UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Efficacy and effectiveness of therapist-guided internet versus face-to-face cognitive behavioural therapy for depression via counterfactual inference using naturalistic registers and machine learning in Finland: a retrospective cohort study

Rosenström, Tom H.; Saarni, Suoma E.; Saarni, Samuli I.; Tammilehto, Jaakko; Stenberg, Jan-Henry

2025-03
Elsevier Ltd.
http://hdl.handle.net/10138/593872
Rosenström, T H, Saarni, S E, Saarni, S I, Tammilehto, J & Stenberg, J-H 2025, 'Efficacy and effectiveness of therapist-guided internet versus face-to-face cognitive behavioural therapy for depression via counterfactual inference using naturalistic registers and machine learning in Finland: a retrospective cohort study', The Lancet psychiatry, vol. 12, no. 3, pp. 189-197. https://doi.org/10.1016/S2215-0366(24)00404-8

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

Authors' Accepted Manuscript version of Rosenström et al., 2025, Lancet Psychiatry, https://doi.org/10.1016/S2215-0366(24)00404-8

(NOTE: The text in this version has differences to the final published version)

Efficacy and effectiveness of therapist-guided internet *versus* face-to-face cognitive behavioural therapy for depression via counterfactual inference using naturalistic registers and machine learning in Finland: a retrospective cohort study

Tom H. Rosenström (PhD)[1,*], Suoma E. Saarni (MD, PhD, Prof.)[2,3,4], Samuli I. Saarni (MD, PhD, Prof.)[3], Jaakko Tammilehto (PhD)[1], Jan-Henry Stenberg (PhD)[2]

1 Department of Psychology, Faculty of Medicine, University of Helsinki, Helsinki, Finland (Tom Rosenström, PhD, Docent; Jaakko Tammilehto, PhD)

2 Department of Psychiatry, Helsinki University Hospital, Helsinki, Finland (Prof Suoma E. Saarni; Jan-Henry Stenberg, PhD)

3 Clinical Medicine, Tampere University, Tampere, Finland (Prof Suoma E. Saarni; Prof Samuli I. Saarni)

4 Family and Social Services, Psychiatry, Wellbeing Services County of Päijät-Häme, Finland

* Correspondence to Dr. Tom H. Rosenström, Department of Psychology, University of Helsinki, PL 21 (Haartmaninkatu 3), 00014, Helsinki, Finland: tom.rosenstrom@helsinki.fi

Tom H. Rosenström, Ph.D., https://orcid.org/0000-0001-8277-3776

Suoma E. Saarni, M.D., Ph.D., https://orcid.org/0000-0003-3555-9958

Samuli I. Saarni, M.D., Ph.D., https://orcid.org/0000-0002-9348-5435

Jaakko Tammilehto, Ph.D., https://orcid.org/0000-0002-6920-4322

Jan-Henry Stenberg, Ph.D., https://orcid.org/0000-0003-1327-7757

Summary

Background

According to meta-analyses of randomized controlled trials (RCTs), therapist-guided internet-delivered cognitive behavioural therapy (iCBT) is as effective treatment for depression as traditional face-to-face CBT (fCBT), despite its substantially lower costs. RCTs are not always representative of routine practice, however, which may inflate effectiveness estimates. We combined rich data with counterfactual causal statistical reasoning to provide an fCBT-iCBT comparison complementary to RCTs.

Methods

We linked full historical guided-iCBT and fCBT cohort registries with multiple Finnish social and healthcare registries. The iCBT programme with third-wave principles and the fCBTs were provided by HUS Helsinki University Hospital to depression patients without acute suicide or substance-abuse risk; fCBT was delivered in the Uusimaa region, whereas iCBT was nationwide, excluding patients with treatment-interfering psychotic, neurological or personality disorder, chronic or bipolar depression, or age under 15. Primary outcome was causal average treatment effect (ATE) differential between fCBT and iCBT, with effect measured by change in Patient Health Questionnaire -9 (PHQ-9) over treatment or to the last recorded report (if same as the first report, no change was recorded, thereby penalizing dropout). For its optimal and robust (*a posteriori* balanced) estimate, we applied targeted maximum likelihood machine learning.

Findings

The guided-iCBT registry spanned from 2018-12-12 to 2022-12-22 and fCBT from 2018-08-28 to 2022-09-28. From the total of 32,343 registered therapies, 392 fCBT and 5,467 iCBT depression treatments were extracted. Main sample contained 1733 men (30%) and 4101 women (70%) aged 35 on average (s.d. 12 years, range 15–80^{+}). ATE estimate indicated that the iCBT program was at least as effective and efficacious as fCBT (average PHQ-9 score declined 0.745 points more in iCBT; 95% confidence interval 0.156, 1.334). Sensitivity analyses concurred.

Interpretation

Considered alongside prior RCTs that strengthen causal evidence, these findings underscore iCBT as a cost-effective first-line treatment option.

Keywords: psychotherapy; delivery format; causality; depressive disorder; targeted maximum likelihood; inverse-probability weighting; super learner; digital intervention; psychosocial treatments

Funding:

This study was supported by the Research Council of Finland (a.k.a., Academy of Finland) with grant numbers 334057, 335901, and 358138 to THR and the project Machine-learned causal factors for psychotherapy.

Research in context (panel)

Evidence before this study.

Multiple meta-analyses (see references for 3 recent ones) indicate that cognitive behavioural therapy (CBT) is equally efficacious in treating depressive disorders when implemented in guided internet-delivered format (iCBT) compared to when delivered in the usual face-to-face format (fCBT). Because iCBT programs use automated videos and exercises, the online therapists can spend much less time per patient compared to fCBT therapists, which can lead to significant costs savings. However, the meta-analyses rely on randomized controlled trials (RCTs), and the possibility remains that the trials fail to represent the full routine-treatment population or to adequately take in account the large differences in treatment adherence or other differences between fCBT and guided iCBT. Using complementary sources of causal information have recently been forcefully endorsed for all of medicine. One of the state-of-art methods for such causal triangulation is the targeted maximum likelihood estimation with machine learning. On September 16, 2024, our PubMed (MEDLINE) search gave zero hits for the following query: "targeted maximum likelihood" AND depression AND (internet OR digital) AND ("cognitive behavioural therapy" OR "cognitive behavioural therapy"). Overall, causal triangulation of the fCBT-iCBT efficacy differential in depression treatment is important given the substantial societal and monetary costs associated with depressive disorders and their treatment, which jointly imply need to comprehensively understand the health effects of seemingly low-cost treatment policies [the above query gave zero pertinent hits also when replacing "targeted maximum likelihood" with “causal triangulation” or with “causality”, or with “quasi-experiment”].

Added value of this study

To our knowledge, this is the first study to quantitatively evaluate the efficacy differential between fCBT and guided iCBT that is *not* an RCT or a meta-analysis of such trials. This adds value because RCTs may be prone to shared sources of bias, meaning that complementary sources of information are needed. Our complementary source was [counterfactual] causal inference using targeted maximum likelihood (machine)

learning and a unique linked registry from five distinct registry owners, combined using Finnish laws and regulations on secondary use of health and social registries. This recently developed method can integrate features from both parametric and non-parametric methods, with the aim of providing precise and reliable estimates of causal effects in situations where the data is complex and may contain multiple confounding factors. We observed that the results from the RCTs held true both in routinely collected data and in *a posteriori* balanced (i.e., causally informative) data – the guided iCBT program was as effective and efficacious as fCBT in terms of depression symptom decline from pre- to post-treatment (or to the last available measurement).

Implications of all the available evidence

Previous RCTs have suggested that guided internet delivery of CBT is equally effective and more cost-effective in treating depression compared to face-to-face delivery. In this study, the result was replicated in naturalistic register data and with counterfactual causal inference. Consequently, from multiple convergent lines of empirical evidence, internet-delivery of short first-line treatments emerges as a sensible default-treatment option and an apt response to the high demand. Future research may consider possible further gains from personalization of the delivery method.

Introduction

Meta-analyses of available randomized controlled trials (RCTs) indicate that therapist-guided internet-delivered cognitive behavioural therapy (iCBT) appears to be similarly effective treatment compared to the more costly and traditional face-to-face CBT (fCBT).[1,2] Much of the cost differential between guided iCBT and fCBT derives from therapists' differential time investment per patient. However, a recent meta-analysis found no association between contact time and treatment effect in the psychotherapy for depression.[3] When patients are directed to digital therapy supported by a therapist, they may accept it as readily as any recommended treatment.[4] They may benefit regardless of their symptom severity at intake.[5,6] Some clinicians, however, may find it difficult to accept that partially automated treatments could yield as good treatment outcomes as face-to-face therapies. The risk here is that clinicians may inadvertently impart their attitudes to patients, potentially leading to diminished treatment efficacy.[4]

RCTs are not always representative of routine practice, which may inflate effectiveness estimates.[7,8] For example, patients objecting to digital treatments may not enrol in an RCT offering iCBT. Furthermore, samples in RCTs tend to be small. Although meta-analyses combine RCTs to achieve larger samples, they often lack commensurate and comprehensive data on complexity of patients' condition and other background factors that could assess representativeness. Thus, it remains unclear whether guided iCBT truly is as effective as fCBT in real-world settings. The central counterfactual question is: Would the Average Treatment Effect (ATE) still be equal if we could allocate the exact replicas of the same patients simultaneously to both guided iCBT and fCBT and compare their efficacies?

Although considered a gold standard, RCTs are neither the only nor a perfect method for balancing intervention-group composition to avoid confounding. Researchers in medicine, psychology, and statistics forcefully argue that reliable causal inference often requires multiple sources of evidence rather than RCTs exclusively.[9–11] Figure 1 summarizes the assumptions many such methods make. For example, estimates weighted by (known) inverse probabilities of treatment are free from confounding.[12,13]

Alternatively, if the outcome can be accurately modelled as a function of confounders and treatment allocation, an unbiased estimate for treatment effect is achievable via "G-computation".[14] If models for *both* outcome and treatment-assignment probability exist, they can be combined to achieve an augmented inverse-probability of treatment weighted (AIPTW) estimate that is "doubly robust". This means that a consistent estimate of the ATE is obtained if at least one of the models is correctly specified.[14] Finally, ensemble of machine learners (i.e., super learners) maximize the probability of at least one model being correct.[13,14]

A recently developed "targeted maximum likelihood estimator" (TMLE) based on super learners enjoys the asymptotic double-robustness property but is additionally efficient and well-defined and well-behaved in finite samples.[13,14] TMLE integrates features from modern statistics, machine learning, and causal inference. TMLE can optimize bias-variance trade-off for target parameter (e.g., ATE) instead of the treatment and outcome prediction models.[13,14] Primary motivation to use TMLE is its ability to combine multiple desirable properties, perhaps representing the state of the art in the field of pertinent counterfactual causal inference. The recent availability of such procedures enhances the value of administrative, population-based registries, which often contain abundant etiological information.[15]

Recent developments in Nordic countries enable linking psychotherapy data with background information in national registries.[16–18] In this study, we use TMLE on these rich information sources to estimate fCBT-iCBT efficacy difference in a causally informative design that is not an RCT, thus circumventing typical challenges of RCTs while providing triangulating evidence on the same causal question. Our naturalistic data informs about effectiveness, and their counterfactual analysis informs about both effectiveness and efficacy when all real patients would receive fCBT *vs*. guided iCBT. The new gold standard for evidence will then consist of combined inferences from RCT meta-analyses and TMLE applied to naturalistic datasets. Our investigation serves as a pragmatic proof of concept for using TMLE and related methods to address the question of ATE between fCBT and guided iCBT.

