting optimal treatment choice gave the same recommendations in 32.3% of cases, whereas in 10.6% of cases, all recommendations were different. Based on change alone (i.e., if the treatment recommendations were completely random), the treatment recommendations across the three models would have been the same in only 6.3% of cases. Inter-rater reliability between the optimal treatment option of different TMLE models represented a fare agreement (Cohen's κ for CORE-10 and PHQ-9 models was 0.325, for CORE-10 and SOFAS models κ = 0.213, for SOFAS and PHQ-9 models κ = 0.291; note, agreement is not required as achieving different outcomes may genuinely require different interventions). The congruence improved considerably if we constrained the

Table III. Number of Patients in Different Psychotherapeutic frameworks.

	CORE-10		PHQ-9		SOFAS	
Psychotherapy framework	Observed n (%)	Optimal n (%)	Observed n (%)	Optimal n (%)	Observed n (%)	Optimal n (%)
Solution-focused	581 (28.1%)	763 (36.9%)	578 (28.0%)	1059 (51.3%)	595 (28.0%)	1151 (54.2%)
Cognitive behavioral therapy	749 (36.2%)	709 (34.3%)	751 (36.4%)	423 (20.5%)	756 (35.6%)	454 (21.3%)
Integrative & CAT	479 (23.2%)	376 (18.2%)	477 (23.1%)	456 (22.1%)	502 (23.6%)	394 (18.5%)
Psychodynamic	258 (12.5%)	219 (10.6%)	257 (12.5%)	125 (6.1%)	271 (12.8%)	125 (5.9%)

Optimal treatment is estimated with TMLE (targeted maximum likelihood estimation). Our analysis required separate datasets, but samples for SOFAS (the Social and Occupational Functioning Assessment Scale), PHQ-9 (the Patient Health Questionnaire), and CORE-10 (the Clinical Outcomes In Routine Evaluation −10) shown here are essentially overlapping. n(CORE-10) = 2067, n(PHQ-9) = 2063, n(SOFAS) = 2124, n(all) = 2255.

sample to a 30% of the patient population who received most pronounced recommendation for any specific treatment framework (selecting based on the CORE-10 model led to $n = 553$ and to $\kappa = 0.644$ between CORE-10 and PHQ-9 models, $\kappa = 0.426$ between CORE-10 and SOFAS models, and $\kappa = 0.414$ between SOFAS and PHQ-9 models). Altogether 49% of these patients were allocated to the same treatment by the three allocation algorithms; that is, the models were more congruent across outcomes when excluding difficult-to-assign cases. The results were similar when the subsamples were based on other outcomes (see Supplementary materials, p. 11, Tables S5–S9).

Predictors – What Kind of Patient is Predicted to Benefit from Which Treatment?

To better understand our prediction models for optimal individualized treatments, we used baseline covariates to depict the "average patient" for whom each referred framework was optimal, although an average does not necessarily represent any real