Methods

Study design

The study participants of this partly Finland-wide inverse-probability of treatment weighted cohort study were derived from naturalistic hospital treatment registries, from HUS Helsinki University Hospital iCBT (HUS-iCBT) registry and from the Finnish Psychotherapy Quality Register (FPQR).[16] The HUS-iCBT for depression largely corresponds to an iCBT we have previously investigated,[18] but represents more modern iteration of the online platform. The iCBT registry spanned from 2018-12-12 to 2022-12-22 and FPQR from 2018-08-28 to 2022-09-28. HUS was the only public-sector iCBT provider, providing iCBTs nationwide, but fCBTs mainly for Uusimaa, the Southwest region of Finland. For maximal comparability across different delivery formats for CBT for depression, we sampled the patients from these registries according to the flow chart of Figure 2A.

Each patient from both the registries was linked with other national registries using their social-security identification number. The national Findata Data Permit Authority and the Digital and Population Services Agency linked the patients with their relatives (parents and siblings), after which both the patients and their relatives were linked with data from other registry owners as summarised in Figure 2B. The registries contain accurate event dates for public-sector diagnoses and pharmaceutic purchases (F and ATC-N codes), but not (yet) private-sector diagnoses. Furthermore, 4 out of 40,263 patients we accessed had objected use of their registry data to Findata (present inclusion criteria unknown). Our research permission was granted by Findata according to national rules and regulations on secondary use of social and health data (permissions THL/4810/14.02.00/2020 and THL/1303/14.06.00/2023 to THR). These relinquish the requirement for informed consent. This study was conducted in accordance with the Sixth World Medical Association Declaration of Helsinki and evaluated by The Ethics Committee of HUS (HUS/1861/2020 and HUS/3150/2020). The study was not pre-registered elsewhere.

Patients and procedures

The fCBT patients were from FPQR, undergoing primary- and specialty-care CBTs outsourced by public healthcare: We selected treatments with patients whose primary diagnosis for the treatment was depression (ICD-10 codes F32, F33, or F34.1) and therapist was trained in CBT framework. All FPQR therapists have a 60-credit, 3-year, training in their framework alongside their clinical work. The fCBT treatments typically contained up to 20 sessions, but in rare cases, could maximally contain 40 sessions per year, up to three years.[16] A session was 45 minutes, and a conservative estimate of worktime for each session would be a total 60 minutes, including preparation and documentation. This and the observed mean number of 14 sessions in short FPQR treatments[19] implies an 840-min average work-time expenditure per patient. Weekly sessions were recommended.

Patients entered iCBT with a doctor's prescription, inclusion criteria being mild to moderate depression (F32.0 or F32.1), age over 15 years, and Finnish or Swedish proficiency with motivation and technical means to work online independently. Exclusion criteria for both fCBT and iCBT were acute suicide risk and current substance use. Further exclusion criteria for iCBT were treatment-interfering psychotic, neurological, or personality disorder, or chronic or bipolar depression. The suitability for treatment was assessed by the attending physician. All the iCBTs were performed using the same online program that aligns with CBT principles (including third-wave techniques like mindfulness): the patient proceeds on their own pace but asynchronously chatting with a trained therapist (a psychologist or a nurse), who also provides unprompted guidance and an occasional phone call (1 per patient ideally). The recommended time spent by a therapist per patient is 11 min/week, or 77 min in total for the seven weekly sessions. The pace of one session per week was a recommendation to the patients and they could proceed slower or faster, up to 1 session per day. However, not progressing in two weeks led to a call from the therapist and, ultimately, cessation of treatment if progress did not ensue despite encouragement.

In both treatment arms, we included treatments that had started 3 months before the end of our research registries (end of 2021), thus giving time for therapy completion without requiring it. Both iCBT and fCBT were free to the patient, offered by the national public healthcare. There was, on average,

5.5 treatment sessions in the analytic iCBT group of Table 1 (s.d. 2.1, median 7), whilst in only 352 out of the 388 fCBTs (91%) the therapist reported number of sessions (mean 12.3, s.d. 9.8, median 14; reporting was not obligatory due to coupon-system for predefined number of sessions).

Choice of primary measure and covariates

As the primary outcome, we used change in Patient Health Questionnaire -9 (PHQ-9) score, a widely-used and -translated brief public-domain measure with good psychometric properties.[20,21] PHQ-9 score ranges from 0 to 21, with ≥ 1.7 point declines being subjectively detectable (as "feeling better").[22] We computed treatment response as change in PHQ-9 from baseline to last measurement right at post-treatment, denoted ΔPHQ-9 [we did not focus on the follow ups 3 months after iCBT]. On average, 9.4 weeks passed between baseline and follow-up assessments in iCBT for those who completed the last session and 7.7 weeks for all together (including cases where first and last sessions were the same). We estimated 35-week average interval from the less accurate fCBT reports. If a patient discontinued the treatment after providing only a single measurement, a ΔPHQ-9 of 0 was assigned to that patient [note, proceeding in the iCBT program required symptom reports; fCBT therapists were advised to encourage patients to report]. This conservative approach effectively penalises treatments with high drop-out rates compared to complete-data or imputation analysis. Both the fCBTs and the iCBTs also recorded Overall Anxiety Severity and Impairment Scale (OASIS), allowing us to use this continuous-valued score as a baseline predictor of the outcome.[23] Other predicting variables were drawn from the multiple linked administrative registries (see Figure 2B). For privacy and variance-reduction, we only included categorical covariates that resulted in cell sizes 10 observations or above when cross-tabulated with treatment allocation. Table 1 summarises the included background variables.

Statistical analysis

We arrived at our sample size by taking all eligible patients from the register. Our aim was to estimate ATE between fCBT and iCBT. An estimate with 95% confidence intervals (CI) no wider than 1.7 PHQ-9 points [as occurred here] can be considered accurate in the sense of falling below a previously estimated detection threshold between "feeling better" vs. "feeling the same".[22]

For TMLE, we implemented super learners with a broad set of base learners (cf. Examples 6 and 7 in ref[24]), including generalized linear models, Bayesian generalized linear models, generalized additive models, lasso, elastic net and ridge regressions, polynomial splines, multivariate regression splines, extreme gradient boosting, Bayesian additive regression trees, and highly adaptive lasso.[24–27] TMLE requires two super learners, one predicting treatment allocation and another predicting treatment outcome.[13,14] Our super learners used cross-validation (CV) to automatically weight the base learners with the non-negative least squares regression algorithm, having guaranteed of asymptotically better performance compared to the base learners.[13,24] The super learners were used to construct an estimate for ATE, using TMLE to optimize the bias-variance trade-off for the ATE parameter.[13,26] Thus, TMLE controlled for confounding by variables of Table 1 and their machine-learned nonlinear combinations and interactions. We followed the online and published guidance by van der Laan and colleagues,[24,26] using their `tlverse` `R` environment, including packages `sl3` (version 1.4.4), `tmle3` (v.0.2.0), `cvAUC` (v.1.1.4). This implied nested 10-fold and nested stratified [by delivery format] 10-fold CV for treatment outcome and allocation, respectively, when investigating super-learner performance (nesting train-weight base learners CV within select-validate super learner CV; see supplement for illustration). Non-nested (train-select) CV was used within TMLE estimation (target parameter is a counterfactual quantity, precluding validation samples). Multiple imputation applied `superMICE` (v.1.1.1) package (5 chains, 5 iterations). Learners, covariates and imputation choices and examples of the central analysis scripts are further discussed supplementary materials.

As sensitivity analyses, we additionally computed the more classic G-computation and AIPTW estimates of ATE, and a version of TMLE with simple models.[14] Linear regression was used as the outcome model and logistic regression for estimating the probability of receiving fCBT as opposed to iCBT. Online

supplement shows covariate balance investigation recommended for (A)IPTW.[28] All the analyses took place in the Findata authority's secure Kapseli® computing environment, in R version 4.2.2.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Of the 5467 iCBT patients, those who finished the treatment by reaching session #7 (*n* = 3333, 61%) reported a PHQ-9 decline of 4.8 points (s.d. 5.3; paired-sample Cohen's *d* = −0.91; for 14%, symptoms increased). The PHQ-9 decline defined via ΔPHQ-9 in the full sample was 3.6 points (s.d. 5.0; *d* = −0.71). However, the decline was 4.2 points (s.d. 5.2, *d* = −0.81) if excluding the 819 (15%) iCBT patients who only reported the baseline PHQ-9 and therefore had null change in the score. If we were to include as null changes both the patients with only baseline PHQ-9 and the 1800 patients who were prescribed the program but never submitted a single PHQ-9 report (Figure 2A), the PHQ-9 decline would have been 2.7 points (s.d. 4.6, *d* = -0.58).

Of the 392 fCBT patients, those who filled the post-treatment PHQ-9 (n = 316; 81%) reported a mean PHQ-9 score decline of 3.0 points (s.d. 5.2; *d* = −0.59). For 23%, symptoms increased; a greater proportion compared to iCBT ($p < .001$, $\chi_1^2 = 16.96$). The full-data ΔPHQ-9 was 2.4 points (s.d. 4.8; *d* = −0.51). The raw iCBT-fCBT difference in this primary outcome metric in the total sample (Figure 2A) was statistically significant [1.1 points in favor of iCBT; in Welch's t-test, $t = -4.55$, d.f. = 455.04, $p < .001$, Hedge's $g = -0.228$, CI = (−0.331, −0.126)]. If including those without the baseline PHQ-9 as non-changers (resulting in *n* = 413 for fCBT and 7194 for iCBT), the difference between the treatments was 0.4 points and not statistically significant [$g = -0.083$, CI = (−0.182, 0.016)]. In iCBT, PHQ-9 points declined less for men (3.1, s.d. 4.9) compared to female gender (3.8, s.d. 5.0, $p < 0.001$ for difference), but not amongst iCBT-

treatment completers (4.5 vs. 4.9; s.d. 5.4 and 5.3, $p = 0.095$) nor in fCBT (2.4 vs. 2.5; s.d. 4.5 and 4.9, $p = 0.991$).