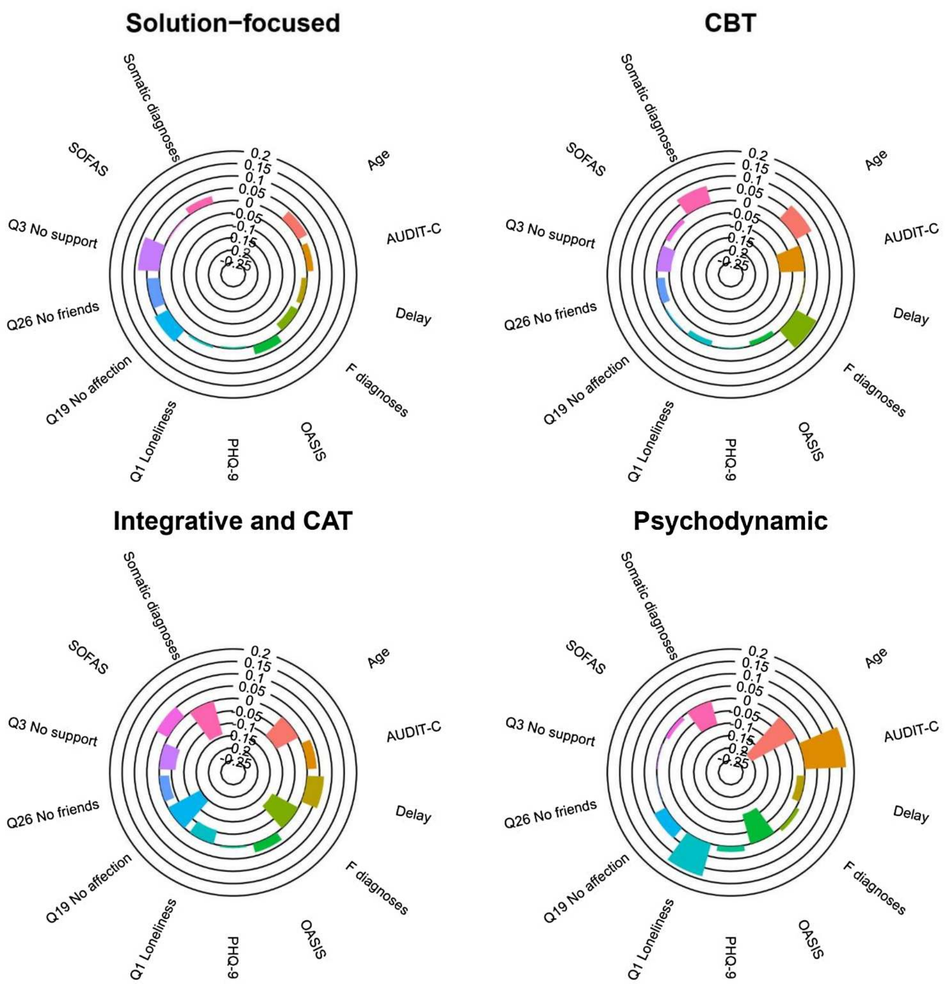

Figure 1. Patient profiles for the "average patient" allocated to each framework according to the optimal treatment rule estimated based on predicted improvement in CORE-10 score (the Clinical Outcomes In Routine Evaluation −10). All variables are standardized to mean = 0 and SD = 1. The zero line represents the average value across the whole sample. Note that the same-sized deviations might still have different meanings, depending on the variable's distribution and original scale. For example, a difference of one standard deviation in the number of previously diagnosed mental disorders (mean = 2.3, sd = 3.3) should be given much more significance compared to the same-size difference in AUDIT-C (mean = 1.8, sd = 1.4). An AUDIT-C score of 3.2 still falls below the threshold for risky alcohol use (5 for women, 6 for men), whereas a history of over five different mental disorders indicates a more severe etiology. The group differences of 0.2 and above are statistically significant ($p < 0.05$) even between the smallest groups ($n_{\text{psychodynamic}} = 219$, $n_{\text{integrative and CAT}} = 376$). Somatic diagnoses = the number of somatic diagnoses, SOFAS = Social and Occupational Functioning Assessment Scale, Q3 No support = CORE-OM item 3 assessing social support (higher values indicating less support, similarly to Q26, Q19 and Q1), Q26 No friends = CORE-OM item 26 assessing social relationships, Q19 No affection = CORE-OM item 19 assessing feelings of affection, Q1 Loneliness = CORE-OM item 1 assessing loneliness, PHQ-9 = Patient Health Questionnaire, OASIS = the Overall Anxiety Severity and Impairment Scale, F diagnoses = number of mental health diagnoses, Delay = treatment onset delay, AUDIT-C = The Alcohol Use Disorders Identification Test Consumption.

patient accurately (CORE-10 shown in Figure 1; see Supplementary Figures S8 and S9 for patient profiles stemming from PHQ-9 and SOFAS models). All variables were standardized to have a mean of 0 and a standard deviation of 1, and Figure 1 shows how the patient profiles deviate from the whole-sample means.

For the average patient allocated to solution-focused therapy based on the optimal individualized rule for CORE-10, most baseline variables were in a close range to the mean, with slightly higher levels of CORE-OM measures (Q3, Q26, and Q19) describing fewer feelings of affection and social support as exceptions. The mean patient allocated to psychodynamic therapy was younger, and reported higher levels of alcohol consumption, more loneliness, and fewer anxiety symptoms, compared to the mean scores across all patients, whereas patients referred to CBT experienced more comorbidity, reported less alcohol use, and were older compared to the average. These profiles were both outcome- and treatment-dependent, indicating that patients' baseline characteristics influence how each framework affects different measures of improvement.

Figure 2 presents the most significant predictors of optimal individualized treatment, as identified through SHAP analysis and illustrated using beeswarm plots. The five most important variables across the models included the psychotherapeutic framework, baseline measures of PHQ-9, OASIS, and CORE-10, number of previously prescribed somatic diagnoses, delay in treatment onset, age,

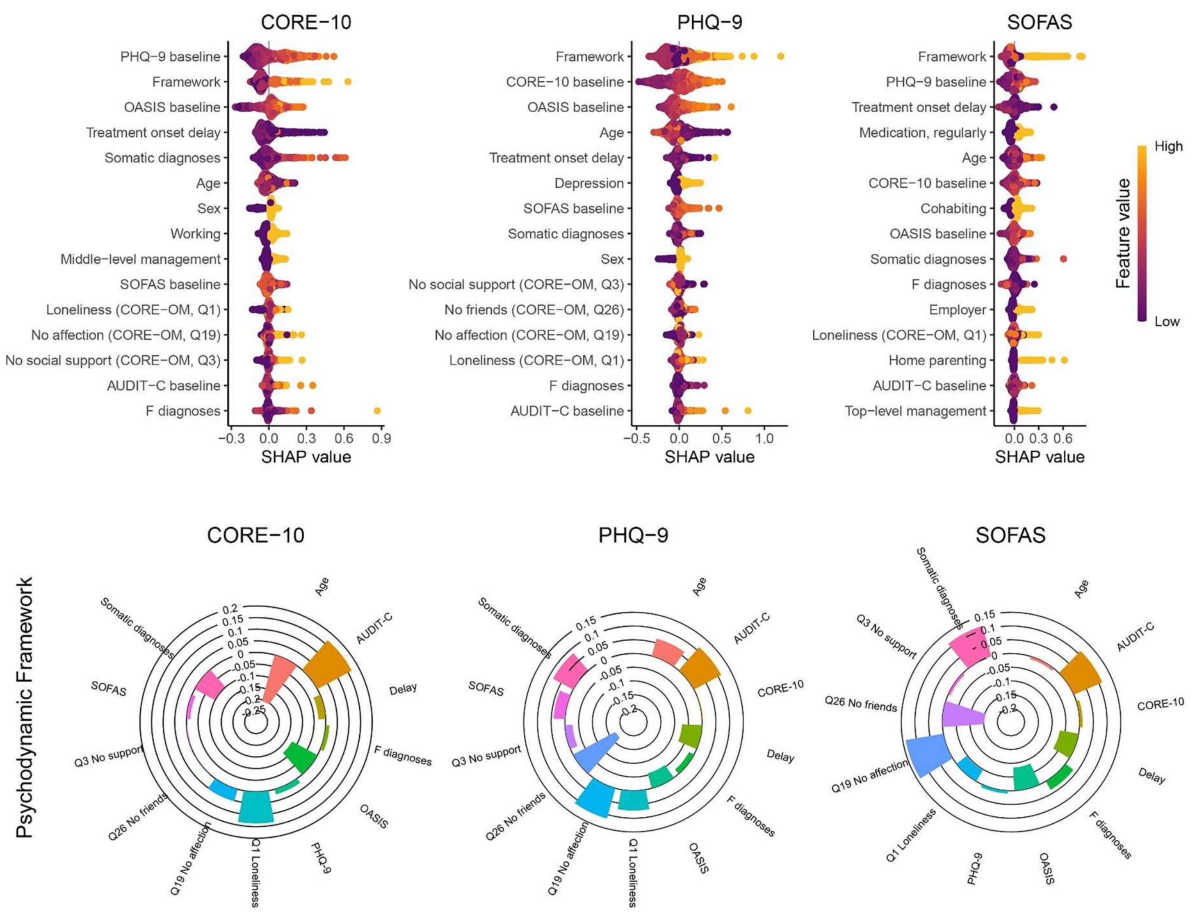

Figure 2. Beeswarm plots displaying SHAP (Shapley Additive Explanations) values for baseline predictors of optimal individualized treatment rules, and radar plots highlighting the baseline profiles of patients predicted to benefit most from psychodynamic psychotherapy. The optimized treatment rules were derived from predicted improvements in SOFAS (Social and Occupational Functioning Assessment Scale), CORE-10 (the Clinical Outcomes In Routine Evaluation −10), or PHQ-9 (Patient Health Questionnaire) scores. Across all models, the psychotherapeutic framework, baseline symptom measures, and treatment onset delay were identified as the most significant predictors. In particular, the psychodynamic framework (characterized by the highest value in the framework variable) held important information for the prediction models. Patients predicted to benefit most from the psychodynamic framework showed distinct baseline profiles across the three models. A consistent finding was that patients identified as most likely to benefit from psychodynamic therapy had higher baseline AUDIT-C scores (The Alcohol Use Disorders Identification Test Consumption) and lower baseline OASIS scores (the Overall Anxiety Severity and Impairment Scale) compared to those referred to other therapeutic frameworks.

and regular use of psychotropics. The framework was the most important variable in every model, with the highest SHAP values given for psychodynamic therapy. Therefore, we examined the average baseline profiles of patients allocated to psychodynamic treatment with radar plots. Across all models, patients for whom psychodynamic therapy was recommended as the optimal individualized treatment exhibited diverse baseline characteristics on average but reported fewer anxiety symptoms and higher alcohol consumption compared to those predicted to benefit most from other frameworks.

Discussion

We used targeted learning to predict the optimal treatment assignment of each patient to different psychotherapeutic frameworks. Consistent with our hypothesis (2), our results indicate that a machine learning-based treatment selection system could help to improve treatment efficacy, compared to the current practice. By employing targeted treatment selection, we witnessed significant enhancements in predicted treatment effects, translating to an increase of 0.28–0.29 standard deviations in improvements across CORE-10, PHQ-9, and SOFAS measures. Based on modeling studies, the effect of this size would be generally translated as clinically relevant (Cuijpers et al., 2014). To our knowledge, this is the first study using formal statistical theory to estimate the benefit of psychotherapy framework personalization in a standard metric.

Our findings suggest that implementing algorithm-based optimization of treatment targeting could benefit the treatment planning process for adult patients undergoing individual short psychotherapies. For example, compared to the observed effect size of CORE-10 (Cohen's $d = 0.66$), the treatment effect would increase by 0.29 standard deviation (to $d = 0.95$) if all patients had received their optimal treatment. This is a significant improvement since the average treatment effects for optimal treatment allocation are similar to the effect sizes observed in research settings that compare psychotherapy with other usual treatments (Cuijpers et al., 2021; Munder et al., 2018) when, in contrast, comparisons between different psychotherapy frameworks usually produce non-significant differences (Cuijpers, 2017). Efficient clinical implementation of this decision-aiding algorithm might not require much work, as the overall effect of optimal allocation was likely driven by a minority of patients predicted to benefit significantly from personalization rather than by slight improvements in most patients, as also previously observed (Schwartz et al., 2021).