Table 1 summarises treatment group characteristics at baseline, after removing 25 patients with missing (mostly demographic) baseline data (4 from fCBT, 21 from iCBT; other missingness pertain treatment dropout not affecting Table 1). On average, patients' depression was mild (Table 1). Overall, depression patients directed to fCBT differed from those directed to iCBT. Thus, it is prudent to ask, what would be the treatment effect if giving all the patients fCBT relative to giving them iCBT. This ATE differential is analysed next.

The super learners for treatment outcome and treatment assignment, weighted partly distinct base learners (algorithms) and designated as important only partly overlapping sets of variables (Supplementary Table S2 and Figure S1). Cross-validated area under the curve (CV-AUC) performance measure for the model predicting treatment assignment to fCBT vs. iCBT was .77 (CI = .74, .79), meaning that the model was informative (value .5 indicates chance performance and value 1 perfect performance). Cross-validated R^2 for the outcome prediction model was .031 (~.06 without cross-validation, or $r = .24$). When we targeted the outcome and treatment-assignment models to ATE between fCBT *versus* iCBT, the result was close to the naturalistic treatment differential (Table 2). If all the patients could have been assigned to the same treatment, the ATE would have been nearly the same if all the patients had been given iCBT as when all would have gotten the fCBT – slightly better with iCBT (Table 2, row 2). The result was robust when restricting iCBT treatments to fCBT delivery region (i.e., to residents of Uusimaa) and for alternative counterfactual definitions of dropout (Table 2).

Admitting a certain black-box nature in the TMLE, we computed simpler estimates for ATE as sensitivity analyses. We took 10 most important predictors based on the super learners (Supplementary Figure S1) and instead fitted a logistic regression model predicting treatment assignment and a linear regression model predicting ΔPHQ-9 (Supplementary Table S3 and S4). The assignment model performance was well above chance (CV-AUC = .70, CI = 0.67, 0.73), although inferior to the super learner. Adjusted R^2

for the outcome prediction model was .03. After removal of predictors that were non-significant in both the models (Supplementary Table S3 and S4), their performance remained similar [CV-AUC = .70 (CI = 0.67, 0.73), adjusted R^2 = .03]. Treatment-assignment model-based inverse-probability weighting resulted in much-improved covariate balance across fCBT-iCBT samples (Supplementary Figure S3). The finding of similar efficacy for iCBT and fCBT was robust to all the methodological choices: G-computed ATE 1.073; AIPTW-ATE 0.735 (CI 0.109, 1.361); simple TMLE-ATE 0.790 (CI 0.164, 1.415). Our prediction models did not use baseline PHQ-9 to avoid artificial dependence with outcome (it is part of the outcome already). A sensitivity analysis including it to assignment model had only modest effect to above estimates [G-ATE 1.07, AIPTW-ATE 0.12 (CI = -0.59, 0.85), simple TMLE-ATE 0.11 (CI = -0.57, 0.80)]. When including baseline PHQ-9 in both outcome and assignment models of full TMLE, we estimated ATE 0.06 (CI = -0.64, 0.51).

Discussion

We found no mentionable difference between treatment effectiveness in guided internet *versus* face-to-face delivery of CBT for depression. This finding was consistent across raw data from hospital registries and multiple causal-effect estimates that retrospectively balanced the treatment groups. Our findings encourage wider dissemination of cost-effective and -efficacious guided iCBTs.

Our findings need to be interpreted considering the following limitations. First, the employed methodology achieves causal inference when relevant confounders are observed. Hence, we strived to observe much registry-based background information and interpreted the results relative to RCTs. Nevertheless, variables missing from our set, such as digital affinity, could be of future interest, as well as more accurate temporal and follow-up analyses. Second, the fCBT and iCBT treatments had qualitative differences other than delivery method: The fCBTs were an outsourced service from public mental health care, produced in private practices by licensed psychotherapists. In contrast, the trained iCBT therapists were employees in psychiatric specialty care (psychologists and nurses), but not necessarily licensed psychotherapists. Whereas the fCBT therapists could adjust the treatment specifics during the treatment,

the online platform largely standardized the iCBT program. On average, the fCBTs had more treatment sessions than the iCBTs. However, the number of sessions was uncorrelated with effect sizes in past trials.[3] Third, the study was not pre-registered. However, sensitivity analyses demonstrated robustness of the result to many alternative analytic choices. Fourth, while HUS-iCBT clinical development extensively collaborates with individuals with lived experience, these individuals were not involved in this research specifically. Including their perspectives in future research might add insights.

Finally, many more iCBT patients dropped out from their treatment program compared to the fCBT patients. We took this in account by using the last measure before dropping out to compute the treatment outcome. Thus, iCBT was heavily penalized for its dropouts. Patients dropping out without a second measurement had zero change in their PHQ-9 score, which is less than average spontaneous recovery in a wait-list.[29] Furthermore, one does not always attach a special clinical meaning to dropping out from treatment. For example, the "stepped care model" may improve treatment outcomes by starting patients with a low-intensity treatment and by subsequently stepping non-responders up to a higher-intensity treatment.[30] We found that the average symptom change that *included* those who discontinued was no worse for iCBT than for fCBT, and iCBT had fewer treatment sessions in general. Thus, when dropout is not considered a primary indicator of treatment failure, iCBT provided comparable treatment outcomes to fCBT but achieved these results more quickly.

Our study provides important additional support for guided iCBTs. Research has shown guided iCBTs to attain comparable effects to fCBT in RCTs and in their meta-analyses (efficacy).[1,2] We complemented these evidence by showing that an iCBT program attained comparable effects to fCBT also in naturalistic settings (effectiveness) and when the naturalistic patients were statistically balanced across the treatments (efficacy and effectiveness triangulated). The analytic frameworks of experiments (e.g., RCTs) and counterfactual causal inference (e.g., TMLEs) build on different assumptions but provide consistent evidence. Conceptually, such a situation of causal triangulation is considered the best-case scenario in modern causal-inference literature.[9–11] This is because it is difficult to guarantee the

assumptions of any single method, RCTs included, while it is easier to guarantee that multiple methods did not all fail in distinct ways to provide the same and yet biased finding. If accepting equal treatment effects, a treatment that takes 77 min of therapist time (iCBT) on average should be more *cost*-effective than a treatment that takes 840 min of therapist time (fCBT) – 11-fold more cost-effective.

Given the consistency of the present and past findings across various approaches to evaluating the efficacy of iCBT versus fCBT, our study offers reassurance regarding the validity of the analytical frameworks employed. Similar formal causal inference methods might represent cost-effective means for detecting active treatment components and predictors in mental health research, where confounding issues are pronounced.

We conclude that, for depressive disorders, evidence from RCTs, naturalistic studies and quasi-experimental research all demonstrate comparable treatment efficacy and effectiveness between guided iCBT and fCBT. Consequently, internet delivery of short first-line treatments emerges as a cost-effective and scalable default treatment option. Nevertheless, maintaining a range of treatment alternatives remains important, considering the less-than-optimal efficacy of all available treatments,[8] special needs, and the emerging opportunities in treatment targeting.

Contributors

THR wrote the first draft of the report and conducted the data-analyses. SES, SIS, JT, and JS wrote critical revisions to the manuscript and accepted the final version. THR, SES, and JT have accessed and independently verified the research data, collection of which was monitored by the administrative personnel at Findata and HUS Tietohallinto; respective registry owners have verified the original registries. Only Findata authority had access to the full personal data, with THR, SES, and JT only accessing pseudonymized research data, preprocessed by the Findata authority after consultation of the national Office of the Data Protection Ombudsman. THR and JT independently verified (re-derived) results for all the naturalistic observations (only THR conducted the machine learning, with the code being in the online

supplement). In addition, SES and JS can verify general trends in primary outcome data as employees of HUS Psychiatry. According to CRediT taxonomy, Conceptualization: THR, SES, JS; Data curation: THR, JT; Formal analysis: THR, JT; Investigation: THR, SES, JS; Methodology: THR; Project administration: THR; Resources: THR, SES, JS; Software: THR, JT; Supervision: THR, SES; Validation: THR, SES, JT; Visualization: THR; Writing - original draft: THR; Writing - review & editing: THR, SES, SIS, JT, JS; Literature search: THR, SES, SIS, JT, JS.

Declaration of interests

We declare no competing interests.

Data sharing

As per legislations, the data is available only through a research permission from the Findata, Social and Health Data Permit Authority (https://findata.fi/en/), whose activities are based on the Act on the Secondary Use of Health and Social Data (https://stm.fi/documents/1271139/1365571/The+Act+on+the+Secondary+Use+of+Health+and+Social+Data/a2bca08c-d067-3e54-45d1-18096de0ed76/The+Act+on+the+Secondary+Use+of+Health+and+Social+Data.pdf). The authors have no permission to directly share data with third parties. The analysis scripts are available as parts of the supplementary online material.

Acknowledgements

This study was supported by the Research Council of Finland (a.k.a., Academy of Finland) with grant numbers 334057, 335901, and 358138 to THR and the project Machine-learned causal factors for psychotherapy. The funder of the study had no role in study design, data collection, data analysis, data

interpretation, or writing of the report. We thank Ville Ritola for comments on an earlier version of the manuscript.

References

1 Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *World Psychiatry* 2023; **22**: 305–14.

2 Andersson G, Cuijpers P, Carlbring P, Riper H, Hedman E. Guided internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. *World Psychiatry* 2014; **13**: 288–95.

3 Ciharova M, Karyotaki E, Miguel C, *et al.* Amount and frequency of psychotherapy as predictors of treatment outcome for adult depression: A meta-regression analysis. *J Affect Disord* 2024; **359**: 92–9.

4 Pontén M, Jonsjö M, Vadenmark V, *et al.* Association between expectations and clinical outcomes in online v. face-to-face therapy – an individual participant data meta-analysis. *Psychol Med* 2024; **54**: 1207–14.

5 Moshe I, Terhorst Y, Philippi P, *et al.* Digital interventions for the treatment of depression: A meta-analytic review. *Psychol Bull* 2021; **147**: 749–86.

6 Tröger A, Miguel C, Ciharova M, *et al.* Baseline depression severity as moderator on depression outcomes in psychotherapy and pharmacotherapy. *J Affect Disord* 2024; **344**: 86–99.

7 van der Lem R, de Wever WW, van der Wee NJ, van Veen T, Cuijpers P, Zitman FG. The generalizability of psychotherapy efficacy trials in major depressive disorder: an analysis of the influence of patient selection in efficacy trials on symptom outcome in daily practice. *BMC Psychiatry* 2012; **12**: 192.

8 Ormel J, Hollon SD, Kessler RC, Cuijpers P, Monroe SM. More treatment but no less depression: The treatment-prevalence paradox. *Clin Psychol Rev* 2022; **91**: 102111.

9 Hammerton G, Munafò MR. Causal inference with observational data: the need for triangulation of evidence. *Psychol Med* 2021; **51**: 563–78.

10 Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol* 2016; **45**: 1866–86.