The treatment recommendations varied across our models, as each model optimizes treatment choices based on different criteria (maximum improvement in depression, general distress, or psychosocial functioning). This finding is not entirely surprising, as, for instance, symptom reduction does not necessarily correspond to improved functioning (Sheehan et al., 2017). For example, returning to work—often seen as both an exposure and an indicator of improved functioning—may temporarily increase anxiety symptoms. Additionally, our outcome measures differ not only in content but also in their sources of information; the SOFAS reflects the therapist's perception of the patient's functional problems, the CORE-10 captures the patient's perception of symptoms and functioning, and the PHQ-9 measures patient-reported depressive symptoms, which not all patients necessarily suffered from initially.

Nevertheless, the fact that only 32.3% of treatment recommendations were consistent across all outcomes (49% in the top 30 percentiles in allocation-recommended strength) highlights the need to carefully consider the criteria used to label any treatment as clinically optimal, even if it is statistically optimal for an outcome. Thus, our results underscore that the optimal treatment for an individual patient is influenced by the choice of outcome measures and, consequently, by the treatment's intended goal. The clarity of treatment recommendations could be enhanced by using composite outcomes (for example, see Zainal et al., 2024) or broad-spectrum outcomes with multiple functional correlates (Pettersson et al., 2020). As a simple solution, clinicians could implement only the strongest and most consistent of multiple available algorithmic recommendations.

The baseline profiles of patients assigned to different psychotherapy frameworks determined by the individualized treatment rule varied depending on the treatment and the outcome measure used. Although, the most important variables for predicting optimal treatments were similar across all our models. According to the SHAP values, the psychotherapeutic framework, baseline symptom measures, and delay in treatment onset were the most informative variables for predictions of optimal treatment; variables that are easily collected along the routine treatment process.

The optimal rule likely involves complex interactions and nonlinearities that are difficult to interpret through individual variables, which may in part explain why machine learning approaches suggest greater benefits from personalization compared to traditional empirical research or clinical judgment

(Hannan et al., 2005; Lutz et al., 2022). Yet, our study demonstrates that, through advanced predictive modeling, it is possible to identify optimal treatments despite the absence of identifiable patient profiles or treatment moderators. This complements the view from meta-analyses of univariate associations that have found outcome predictive role only for older age and higher baseline symptoms (Cuijpers et al., 2022) and some evidence for social support and comorbidity (Tanguay-Sela et al., 2022).

Our findings on the probable nonlinearities of optimal treatment rules and their sensitivity to outcome measures underscore that care must be taken with context standardization in future efforts of implementing personalized treatment models. For instance,van Bronswijk et al. (2021) failed to validate personalization algorithms across two study samples that had some substantial differences, necessitating modifications to the predictive models before their application to validation data. Overall, the generalizability of clinical prediction algorithms has been worse than expected (Chekroud et al., 2024), and only a few attempts have been made to implement personalized treatment models into clinical practice (Delgadillo et al., 2020, 2021; Lutz et al., 2022a). Therefore, more efforts should be directed toward testing precision-treatment models in a clinical context. However, an ATE estimate may generalize even when underlying prediction models would not, as they pertain to different target parameters.

Against our first hypothesis, we did observe some minor differences in the counterfactual treatment effects between frameworks (average treatment effect if all patients received the same form of psychotherapy). If all patients had undergone solution-focused therapy, psychological functioning (SOFAS) would have improved more compared to the mean treatment effect observed across different frameworks. Conversely, if all patients had received psychodynamic therapy, the average change in psychological functioning would have been smaller than the observed mean change. However, for PHQ-9 or CORE-10, none of the counterfactual treatment-specific means differed from the observed ones, suggesting that the differences were therapist-perceived rather than an experience shared with patients. We interpret these results as being broadly consistent with the previous literature, indicating that, on average, there are no significant differences in the effectiveness of various psychotherapeutic frameworks within the most common diagnostic groups (e.g., Cuijpers et al., 2020; Stiles et al., 2008). Together, our findings demonstrate that different treatments may have varying effects on individuals, despite their comparative effects on average.

Our findings should be interpreted in the light of at least the following limitations. First, our results might be somewhat diminished if evaluated using an experimental design. For instance, a meta-analysis by Nye et al. (2023) found that the effect size for prospective treatment personalization was modest, both for personalization in general ($d = 0.22$) and for treatment matching ($d = 0.13$). However, the authors emphasize that even small effects can have a significant impact on the population level, and thus should not be dismissed as trivial.

Second, our study could have missed important predictors of treatment outcome, although this possibility was reduced by using a large set of predictors and efficient algorithms for selecting and weighing them. Third, while we cross-validated our machine learning models, we did not validate the ITR rule using independent training and test samples, as it pertains to a latent variable without an observed criterion, and as further sub-sampling would have violated the sample size requirements (Luedtke et al., 2019). Instead, large-sample theoretical confidence intervals of the implied ATE were computed as per TMLE theory (van der Laan & Rose, 2011).

Finally, clinicians are known to tailor their treatments according to patients' individual needs (Wester et al., 2025), and they often receive additional training in multiple psychotherapeutic frameworks. As a result, there is probably content overlap among the different psychotherapeutic treatments examined in this study. This kind of overlap may lead to more comparable treatment approaches, and therefore, smaller differences in average effects would be expected between bona fide therapies than between manualized ones. However, it is interesting that despite the likely presence of overlap in techniques and content across frameworks in naturalistic settings, we still observed variation in which treatments are most optimal for different individuals.

One of the strengths of our study is the large naturalistic sample and a comprehensive set of possible predictors. Furthermore, we were able to compare four different treatment options for each patient, whereas most previous studies on pre-treatment recommendations have only included pairwise comparisons (Bettega et al., 2024; Deisenhofer et al., 2018; Huibers et al., 2015; Schwartz et al., 2021; van Bronswijk et al., 2021). An algorithm offering diverse treatment options has more flexibility in selecting the most optimal treatment for each patient. In the development of a clinical decision support system, it could be useful to include many available treatments.

To conclude, we found evidence that clinically significant improvements in treatment outcomes can be achieved via targeting (personalization) of the psychotherapy framework based on individualized treatment rules. We consistently estimated a difference of ~0.3 standard units between personalized and non-personalized psychotherapy frameworks. Unlike the treatment effects size, the optimal individualized treatment rule depended on the outcome for which it was optimized and appeared non-linear, meaning that a clinical implementation requires careful consideration and testing. Here, we estimated likely benefits from such efforts and illustrated some flexible and statistically rigorous tools to help realize those benefits. Our results showed that each psychotherapy framework included in our model was optimal for some patients, underscoring the possible benefits of having multiple treatment options.

Author Contributions

VKM, THR, and SES participated in designing the study. VKM conducted the statistical analysis under consultation by THR, while all authors collaborated on interpreting the results. VKM wrote the original draft of the manuscript, and all authors provided critical revisions and accepted the final version for publication.

Funding

This study was supported by the Social Insurance Institution of Finland (Kela 140/331/2021, to SES), the Research Council of Finland (formerly Academy of Finland; grant numbers 334057, 335901, and 358138, to THR), and by the European Union – NextGenerationEU (SES).

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Ethical Approval

This research has received authorization from the HUS Ethical Review Board (HUS/1861/2020 & HUS/3150/2020) and Findata, the national social and health data permit authority (THL/4810/14.02.00/2020 & THL/1303/14.06.00/2023).

Supplemental Data

Supplemental data for this article can be accessed https://doi.org/10.1080/10503307.2025.2517567.

ORCID

Veera K. Malkki http://orcid.org/0000-0001-5138-8103
Suoma E. Saarni http://orcid.org/0000-0003-3555-9958
Wolfgang Lutz http://orcid.org/0000-0002-5141-3847
Tom H. Rosenstöm http://orcid.org/0000-0001-8277-3776

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

Aarnisalo, P., & Mattila, A. S. (2020). Ratkaisukeskeinen psykoterapia teoksessa M. In O. Huttunen, & H. Kalska (toim.), *Psykoterapiat* (ss. 178–192, 4. uudistettu painos). Kustannus Oy.

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

Barber, J. P., Muran, J. C., Mccarthy, K. S., Keefe, J. R., & Zilcha-Mano, S. (2021). Research on dynamic therapies. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (pp. 387–388). John Wiley & Sons.

Barkham, M., Bewick, B., Mullin, T., Gilbody, S., Connell, J., Cahill, J., Mellor-Clark, J., Richards, D., Unsworth, G., & Evans, C. (2013). The CORE-10: A short measure of psychological distress for routine use in the psychological therapies. *Counselling and Psychotherapy Research*, *13*(1), 3–13. https://doi.org/10.1080/14733145.2012.729069

Barkham, M., Lutz, W., & Castonguay, L. G. (2021). *Bergin and garfield's handbook of psychotherapy and behavior change*. John Wiley & Sons.

Benjet, C., Zainal, N. H., Albor, Y., Alvis-Barranco, L., Carrasco-Tapias, N., Contreras-Ibáñez, C. C., Cudris-Torres, L., de la Peña, F. R., González, N., Guerrero-López, J. B., & Gutierrez-Garcia, R. A. (2023). A precision treatment model for internet-delivered cognitive behavioral therapy for anxiety and depression Among university students: A secondary analysis of a randomized clinical trial. *JAMA Psychiatry*, *80*(8), 768–777. https://doi.org/10.1001/jamapsychiatry.2023.1675

Bettega, F., Mendelson, M., Leyrat, C., & Bailly, S. (2024). Use and reporting of inverse-probability-of-treatment weighting for multicategory treatments in medical research: A systematic review. *Journal of Clinical Epidemiology*, *170*), https://doi.org/10.1016/j.jclinepi.2024.111338

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research*, *25*(1), 6–19. https://doi.org/10.1080/10503307.2013.817696

Bush, K., Kivlahan, D. R., McDonell, M. B., Fihn, S. D., Bradley, K. A., & for the Ambulatory Care Quality Improvement Project (ACQUIP), (1998). The AUDIT alcohol consumption questions (AUDIT-C): An effective brief screening test for problem drinking. *Archives of Internal Medicine*, *158*(16), 1789–1795. https://doi.org/10.1001/archinte.158.16.1789

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

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

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology*, *14*(1), 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Coyle, J. R., Hejazi, N. S., Malenica, I., Phillips, R. V., Mertens, A., Benjamin-Chung, J., Cai, W., Dayal, S., Jr, C., & Hubbard, A. E. (2020). *Targeting learning: Robust statistics for reproducible research*.