11 Dahabreh IJ, Bibbins-Domingo K. Causal inference about the effects of interventions from observational studies in medical journals. *JAMA* 2024; **331**: 1845–53.

12 Hernán MA, Robins JM. Causal Inference: What If. Boca Raton, US-FL: Chapman & Hall/CRC, 2020 https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

13 van der Laan MJ, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. New York, US-NY: Springer, 2011.

14 Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Stat Med* 2018; **37**: 2530–46.

15 Torvik FA, Ystrom E, Gustavson K, *et al.* Diagnostic and genetic overlap of three common mental disorders in structured interviews and health registries. *Acta Psychiatr Scand* 2018; **137**: 54–64.

16 Saarni SE, Rosenström TH, Stenberg J-H, *et al.* Finnish Psychotherapy Quality Register: rationale, development, and baseline results. *Nord J Psychiatry* 2023; **77**: 455–66.

17 Ritola V, Lipsanen JO, Pihlaja S, *et al.* Internet-delivered cognitive behavioral therapy for generalized anxiety disorder in nationwide routine care: Effectiveness study. *J Med Internet Res* 2022; **24**: e29384.

18 Rosenström TH, Ritola V, Saarni S, Joffe G, Stenberg J-H. Measurement invariant but non-Normal treatment responses in guided internet psychotherapies for depressive and generalized anxiety disorders. *Assessment* 2023; **30**: 618–32.

19 Rosenström TH, Mylläri S, Malkki V, Saarni SE. Feasibility of generic, short, and easy-to-use assessment of psychological distress during psychotherapy: Longitudinal measurement invariance of CORE-10 and -OM. *Psychother Res* 2022; **32**: 1090–9.

20 Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16**: 606–13.

21 Stochl J, Fried EI, Fritz J, *et al.* On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment* 2022; **29**: 355–66.

22 Kounali D, Button KS, Lewis G, *et al.* How much change is enough? Evidence from a longitudinal study on depression in UK primary care. *Psychol Med* 2022; **52**: 1875–82.

23 Norman SB, Cissell SH, Means-Christensen AJ, Stein MB. Development and validation of an Overall Anxiety Severity And Impairment Scale (OASIS). *Depress Anxiety* 2006; **23**: 245–9.

24 Phillips RV, van der Laan MJ, Lee H, Gruber S. Practical considerations for specifying a super learner. *Int J Epidemiol* 2023; **52**: 1276–85.

25 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. New York, USA: Springer-Verlag, 2009.

26 van der Laan M, Coyle J, Hejazi N, Malenica I, Phillips R, Hubbard A. Targeted Learning in R: Causal Data Science with the tlverse Software Ecosystem. 2024 https://tlverse.org/tlverse-handbook/ (accessed April 17, 2024).

27 Benkeser D, van der Laan M. The highly adaptive Lasso estimator. *IEEE International Conference on Data Science and Advanced Analytics* 2016; **2016**: 689–96.

28 Chesnaye NC, Stel VS, Tripepi G, *et al.* An introduction to inverse probability of treatment weighting in observational research. *Clin Kidney J* 2022; **15**: 14–20.

29 Whiteford HA, Harris MG, McKeon G, *et al.* Estimating remission from untreated major depression: a systematic review and meta-analysis. *Psychol Med* 2013; **43**: 1569–85.

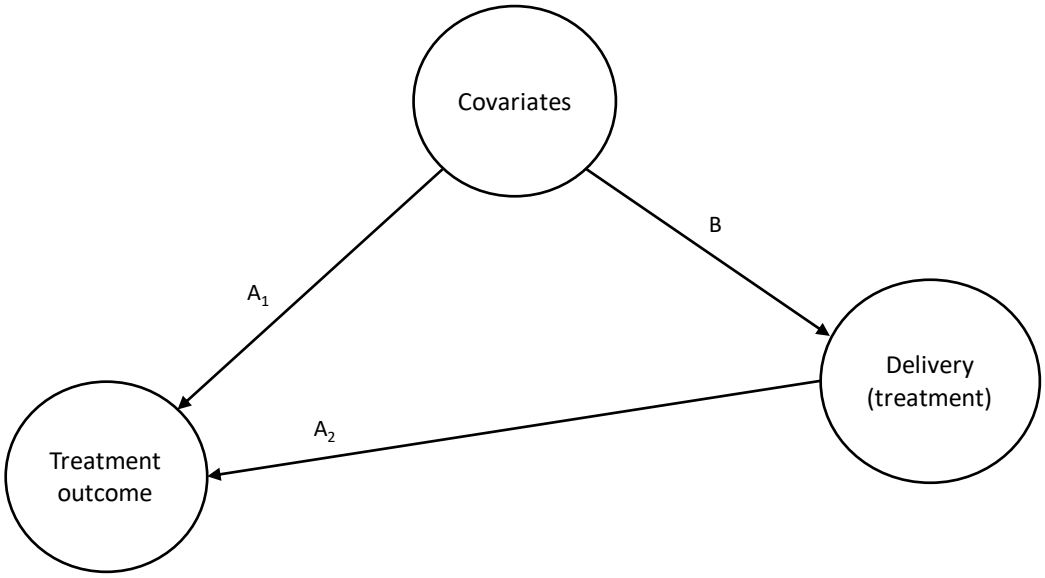
30 van Straten A, Hill J, Richards DA, Cuijpers P. Stepped care treatment delivery for depression: a systematic review and meta-analysis. *Psychol Med* 2015; **45**: 231–46.

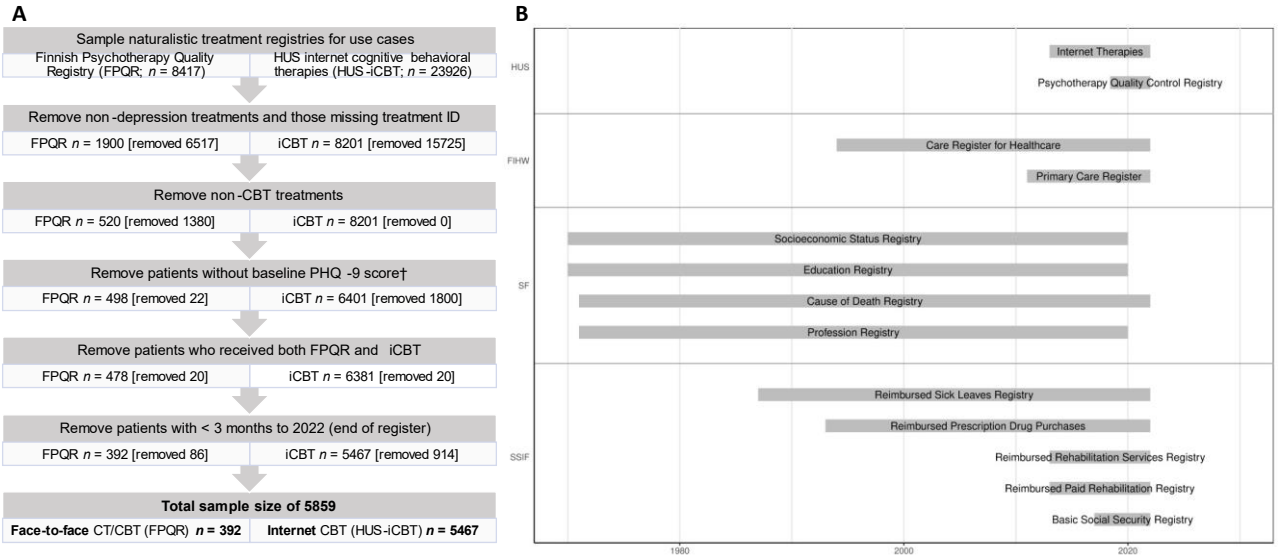
Figure captions

Figure 1. Technical assumptions. Our present methods make the working assumption that treatment outcome depends on the treatment (delivery method). Outcome and treatment may depend on covariates, but all such confounding covariates are assumed fully observed. Other key assumptions are consistency (no other versions of the treatments), no interference (a patient's outcome does not depend on the other patients), and positivity (a positive probability for ending up in both iCBT and fCBT for all predictor covariate values). Finally, one should accurately capture forms of the dependencies indicated by the arrows. Some methods need accuracy for both A and B dependencies, whereas doubly robust methods can achieve consistent estimates with only A's or B and A_2 correctly modelled. Even methods not doubly robust may be preferred over naïve regression analysis approaches, which are biased whenever the non-flexible parametric model is mis-specified relative to reality. Even application of mis-specified models that miss some covariates tends to improve over naïve approaches that estimate average treatment effect with simple regression or mean difference.

Figure 2. Data. **A**) *Patient selection to this study. Note, step "†" also removed duplicate entries to treatment of a same type (FPQR/fCBT or iCBT).* **B**) *Linked health and social security registries (gray lines indicate time spans available to this study). Findata (and Digital and Population Services Agency) linked HUS Helsinki University Hospital (HUS) -owned iCBT and fCBT registries with multiple registries (see figure) from Finnish Institute of Health and Welfare (FIHW), Statistics Finland (SF), and Social Security Institute of Finland (SSIF, a.k.a. Kela). These registries were further reduced to outcome data and to covariate data of Table 1. Some of their contents have been described online (https://aineistokatalogi.fi/catalog/studies). Registry owners have legal obligations to keep registries accurate.*

Figures




A
Sample naturalistic treatment registries for use cases
Finnish Psychotherapy Quality Registry (FPQR; n = 8417)
HUS internet cognitive behavioral therapies (HUS-iCBT; n = 23926)
Remove non-depression treatments and those missing treatment ID
FPQR n = 1900 [removed 6517]
iCBT n = 8201 [removed 15725]
Remove non-CBT treatments
FPQR n = 520 [removed 1380]
iCBT n = 8201 [removed 0]
Remove patients without baseline PHQ -9 score†
FPQR n = 498 [removed 22]
iCBT n = 6401 [removed 1800]
Remove patients who received both FPQR and iCBT
FPQR n = 478 [removed 20]
iCBT n = 6381 [removed 20]
Remove patients with < 3 months to 2022 (end of register)
FPQR n = 392 [removed 86]
iCBT n = 5467 [removed 914]
Total sample size of 5859
Face-to-face CT/CBT (FPQR) n = 392
Internet CBT (HUS-iCBT) n = 5467
B
HUS
FIHW
SF
SSIF
Internet Therapies
Psychotherapy Quality Control Registry
Care Register for Healthcare
Primary Care Register
Socioeconomic Status Registry
Education Registry
Cause of Death Registry
Profession Registry
Reimbursed Sick Leaves Registry
Reimbursed Prescription Drug Purchases
Reimbursed Rehabilitation Services Registry
Reimbursed Paid Rehabilitation Registry
Basic Social Security Registry
1980
2000
2020