Cuijpers, P. (2017). Four decades of outcome research on psychotherapies for adult depression: An overview of a series of meta-analyses. *Canadian Psychology*, *58*(1), 7–19. https://doi.org/10.1037/cap0000096

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), Article 1. https://doi.org/10.3390/jpm12010093

Cuijpers, P., Karyotaki, E., de Wit, L., & Ebert, D. D. (2020). The effects of fifteen evidence-supported therapies for adult depression: A meta-analytic review. *Psychotherapy Research*, *30*(3), 279–293. https://doi.org/10.1080/10503307.2019.1649732

Cuijpers, P., Quero, S., Noma, H., Ciharova, M., Miguel, C., Karyotaki, E., Cipriani, A., Cristea, I. A., & Furukawa, T. A. (2021). Psychotherapies for depression: A network meta-analysis covering efficacy, acceptability and long-term outcomes of all main treatment types. *World Psychiatry*, *20*(2), 283–293. https://doi.org/10.1002/wps.20860

Cuijpers, P., Turner, E. H., Koole, S. L., van Dijke, A., & Smit, F. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety*, *31*(5), 374–378. https://doi.org/10.1002/da.22249

Dahabreh, I. J., & Bibbins-Domingo, K. (2024). Causal inference about the effects of interventions from observational studies in medical journals. *JAMA*, *331*(21), 1845–1853. https://doi.org/10.1001/jama.2024.7741

Deisenhofer, A.-K., Delgadillo, J., Rubel, J. A., Böhnke, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and Anxiety*, *35*(6), 541–550. https://doi.org/10.1002/da.22755

Delgadillo, J., Ali, S., Fleck, K., Agnew, C., Southgate, A., Parkhouse, L., Cohen, Z. D., DeRubeis, R. J., & Barkham, M. (2021). Stratified care vs stepped care for depression: A cluster randomized clinical trial. *JAMA Psychiatry*, https://doi.org/10.1001/jamapsychiatry.2021.3539

Delgadillo, J., Appleby, S., Booth, S., Burnett, G., Carey, A., Edmeade, L., Green, S., Griffin, P., Johnson, E., Jones, R., Parker, P., Reeves-McLaren, L., & Lutz, W. (2020). The Leeds risk index: Field-test of a stratified psychological treatment selection algorithm. *Psychotherapy and Psychosomatics*, *89*(3), 189–190. https://doi.org/10.1159/000505193

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

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The personalized advantage index. *Translating Research on Prediction Into Individualized Treatment Recommendations. A Demonstration. PLoS One*, *9*(1), 1–8. https://doi.org/10.1371/journal.pone.0083875

Goldman, H. H., Skodol, A. E., & Lave, T. R. (1992). *Revising Axis V for DSM-IV: A review of measures of social functioning* (SSRN Scholarly Paper ID 2143992). Social Science Research Network. https://papers.ssrn.com/abstract=2143992.

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*, 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Gruber, S., & van der Laan, M. J. (2009). Targeted maximum likelihood estimation: A gentle introduction.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, *61*(2), 155–163. https://doi.org/10.1002/jclp.20108

Hilsenroth, M. J., Ackerman, S. J., Blagys, M. D., Baumann, B. D., Baity, M. R., Smith, S. R., Price, J. L., Smith, C. L., Heindselman, T. L., Mount, M. K., & Holdwick, D. J. (2000). Reliability and validity of DSM-IV axis V. *American Journal of Psychiatry*, *157*(11), 1858–1863. https://doi.org/10.1176/appi.ajp.157.11.1858

Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, *76*(3), 292–304. https://doi.org/10.1080/00031305.2021.2021984

Huibers, M. J. H., Cohen, Z. D., Lemmens, L. H. J. M., Arntz, A., Peeters, F. P. M. L., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage index approach. *PLoS One*, *10*(11), e0140771. https://doi.org/10.1371/journal.pone.0140771

Imbens, G., & Rubin, D. (2015). Causality: The basic framework. In G. W. Imbens & D. B. Rubin (Eds.), *Causal inference for statistics, social, and biomedical sciences: An introduction* (pp. 3–21). Cambridge University Press.

Jovanović, A. A., GaSic, M. J., Ivković, M., Milovanović, S., & Damjanović, A. (2008). Reliability and validity of DSM-IV axis V scales in a clinical sample of veterans with posttraumatic stress disorder. *Psychiatria Danubina*, *20*, 286–300.

Kalska, H., & Särkelä, L. (2020). Kognitiivis-analyyttinen psykoterapia Teoksessa. In M. O. Huttunen & H. Kalska (toim.), *Psykoterapiat* (ss. 112–127, 4. uudistettu painos). Kustannus Oy Duodecim.

Kazdin, A. E. (2022). Drawing causal inferences from randomized controlled trials in psychotherapy research. *Psychotherapy Research*, *33*(8), 991–1003. https://doi.org/10.1080/10503307.2022.2130112

Kessler, R. C., Furukawa, T. A., Kato, T., Luedtke, A., Petukhova, M., Sadikova, E., & Sampson, N. A. (2022). An individualized treatment rule to optimize probability of remission by continuation, switching, or combining antidepressant medications after failing a first-line antidepressant in a two-stage randomized trial. *Psychological Medicine*, *52*(15), 3371–3380. https://doi.org/10.1017/S0033291721000027

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

Kuusinen, K.-L., Valkonen, H., & Wahlström, J. (2020). Integratiivinen psykoterapia. In M. O. Huttunen, & H. Kalska (toim.), *Psykoterapiat* (ss. 193–209, 4th edition). Kustannus Oy.

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy*, *55*(4), 520–537. https://doi.org/10.1037/pst0000167

Lorenzo-Luaces, L., Peipert, A., De Jesús Romero, R., Rutter, L. A., & Rodriguez-Quintana, N. (2021). Personalized medicine and cognitive behavioral therapies for depression: Small effects, Big problems, and bigger data. *International Journal of Cognitive Therapy*, *14*(1), 59–85. https://doi.org/10.1007/s41811-020-00094-3

Luedtke, A., Sadikova, E., & Kessler, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clinical Psychological Science*, *7*(3), 445–461. https://doi.org/10.1177/2167702618815466

Lundberg, S., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions* (arXiv:1705.07874). arXiv. https://doi.org/10.48550/arXiv.1705.07874.

Luque-Fernandez, M. A., Schomaker, M., Rachet, B., & Schnitzer, M. E. (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, *37*(16), https://doi.org/10.1002/sim.7628

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

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W. B., Evans, C., Noble, R., & Iveson, S. (2005). Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. *Journal of Consulting and Clinical Psychology*, *73*(5), 904–913. https://doi.org/10.1037/0022-006X.73.5.904

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., & Delgadillo, J. (2022b). Measurement-Based and data-informed psychological therapy. *Annual Review of Clinical Psychology*, *18*(1), 71–98. https://doi.org/10.1146/annurev-clinpsy-071720-014821

Munder, T., Flückiger, C., Leichsenring, F., Abbass, A. A., Hilsenroth, M. J., Luyten, P., Rabung, S., Steinert, C., & Wampold, B. E. (2018). Is psychotherapy effective? A re-analysis of treatments for depression. *Epidemiology and Psychiatric Sciences*, *28*(3), 268–274. https://doi.org/10.1017/S2045796018000355

Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: An introduction to super learning. *European Journal of Epidemiology*, *33*(5), 459–464. https://doi.org/10.1007/s10654-018-0390-z

Newman, M. G., Agras, W. S., Haaga, D. A. F., & Jarrett, R. B. (2021). Cognitive, behavioral, and cognitive behavioral therapy. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and garfield's handbook of psychotherapy and behavior change* (pp. 469–470). John Wiley & Sons.

Norman, S. B., Hami Cissell, S., Means-Christensen, A. J., & Stein, M. B. (2006). Development and validation of an overall anxiety severity And impairment scale (OASIS). *Depression and Anxiety*, *23*(4), 245–249. https://doi.org/10.1002/da.20182

Nye, A., Delgadillo, J., & Barkham, M. (2023). Efficacy of personalized psychological interventions: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *91*(7), 389–397. https://doi.org/10.1037/ccp0000820

Pettersson, E., Larsson, H., D'Onofrio, B. M., Bölte, S., & Lichtenstein, P. (2020). The general factor of psychopathology: A comparison with the general factor of intelligence with respect to magnitude and predictive validity. *World Psychiatry*, *19*(2), 206–213. https://doi.org/10.1002/wps.20763

Phillips, R. V., van der Laan, M. J., Lee, H., & Gruber, S. (2023). Practical considerations for specifying a super learner. *International Journal of Epidemiology*, *52*(4), 1276–1285. https://doi.org/10.1093/ije/dyad023

Polley, E., & van der Laan, M. (2010). Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*. https://biostats.bepress.com/ucbbiostat/paper266.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Roehrle, B., & Strouse, J. (2008). Influence of social support on success of therapeutic interventions: A meta-analytic review. *Psychotherapy: Theory, Research, Practice, Training*, *45*(4), 464–476. https://doi.org/10.1037/a0014333

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.2307/2335942

Rosenström, T. H., Mylläri, S., Malkki, V., & Saarni, S. E. (2022). Feasibility of generic, short, and easy-to-use assessment of psychological distress during psychotherapy: Longitudinal measurement invariance of CORE-10 and -OM. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, *32*(8), 1090–1099. https://doi.org/10.1080/10503307.2022.2074807

Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry*, *6*(3), 412–415. https://doi.org/10.1111/j.1939-0025.1936.tb05248.x

Saarni, S. E., Rosenström, T., Stenberg, J.-H., Plattonen, A., Holi, M., Ekelund, J., Granö, N., Komsi, N., & Saarni, S. I. (2022). Finnish psychotherapy quality register: Rationale, development, and baseline results. *Nordic Journal of Psychiatry*, *0*(0), 1–12. https://doi.org/10.1080/08039488.2022.2150788

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E. W., Stahl, D., & Fusar-Poli, P. (2021). Implementing precision psychiatry: A systematic review of individualized prediction models for clinical practice. *Schizophrenia Bulletin*, *47*(2), 284–297. https://doi.org/10.1093/schbul/sbaa120