SUPPLEMENTARY CONTENT FOR *Rosenström et al., 2025, Lancet Psychiatry,*

https://doi.org/10.1016/S2215-0366(24)00404-8

Table of Supplementary Contents

Additional information (supplementary text) 2
Super learners 2
Cross-validation 4
Multiple imputation 5
Supplementary Table S1. Considerations on covariate selection. 6
Supplementary Table S2. Super learners. 13
Supplementary Table S3. Linear models. 14
Supplementary Table S4. Logistic models. 15
Supplementary Table S5. Important variable descriptives. 16
Supplementary Figure S1. Variable importance. 17
Supplementary Figure S2. Imputation diagnostics 18
Supplementary Figure S3. Covariate balance. 19
Supplementary Figure S4. Violin plots of important predictors. 20
Supplementary Figure S5. Time accumulation of included psychotherapies. 21
Supplementary R script on TMLE estimation 22
Supplementary R script on alternative (sensitivity) analyses on ATE estimation 25
Supplementary R scripts on imputation and constructed variables 27
Supplementary R scripts on full-analysis documentation (online repository) 29

Additional information (supplementary text)

Super learners

We briefly explain super learners and our rationale behind some related technical choices. This text is not intended as a tutorial but as help for those readers who do not yet want to invest time on lengthy tutorials. A super learner is an ensemble algorithm that considers multiple other machine-learning algorithms (base learners) as candidates and, in this case, weights them based on their success in cross-validated prediction. Proofs exist that such a super leaner asymptotically achieves at least as good performance as the best of its base learners. With a diverse set of base learners, a super learner then effectively reduces the burden of choice and exploration in predictive inference. To further reduce such burden, we chose our set of base learners from Examples 6 and 7 from Phillips et al. (ref 24 in main text). Example 7 was a recommended set for "complex real data, smaller sample size, causal inference application" and Example 6 motivated minor modifications for "larger sample size". We additionally added a mean predictor for comparison and cross-validated R^2 calculation. Some of the base learners are typical or regularized linear models, whereas others better represent complex interactions (e.g., tree-based methods). Phillips et al. considered that this "library is well-rounded in its learning strategies and can adapt to a diversity of true functional forms in a robust way". It includes "complex machine learning algorithms that are capable of learning complicated, but potentially relevant, interactions" but also "a range of more simplistic algorithms" (their supplement).

A prediction from such a super learner is a (non-negatively) weighted sum of base-learner predictions (see Table S2 for our estimated weights). Table S2 also shows cross-validated performance for each base learner and the super learner, measured as prediction error/loss (standard error is shown besides mean loss). Thus, Table S2 gives information on what kind of algorithms were weighted in prediction and how much. An avid reader may want to study Phillips et al. (ref 24 in main text) and similar sources. However, these details are not very central for the present causal estimate. Rather, here we are more interested in the oracle property and in *not having to worry* about the details. For comprehensiveness, the table below nevertheless specifies some of our considerations on the properties and technical details of the algorithms, but we stress that they are not essential for understanding and may be better described in the respective original literatures.

Acronym in `tlverse` package	**Informal description**	**Reason for inclusion**	**Technical (hyper)parameters**
Lrnr_mean	Predicts validation-data outcome with average training outcome.	Useful for computing R^2. Reveals fundamental unpredictability if weighted.	None.
Lrnr_glm_fast_TRUE_Cholesky	Ordinary linear or logistic regression (computed fast).	A basic model. Ensures one doesn't fool oneself.	Cholesky method, using intercept
Lrnr_bayesglm_TRUE	Bayesian linear or logistic regression. Uses *t*-distribution as a prior for the regression coefficients.	Like above, but encourages some coefficients to small values, countering overfitting.	Defaults from the `arm` package.
Lrnr_gam_NULL_NULL_GCV.Cp	Generalized additive model. The effects are additive as in glm, but can be nonlinear up to a smoothing penalty for wiggliness (to control overfitting).	Allows non-linear effects of covariates in glm. This covers the possibility that some covariate(s) have non-simple univariate relation with the outcome.	Default `gam` in `mgcv` (solves the smoothing parameter estimation problem by using the Generalized Cross Validation (GCV) criterion, and model fitting by Penalized

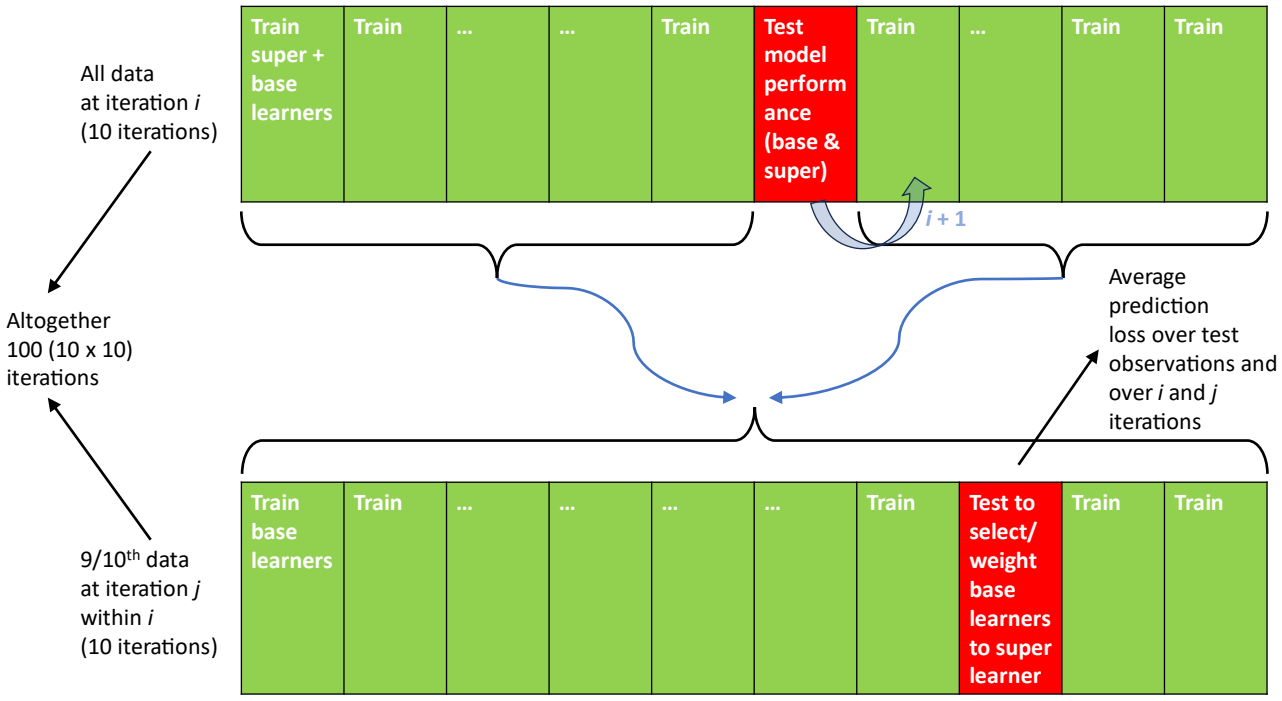
			Iteratively Re-weighted Least Squares).
Lrnr_glmnet_NULL_deviance_ 10_1_100_TRUE_FALSE	Lasso regression. Linear or logistic. Like glm but encourages some coefficients to zero values (automatically excludes bad covariates).	To automatically incorporate variable selection to linear models.	`glmnet` with α = 1 and 10-fold CV to select λ from 100 alternatives.
Lrnr_glmnet_NULL_deviance_ 10_0.5_100_TRUE_FALSE	Regression with Lasso- and Ridge penalty (i.e., like above and like below).	To automatically incorporate variable-group selection to linear models.	`glmnet` with α = 0.5 and 10-fold CV to select λ from 100 alternatives.
Lrnr_glmnet_NULL_deviance_ 10_0_100_TRUE_FALSE	Ridge regression Linear or logistic. Like glm but prevents coefficient trade-off by bounding sum of their absolute values.	To prevent overfitting in linear models without forcing variable selection.	`glmnet` with α = 0 and 10-fold CV to select λ from 100 alternatives.
Lrnr_earth_2_3_backward_0_1_0_0	Multivariate adaptive spline regression. Uses hinge functions of covariates to allow nonlinearities like gam but also interaction effects.	To model nonlinearities and interactions and to automatically select variables and contain outliers (to hinges).	`earth` from `earth` with interaction degree 2, "backward" pruning method, and GCV penalty per knot 3.
Lrnr_polspline_5	Multivariate adaptive polynomial spline regression.	Like above but uses some different technical solutions, which might make a difference in matters of this complexity.	Near-default `polymars` or `polyclass` from `polspline` package (5-fold GCV).
Lrnr_caret_xgbTree_RMSE	eXtreme Gradient Boosting. A regression tree-based method with many technological innovations.	Like above but an award-winning combination of techniques in ML competitions (penalization, boosting, shrinking, randomization, etc.).	`train` model "xgbTree" from `caret` with 10-fold CV and "RMSE" metric.
Lrnr_dbarts_1000_100_200_FALSE_ TRUE_FALSE	Discrete Bayesian Additive Regression Trees	Like above, but in a Bayesian framework: data can overcome assumptions about required tree-depth and shrinkage (i.e., tuning parameters are less rigid).	`bart` from `dbarts`, selecting 1000 posterior draws after burn-in.
Lrnr_hal9001_2_3	Scalable Highly Adaptive Lasso. Builds adaptive basis from covariates and their interactions to optimally estimate functions of bounded variation norm.	Like above but proven optimal estimator for some classes of predictor functions.	`fit_hal` from `hal9001`, with max interaction order 2 and max no of basis function knot points 3.

We strived for both a decent black-box estimate (TMLE with super learners) and simpler interpretable estimates (see sensitivity analyses). These cross-check each other in the following sense. Super learners ensure that simple models do not miss features that affect the balance between iCBT and fCBT or counterfactual predictions of outcome. When resulting in similar inferences as super learners, the simpler models provide interpretability and ensure that inferences from the black box do not contain unforeseen errors. Should there be a conflict between causal inference based on super learners *versus* the same inference from simpler models, that would offer a motivation to investigate the specific complex features behind the difference (a difficult task often). Given the lack of such difference here, the estimates based on super learners can instead be considered as robustness checks for the simpler models. Such a check is important because an arrow in Figure 1 of the main text indicates a dependence relation that can take infinitely many forms rather than one (e.g., a linear) form.

Cross-validation

In validation, one estimates a model using "training" data (selected patients) and then tests it in different datums (here patients not used in model estimation). Test performance is assessed via some loss function, $L(y_{predicted}, y_{test})$, that measures discrepancy between model-predicted and observed values. Typically, loss is computed for each observation (each patient) and then averaged over the test patients. If 1/10 of the patients were set aside for testing, test-sample size is $n/10$, where n is the original sample size. Cross-validation (CV) is like validation, but it repeats the process one by one for each $1/10^{th}$ of the data, so that the test-sample size is n but any datum is never used to both estimate and test *at the same time*. Compared to simple validation, nothing is lost in terms of overfitting propensity, but the test sample gets 10-fold. Hence the name 10-fold CV. Of course, one may choose other value K to get K-fold CV, but 10 is a common choice that offers a good balance between training sample sizes and computational burden.

A super learner selects models based on their CV performance, but it cannot be tested in the same test sets because then the test values would have affected its estimation, causing overfitting to idiosyncratic data ("data-leakage" in validation). The selection of its base learners must instead be nested within its training sets and, because the selection is by CV, the process becomes a "nested 10-fold CV" (see below illustration). Usually, the splits to ten sub-samples are made via random sampling. When such random sampling is carried out separately within classes of a variable (here separately for iCBT and fCBT patients), then the CV is stratified. This may stabilize behaviour when classes are imbalanced in sample size by ensuring sufficient outcome variance for all the CV folds. Here, we used these procedures. We avoided algorithmic *a priori* "feature" (i.e., variable) construction that would apply to both test and training sets in CV, because that would be another form of "data leakage" and risk for low generalizability of the findings.



Multiple imputation

Multiple imputation using chained equations is a standard way to handle missing data (see, e.g., free online book by van Buuren: https://stefvanbuuren.name/fimd/). In such imputation of missing values, one should ensure that dependencies of Figure 1 of the main text are not diluted. Hence, the imputation model should not be too simple (inflexible) compared to the final analysis model. In our case, super-learner-based imputation is thus needed. We used the R package "superMICE" for it. It allowed very much more flexibility compared to other imputation approaches [though some of the base learners we tried (e.g., Ridge) resulted in errors and could not be used]. As per recommendation, we chose a kernel bandwidth containing ~2% of observations within kernel standard deviation. The exact imputation script is given in a supplementary code below.

Multiple imputation is based on the idea of properly estimating the variance of full-data quantities under missing information. Imputing an observation only once and using it for estimation would inflate the certainty of that unknown value relative to the imputing model's estimate of uncertainty. Hence, multiple such imputations are used (here 5 per value) to properly account variability in estimates. Because some missing values depend on missing values in other variables, imputed values must be computed by iterating over variables. Such rounds of iteration are themselves iterated until "convergence" is achieved, meaning that the distribution of the imputed values stays stable and becomes independent of initial values. Supplementary Figure S2 shows standard diagnostic information on the imputation algorithm that verify reasonable imputations and convergence of the algorithm.

Supplementary Table S1. Considerations on covariate selection.

Covariate	Rationale
Continuous or count variables	
Baseline Symptom Severity	
PHQ-9 at 1st assessment (symptom score; included only in sensitivity analyses) OASIS at 1st assessment (symptom score)	Baseline symptom severity has been linked to better iCBT outcomes[1–3], while its role in face-to-face CBT outcomes appears to be more modest[4,5]. Moreover, iCBT in HUS is targeted for individuals with mild to moderate symptoms who may benefit from structured therapy without intensive therapist involvement, whereas fCBT functions as a higher-intensity intervention for patients requiring more comprehensive support[6–10].
Mental Health and Somatic Comorbidity	
Distinct past F diagnoses (count)	The comorbidity of mental health disorders may complicate treatment outcomes[11–14]. Moreover, patients with comorbidities may be more often allocated to fCBT, a higher-intensity intervention designed for those requiring more comprehensive support, whereas iCBT is intended for patients who may benefit from structured therapy without intensive therapist involvement[6–10].
Distinct past non-F diagnoses (count)	Somatic health problems may complicate treatment outcomes (e.g., increased drop-out, less improvement)[15,16].
Family Dynamics	
Number of relatives (count of parents and siblings)	The number of relatives may interact with familial risk for mental health disorders, potentially acting as either a protective (e.g., having a large, supportive extended family) or risk (e.g., exposure to several family members with mental health issues) factor. From a confounding perspective, it is also a safe variable to include in the models (not a mediator or collider).
Familial Risk for Mental Health Disorders	
Maternal F diagnoses (count) Paternal F diagnoses (count) Siblings' F diagnoses per sibling (count)	F diagnoses in relatives indicate a familial risk for mental health disorders[17–23].
Pharmacological Treatment	
Distinct purchased medicine (ATC-N) prescriptions (count) Distinct ATC-N medicine purchase events (count)	CBT with combined pharmacological treatment can lead to better treatment outcomes[5,24–26].
Functional Impairment and Adversity	
Insurance-reimbursed sickness absence days (count) Insurance-reimbursed sickness absence days with F diagnoses (count) Previous psychiatry visits (count)	Patients with more severe functional impairments and adversities —evidenced by increased days of sickness absence and visits to mental health professionals—may be more likely to be allocated to fCBT tailored for patients who need more

	comprehensive support[6,7,9]. In turn, iCBT is designed for patients with relatively good functional capacity and the ability to work independently with minimal professional support [6–10].
Cohort and Maturation Processes	
Age (years)	Patients' age may influence treatment outcomes (e.g., older patients benefit more from iCBT, while young adults may respond best to face-to-face therapy)[2,3,16,27].
Binary-valued variables	
Functional Impairment and Adversity	
Rehabilitation money received Income-support received Rehabilitation-support received Rehabilitation-psychotherapy received	Patients with more severe functional impairments and adversities—evidenced by different social benefits received—may be more likely to be allocated to fCBT tailored for patients who need more comprehensive support[6,7,9]. In turn, iCBT is designed for patients with relatively good functional capacity and the ability to work independently with minimal professional support [6–10].
Biological and Socialization Processes	
Male gender	Female gender increases the risk for depression[28]; it also increases adherence to internet-based therapies[29,30].
Specific vs Unspecific Risk for Depression	
First F diagnosis for depression Most frequent F diagnosis for depression	The number of previous depressive episodes is one of the most robust risk factors for future depression[31–33].
First F diagnosis other* Most frequent F diagnosis other*	Other mental health disorders are risk factors for depression[32,34,35].
Familial Risk for Mental Health Disorders	
Siblings' most frequent F diagnosis for depression Mother's most frequent F diagnosis for depression Father's most frequent F diagnosis for depression	F diagnoses for depression in relatives indicate a familial risk for depression[17,18,20,23,36].
Siblings' most frequent F diagnosis other** Mother's most frequent F diagnosis other* Father's most frequent F diagnosis other*	Other F diagnoses in relatives indicate a familial risk for depression[19,22,36].
Siblings' most frequent F diagnosis for psychotic disorder	F diagnosis for psychotic disorder indicates familial risk for severe mental illness[19,36,37].
Most Commonly Prescribed Medications	
Most frequently purchased medicine was Benzodiazepine	Short-term use of benzodiazepines may be linked to favorable outcomes in patients with depression[38,39]. However, prolonged use carries risks, including addictions and cognitive impairments, which may undermine treatment efficacy[40–42].
Most frequently purchased medicine for migraine	Migraine is a risk factor for depression[43,44] that might complicate its treatment.
Most frequently purchased medicine non-SSRI depression Most frequently purchased medicine SSRI depression medicine	Combined pharmacological treatment may lead to better treatment outcomes[5,24–26].

Most frequently purchased medicine Paracetamol	Physical pain is a risk factor for depression[45], and its presence can complicate treatment outcomes[46].
Most frequently purchased medicine for psychosis	Comorbid severe mental illness complicates the treatment of depression[47,48].
Most frequently purchased medicine for sleep problems	Sleep problems are a risk factor for depression[49,50] that may complicate its treatment[51].
Most frequently purchased medicine Opiate	Opiate use is a risk factor for depression[52–55]. In addition, opiate dependency complicates the treatment of depression[42].
Indicators of Socioeconomic Status	
Entrepreneur	
Upper white-collar employee	
Lower white-collar employee	
Manual labourer	Lower socioeconomic status may be a risk factor for poorer treatment outcomes[1,56–58].
Student	
Retiree	
Unemployed or working status unknown	
Culture and Ethnicity	
Native language Finnish	The iCBT program was available only in Finnish until Nov 2023. However, patients are accepted to the treatment if they are able to understand the program contents and communicate with the therapist in English or Swedish. Native participants may benefit more from iCBT[2], whereas ethnicity has generally not been linked to the effectiveness of face-to-face therapies[59,60].

References

1 Karyotaki E, Efthimiou O, Miguel C, *et al.* Internet-based cognitive behavioral therapy for depression: A systematic review and individual patient data network meta-analysis. *JAMA Psychiatry* 2021; **78**: 361–71.

2 Karyotaki E, Ebert DD, Donkin L, *et al.* Do guided internet-based interventions result in clinically relevant changes for patients with depression? An individual participant data meta-analysis. *Clin Psychol Rev* 2018; **63**: 80–92.

3 Reins JA, Buntrock C, Zimmermann J, *et al.* Efficacy and moderators of internet-based interventions in adults with subthreshold depression: An individual participant data meta-analysis of randomized controlled trials. *Psychother Psychosom* 2020; **90**: 94–106.

4 Furukawa TA, Weitz ES, Tanaka S, *et al.* Initial severity of depression and efficacy of cognitive–behavioural therapy: Individual-participant data meta-analysis of pill-placebo-controlled trials. *Br J Psychiatry* 2017; **210**: 190–6.

5 Cuijpers P, Miguel C, Harrer M, *et al.* Psychological treatment of depression: A systematic overview of a 'Meta-Analytic Research Domain'. *J Affect Disord* 2023; **335**: 141–51.

6 Rivero-Santana A, Perestelo-Perez L, Alvarez-Perez Y, *et al.* Stepped care for the treatment of depression: a systematic review and meta-analysis. *J Affect Disord* 2021; **294**: 391–409.

7 Bower P, Gilbody S. Stepped care in psychological therapies: access, effectiveness and efficiency: Narrative literature review. *Br J Psychiatry* 2005; **186**: 11–7.

8 Mentalhub.fi. What is Internet Therapy? [Internet]. Available from: https://www.mielenterveystalo.fi/en/what-internet-therapy (accessed Nov 12, 2024).

9 Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *Int Rev Psychiatry* 2011; **23**: 318–27.

10 Furukawa TA, Suganuma A, Ostinelli EG, *et al.* Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. *Lancet Psychiatry* 2021; **8**: 500–11.

11 Angelakis I, Huggett C, Gooding P, Panagioti M, Hodkinson A. Effectiveness of cognitive–behavioural therapies of varying complexity in reducing depression in adults: systematic review and network meta-analysis. *Br J Psychiatry* 2022; **221**: 459–67.

12 Flygare A-L, Engström I, Hasselgren M, *et al.* Internet-based CBT for patients with depressive disorders in primary and psychiatric care: Is it effective and does comorbidity affect outcome? *Internet Interv* 2020; **19**: 100303.

13 Goddard E, Wingrove J, Moran P. The impact of comorbid personality difficulties on response to IAPT treatment for depression and anxiety. *Behav Res Ther* 2015; **73**: 1–7.