Schuler, M. S., & Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, *185*(1), 65–73. https://doi.org/10.1093/aje/kww165

Schwartz, B., Cohen, Z. D., Rubel, J. A., Zimmermann, D., Wittmann, W. W., & Lutz, W. (2021). 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

Sheehan, D. V., Nakagome, K., Asami, Y., Pappadopulos, E. A., & Boucher, M. (2017). Restoring function in major depressive disorder: A systematic review. *Journal of Affective Disorders*, *215*, 299–313. https://doi.org/10.1016/j.jad.2017.02.029

Smith, G. N., Ehmann, T. S., Flynn, S. W., MacEwan, G. W., Tee, K., Kopala, L. C., Thornton, A. E., Schenk, C. H., & Honer, W. G. (2011). The assessment of symptom severity and functional impairment With DSM-IV axis V. *Psychiatric Services*, *62*(4), 411–417. https://doi.org/10.1176/ps.62.4.pss6204_0411

Smith, M. J., Phillips, R. V., Luque-Fernandez, M. A., & Maringe, C. (2023). *Application of targeted maximum likelihood estimation in public health and epidemiological studies: A systematic review* (arXiv:2303.07329). arXiv. http://arxiv.org/abs/2303.07329.

Stiles, W. B., Barkham, M., Mellor-Clark, J., & Connell, J. (2008). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary-care routine practice: Replication in a larger sample. *Psychological Medicine*, *38*(5), 677–688. https://doi.org/10.1017/S0033291707001511

Stochl, J., Fried, E. I., Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., Jones, P. B., & Perez, J. (2020). On dimensionality, measurement invariance, and suitability of Sum scores for the PHQ-9 and the GAD-7. *Assessment*, *1073191120976863*), https://doi.org/10.1177/1073191120976863

Tanguay-Sela, M., Rollins, C., Perez, T., Qiang, V., Golden, G., Tunteng, J.-F., Perlman, K., Simard, J., Benrimoh, D., & Margolese, H. C. (2022). A systematic meta-review of patient-level predictors of psychological therapy outcome in major depressive disorder. *Journal of Affective Disorders*, *317*, 307–318. https://doi.org/10.1016/j.jad.2022.08.041

van Bronswijk, S. C., Bruijniks, S. J. E., Lorenzo-Luaces, L., Derubeis, R. J., Lemmens, L. H. J. M., Peeters, F. P. M. L., & Huibers, M. J. H. (2021a). Cross-trial prediction in psychotherapy: External validation of the personalized advantage index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression. *Psychotherapy Research*, *31*(1), 78–91. https://doi.org/10.1080/10503307.2020.1823029

van Bronswijk, S. C. v., DeRubeis, R. J., Lemmens, L. H. J. M., Peeters, F. P. M. L., Keefe, J. R., Cohen, Z. D., & Huibers, M. J. H. (2021b). Precision medicine for long-term depression outcomes using the personalized advantage index approach: Cognitive therapy or interpersonal psychotherapy? *Psychological Medicine*, *51*(2), 279–289. https://doi.org/10.1017/S0033291719003192

van der Laan, M., Coyle, J., Hejazi, N., Malenica, I., Phillips, R., & Hubbard, A. (2023). Targeted learning in R: Causal data science with the tlverse software ecosystem. https://tlverse.org/tlverse-handbook/.

van der Laan, M., & Luedtke, A. R. (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of Causal Inference*, *3*(1), 61–95. https://doi.org/10.1515/jci-2013-0022

van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data | SpringerLink*. https://link.springer.com/book/10.1007978-1-4419-9782-1.

van der Laan, M. J., & Starmans, R. J. C. M. (2014). Entering the Era of data science: Targeted learning and the integration of statistics and computational data analysis. *Advances in Statistics*, *2014*, 1–19. https://doi.org/10.1155/2014/502678

Wester, R. A., Steinbrenner, T., Hanraths, S., Ehrenthal, J. C., Iovoli, F., Fritz, J., & Rubel, J. (2025). Practices of personalized treatment selection among German psychotherapists: A preregistered mixed methods study. *Psychotherapy Research*, *0(0)*, 1–14. https://doi.org/10.1080/10503307.2025.2473931

Zainal, N. H., Bossarte, R. M., Gildea, S. M., Hwang, I., Kennedy, C. J., Liu, H., Luedtke, A., Marx, B. P., Petukhova, M. V., Post, E. P., Ross, E. L., Sampson, N. A., Sverdrup, E., Turner, B., Wager, S., & Kessler, R. C. (2024). Developing an individualized treatment rule for veterans with major depressive disorder using electronic health records. *Molecular Psychiatry*, 1–11. https://doi.org/10.1038/s41380-024-02500-0

Zimmermann, D., Wampold, B. E., Rubel, J. A., Schwartz, B., Poster, K., Schilling, V. N. L. S., Deisenhofer, A.-K., Hehlmann, M. I., Gómez Penedo, J. M., & Lutz, W. (2021). The influence of extra-therapeutic social support on the association between therapeutic bond and treatment outcome. *Psychotherapy Research*, *31*(6), 726–736. https://doi.org/10.1080/10503307.2020.1847344