14 Rayner C, Coleman JRI, Purves KL, *et al.* A genome-wide association meta-analysis of prognostic outcomes following cognitive behavioural therapy in individuals with anxiety and depressive disorders. *Transl Psychiatry* 2019; **9**: 1–13.

15 Miguel C, Karyotaki E, Ciharova M, Cristea IA, Penninx BWJH, Cuijpers P. Psychotherapy for comorbid depression and somatic disorders: a systematic review and meta-analysis. *Psychol Med* 2023; **53**: 2503–13.

16 Catarino A, Bateup S, Tablan V, *et al.* Demographic and clinical predictors of response to internet-enabled cognitive–behavioural therapy for depression and anxiety. *Br J Psych Open* 2018; **4**: 411–8.

17 Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: Review and meta-analysis. *Am J Psychiatry* 2000; **157**: 1552–62.

18 Gronemann FH, Jacobsen RK, Wium-Andersen MK, Jørgensen MB, Osler M, Jørgensen TSH. Association of familial aggregation of major depression with risk of major depression. *JAMA Psychiatry* 2023; **80**: 350–9.

19 Rasic D, Hajek T, Alda M, Uher R. Risk of mental illness in offspring of parents with schizophrenia, bipolar disorder, and major depressive disorder: A meta-analysis of family high-risk studies. *Schizophr Bull* 2014; **40**: 28–38.

20 van Sprang ED, Maciejewski DF, Milaneschi Y, *et al.* Familial risk for depressive and anxiety disorders: associations with genetic, clinical, and psychosocial vulnerabilities. *Psychological Medicine* 2022; **52**: 696–706.

21 Sellers R, Collishaw S, Rice F, *et al.* Risk of psychopathology in adolescent offspring of mothers with psychopathology and recurrent depression. *Br J Psychiatry* 2013; **202**: 108–14.

22 Hindley G, Frei O, Shadrin AA, *et al.* Charting the landscape of genetic overlap between mental disorders and related traits beyond genetic correlation. *Am J Psychiatry* 2022; **179**: 833–43.

23 Rhee SJ, Abrahamsson L, Sundquist J, Sundquist K, Kendler KS. The risks for major psychiatric disorders in the siblings of probands with major depressive disorder. *Mol Psychiatry* 2024; published online July 07. DOI: 10.1038/s41380-024-02650-1.

24 Cuijpers P, Miguel C, Harrer M, *et al.* Cognitive behavior therapy vs. control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: a comprehensive meta-analysis including 409 trials with 52,702 patients. *World Psychiatry* 2023; **22**: 105–15.

25 Cuijpers P, Noma H, Karyotaki E, Vinkers CH, Cipriani A, Furukawa TA. A network meta-analysis of the effects of psychotherapies, pharmacotherapies and their combination in the treatment of adult depression. *World Psychiatry* 2020; **19**: 92–107.

26 Cuijpers P, Oud M, Karyotaki E, *et al.* Psychologic treatment of depression compared with pharmacotherapy and combined treatment in primary care: A network meta-analysis. *Ann Fam Med* 2021; **19**: 262–70.

27 Cuijpers P, Karyotaki E, Eckshtain D, *et al.* Psychotherapy for depression across different age groups: A systematic review and meta-analysis. *JAMA Psychiatry* 2020; **77**: 694–702.

28 Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychol Bull* 2017; **143**: 783–822.

29 Karyotaki E, Kleiboer A, Smit F, *et al.* Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. *Psychol Med* 2015; **45**: 2717–26.

30 Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. *IntJ Behav Med* 2016; **23**: 776–94.

31 Hardeveld F, Spijker J, De Graaf R, Nolen WA, Beekman ATF. Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatrica Scand* 2010; **122**: 184–91.

32 Buckman JEJ, Underwood A, Clarke K, *et al.* Risk factors for relapse and recurrence of depression in adults and how they operate: A four-phase systematic review and meta-synthesis. *Clin Psychol Rev* 2018; **64**: 13–38.

33 Wojnarowski C, Firth N, Finegan M, Delgadillo J. Predictors of depression relapse and recurrence after cognitive behavioural therapy: a systematic review and meta-analysis. *Behav Cogn Psychother* 2019; **47**: 514–29.

34 Jacobson N, Newman M. Anxiety and depression as bidirectional risk factors for one another: A meta-analysis of longitudinal studies. *Psychol Bull* 2017; **143**: 1155–200.

35 Saha S, Lim CCW, Cannon DL, *et al.* Co-morbidity between mood and anxiety disorders: A systematic review and meta-analysis. *Depress Anxiety* 2021; **38**: 286–306.

36 Wilde A, Chan H-N, Rahman B, *et al.* A meta-analysis of the risk of major affective disorder in relatives of individuals affected by major depressive disorder or bipolar disorder. *J Affect Disord* 2014; **158**: 37–47.

37 Lo LE, Kaur R, Meiser B, Green Mj. Risk of schizophrenia in relatives of individuals affected by schizophrenia: A meta-analysis. *Psychiatry Res* 2020; **286**: 112852.

38 Benasi G, Guidi J, Offidani E, Balon R, Rickels K, Fava GA. Benzodiazepines as a monotherapy in depressive disorders: A systematic review. *Psychother Psychosom* 2018; **87**: 65–74.

39 Duan Y, Wei J, Geng W, *et al.* The effect of short-term use of benzodiazepines on cognitive function of major depressive disorder patients being treated with antidepressants. *J Affect Disord* 2019; **256**: 1–7.

40 Crowe SF, Stranks EK. The residual medium and long-term cognitive effects of benzodiazepine use: An updated meta-analysis. *Arch Clin Neuropsychol* 2018; **33**: 901–11.

41 Lader M. Benzodiazepines revisited—will we ever learn? *Addiction* 2011; **106**: 2086–109.

42 Kelly TM, Daley DC, Douaihy AB. Treatment of substance abusing patients with comorbid psychiatric disorders. *Addict Behav* 2012; **37**: 11–24.

43 Amiri S, Behnezhad S, Azad E. Migraine headache and depression in adults: a systematic review and reta-analysis. *Neuropsychiatr* 2019; **33**: 131–40.

44 Falla K, Kuziek J, Mahnaz SR, Noel M, Ronksley PE, Orr SL. Anxiety and depressive symptoms and disorders in children and adolescents with migraine: A systematic review and meta-analysis. *JAMA Pediatr* 2022; **176**: 1176–87.

45 Stubbs B, Vancampfort D, Veronese N, *et al.* Depression and pain: primary data and meta-analysis among 237 952 people across 47 low- and middle-income countries. *Psychol Med* 2017; **47**: 2906–17.

46 Liu JJ, Huang X, Bao Y-P, *et al.* Painful physical symptoms and antidepressant treatment outcome in depression: a systematic review and meta-analysis. *Mol Psychiatry* 2024; **29**: 2560–7.

47 Dubovsky SL, Ghosh BM, Serotte JC, Cranwell V. Psychotic depression: Diagnosis, differential diagnosis, and treatment. *Psychother Psychosom* 2021; **90**: 160–77.

48 van Os J, Reininghaus U. Psychosis as a transdiagnostic and extended phenotype in the general population. *World Psychiatry* 2016; **15**: 118–24.

49 Zhang M-M, Ma Y, Du L-T, *et al.* Sleep disorders and non-sleep circadian disorders predict depression: A systematic review and meta-analysis of longitudinal studies. *Neurosci Biobehav Rev* 2022; **134**: 104532.

50 Baglioni C, Battagliese G, Feige B, *et al.* Insomnia as a predictor of depression: A meta-analytic evaluation of longitudinal epidemiological studies. *J Affect Disord* 2011; **135**: 10–9.

51 Troxel WM, Kupfer DJ, Reynolds CF, *et al.* Insomnia and objectively measured sleep disturbances predict treatment outcome in depressed patients treated with psychotherapy or psychotherapy-pharmacotherapy combinations. *J Clin Psychiatry* 2011; **73**: 478.

52 Scherrer JF, Salas J, Copeland LA, *et al.* Prescription opioid duration, dose, and increased risk of depression in 3 large patient populations. *Ann Fam Med* 2016; **14**: 54–62.

53 Scherrer JF, Svrakic DM, Freedland KE, *et al.* Prescription opioid analgesics increase the risk of depression. *J Gen Intern Med* 2014; **29**: 491–9.

54 Leung J, Santo T Jr, Colledge-Frisby S, *et al.* Mood and anxiety symptoms in persons taking prescription opioids: A systematic review with meta-analyses of longitudinal studies. *Pain Med* 2022; **23**: 1442–56.

55 Rosoff DB, Smith GD, Lohoff FW. Prescription opioid use and risk for major depressive disorder and anxiety and stress-related disorders: A multivariable mendelian randomization analysis. *JAMA Psychiatry* 2021; **78**: 151–60.

56 Finegan M, Firth N, Wojnarowski C, Delgadillo J. Associations between socioeconomic status and psychological therapy outcomes: A systematic review and meta-analysis. *Depress Anxiety* 2018; **35**: 560–73.

57 Leppänen H, Kampman O, Autio R, *et al.* Socioeconomic factors and use of psychotherapy in common mental disorders predisposing to disability pension. *BMC Health Serv Res* 2022; **22**: 983.

58 Buckman JEJ, Saunders R, Stott J, *et al.* Socioeconomic indicators of treatment prognosis for adults with depression: A systematic review and individual patient data meta-analysis. *JAMA Psychiatry* 2022; **79**: 406–16.

59 Cougle JR, Grubaugh AL. Do psychosocial treatment outcomes vary by race or ethnicity? A review of meta-analyses. *Clin Psychol Rev* 2022; **96**: 102192.

60 Ünlü Ince B, Riper H, van 't Hof E, Cuijpers P. The effects of psychotherapy on depression among racial-ethnic minority groups: A metaregression analysis. *Psychiatr Serv* 2014; **65**: 612–7.

Supplementary Table S2. Super learners. **Weights and cross-validation errors for the super learners for (A) treatment outcome and (B) treatment assignment**

(A) Outcome-prediction learners/models	**Weights**	**MSE**	**SE**
Lrnr_mean	0	25.094	0.524
Lrnr_glm_fast_TRUE_Cholesky	0	24.390	0.514
Lrnr_bayesglm_TRUE	0.123	24.390	0.514
Lrnr_gam_NULL_NULL_GCV.Cp	0	24.470	0.518
Lrnr_glmnet_NULL_deviance_10_1_100_TRUE_FALSE	0	24.345	0.513
Lrnr_glmnet_NULL_deviance_10_0.5_100_TRUE_FALSE	0	24.364	0.513
Lrnr_glmnet_NULL_deviance_10_0_100_TRUE_FALSE	0.257	24.328	0.512
Lrnr_polspline_5	0.184	24.405	0.518
Lrnr_earth_2_3_backward_0_1_0_0	0.093	24.851	0.528
Lrnr_caret_xgbTree_RMSE	0.346	24.320	0.518
Lrnr_dbarts_1000_100_200_FALSE_TRUE_FALSE	0	24.452	0.516
Lrnr_hal9001_2_3	0	24.917	0.526
SuperLearner	–	24.316	0.515
(B) Treatment-prediction learners/models	**Weights**	**NLL**	**SE**
Lrnr_mean	0	0.245	0.009
Lrnr_glm_fast_TRUE_Cholesky	0.036	0.220	0.009
Lrnr_bayesglm_TRUE	0	0.220	0.008
Lrnr_gam_NULL_NULL_GCV.Cp	0.113	0.220	0.009
Lrnr_glmnet_NULL_deviance_10_1_100_TRUE_FALSE	0	0.220	0.008
Lrnr_glmnet_NULL_deviance_10_0.5_100_TRUE_FALSE	0	0.219	0.008
Lrnr_glmnet_NULL_deviance_10_0_100_TRUE_FALSE	0	0.219	0.008
Lrnr_polspline_5	0.240	0.217	0.008
Lrnr_earth_2_3_backward_0_1_0_0	0.585	0.215	0.009
Lrnr_dbarts_1000_100_200_FALSE_TRUE_FALSE	0	0.732	0.001
Lrnr_hal9001_2_3	0.070	0.223	0.009
SuperLearner	–	0.204	0.008

Note: Learner names use `tlverse` convention (https://tlverse.org). Weights are for non-negative least squares coefficient in the ensemble learner (i.e., super learner), indicating the importance of the base learner in it. Mean squared error (MSE) is cross-validated prediction error for the continuous-valued outcome prediction and negative log-likelihood (NLL) is the cross-validated prediction error for the binary treatment-assignment prediction (SE stands for standard error). The super learner was cross-validated by nesting two cross-validation procedures. See above supplementary text for longer explanations on the quantities.

Supplementary Table S3. Linear models. **Linear regression models for outcome (ΔPHQ-9) prediction**

	Model 1 (initial)			Model 2 (simplified)		
Independent variable	**β**	**95% CI**	**p**	**β**	**95% CI**	**p**
Intercept	-0.76	(-1.36, -0.15)	0.014	-0.80	(-1.40, -0.20)	0.009
fCBT (ref. iCBT)	1.06	(0.54, 1.59)	<0.001	1.07	(0.55, 1.60)	<0.001
OASIS at 1st assessment (score)	-0.16	(-0.19, -0.12)	<0.001	-0.16	(-0.19, -0.12)	<0.001
Age (years)	-0.05	(-0.06, -0.03)	<0.001	-0.04	(-0.06, -0.03)	<0.001
Sex (male)	0.55	(0.27, 0.84)	<0.001	0.56	(0.27, 0.84)	<0.001
Distinct purchased medicine (ATC-N) prescriptions (count)	0.03	(-0.02, 0.08)	0.221	0.03	(-0.01, 0.07)	0.158
Distinct past F diagnoses (count)	0.10	(0.03, 0.17)	0.006	0.10	(0.03, 0.17)	0.008
Distinct past non-F diagnoses (count)	0.00	(-0.01, 0.02)	0.797	0.01	(-0.01, 0.02)	0.421
Previous psychiatry visits (count)	0.00	(0.00, 0.00)	0.979	0.00	(0.00, 0.00)	0.850
Distinct ATC-N medicine purchase events (count)	0.00	(0.00, 0.00)	0.949	–	–	–
Insurance-reimbursed sickness absence days with F diag. (count)	0.00	(0.00, 0.00)	0.091	–	–	–
Insurance-reimbursed sickness absence days (count)	0.00	(0.00, 0.00)	0.029	0.00	(0.00, 0.00)	0.160
Rehabilitation money received	-0.46	(-2.07, 1.16)	0.579	-0.47	(-2.09, 1.14)	0.567
Lower white-collar employee	-0.64	(-0.95, -0.34)	<0.001	-0.64	(-0.94, -0.34)	<0.001
Upper white-collar employee	-0.50	(-0.89, -0.12)	0.01	-0.50	(-0.88, -0.13)	0.009
Entrepreneur	-0.20	(-0.90, 0.50)	0.579	–	–	–
Income-support received	0.45	(0.11, 0.78)	0.009	0.44	(0.11, 0.78)	0.009

Note: Values are regression coefficients and their 95% confidence intervals and p-values. “Model 1” uses covariates that attained “importance” in Supplementary Figure S1 (10 from both A and B panels, which partly overlap), whereas “Model 2” uses only a subset that were statistically significant Model 1 predictors for outcome, treatment assignment, or both (i.e., in Models 1 of Table S2 and/or S3).

Supplementary Table S4. Logistic models. **Logistic regression models for fCBT treatment-group membership**

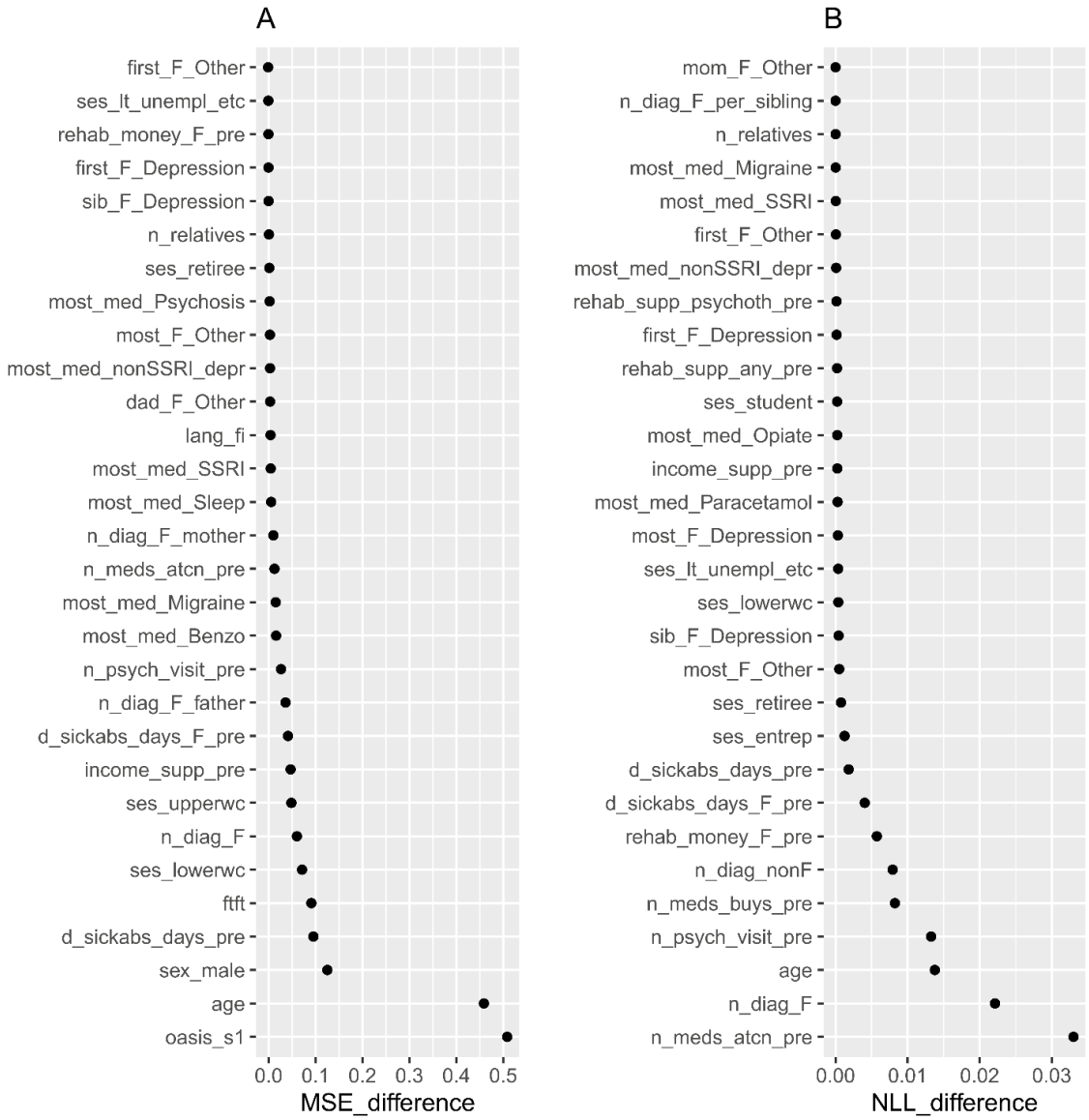
	Model 1 (initial)			Model 2 (simplified)		
Independent variable	**OR**	**95% CI**	**p**	**OR**	**95% CI**	**p**
Intercept	0.02	0.01–0.03	<0.001	0.02	0.01–0.03	<0.001
OASIS at 1st assessment	0.97	0.94–1	0.034	0.97	0.94–1.00	0.033
Age (years)	1.04	1.03–1.05	<0.001	1.04	1.03–1.05	<0.001
Sex (male)	0.84	0.65–1.08	0.171	0.85	0.66–1.09	0.196
Distinct purchased medicine (ATC-N) prescriptions (count)	0.95	0.91–0.99	0.010	0.95	0.92–0.98	0.001
Distinct past F diagnoses (count)	1.12	1.07–1.17	<0.001	1.12	1.07–1.17	<0.001
Distinct past non-F diagnoses (count)	1.04	1.03–1.05	<0.001	1.04	1.03–1.05	<0.001
Previous psychiatry visits (count)	1.00	1.00–1.01	0.001	1.00	1.00–1.01	0.002
Distinct ATC-N medicine purchase events (count)	1.00	1.00–1.00	0.888	–	–	–
Insurance-reimbursed sickness absence days with F diag. (count)	1.00	1.00–1.00	0.219	–	–	–
Insurance-reimbursed sickness absence days (count)	1.00	1.00–1.00	0.022	1.00	1.00–1.00	<0.001
Rehabilitation money receiver	6.75	3.18–14.30	<0.001	6.75	3.19–14.28	<0.001
Lower white-collar employee	0.53	0.40–0.7	<0.001	0.51	0.39–0.67	<0.001
Upper white-collar employee	0.65	0.47–0.9	0.009	0.62	0.45–0.84	0.002
Entrepreneur	1.45	0.91–2.29	0.117	–	–	–
Income-support receiver	0.73	0.54–0.99	0.045	0.71	0.53–0.96	0.027

Note: Values are odds ratios (ORs) and their 95% confidence intervals and p-values, with log(OR) being the original Logistic regression coefficient. “Model 1” uses covariates that attained “importance” in Supplementary Figure S1, whereas “Model 2” uses only a subset that were statistically significant Model 1 predictors for outcome, treatment assignment, or both (i.e., in Models 1 of Table S2 and/or S3).

Supplementary Table S5. Important variable descriptives. **Extended descriptive statistics for internet-delivered cognitive behavioral therapy (iCBT) and face-to-face cognitive behavioral therapy (fCBT) samples**

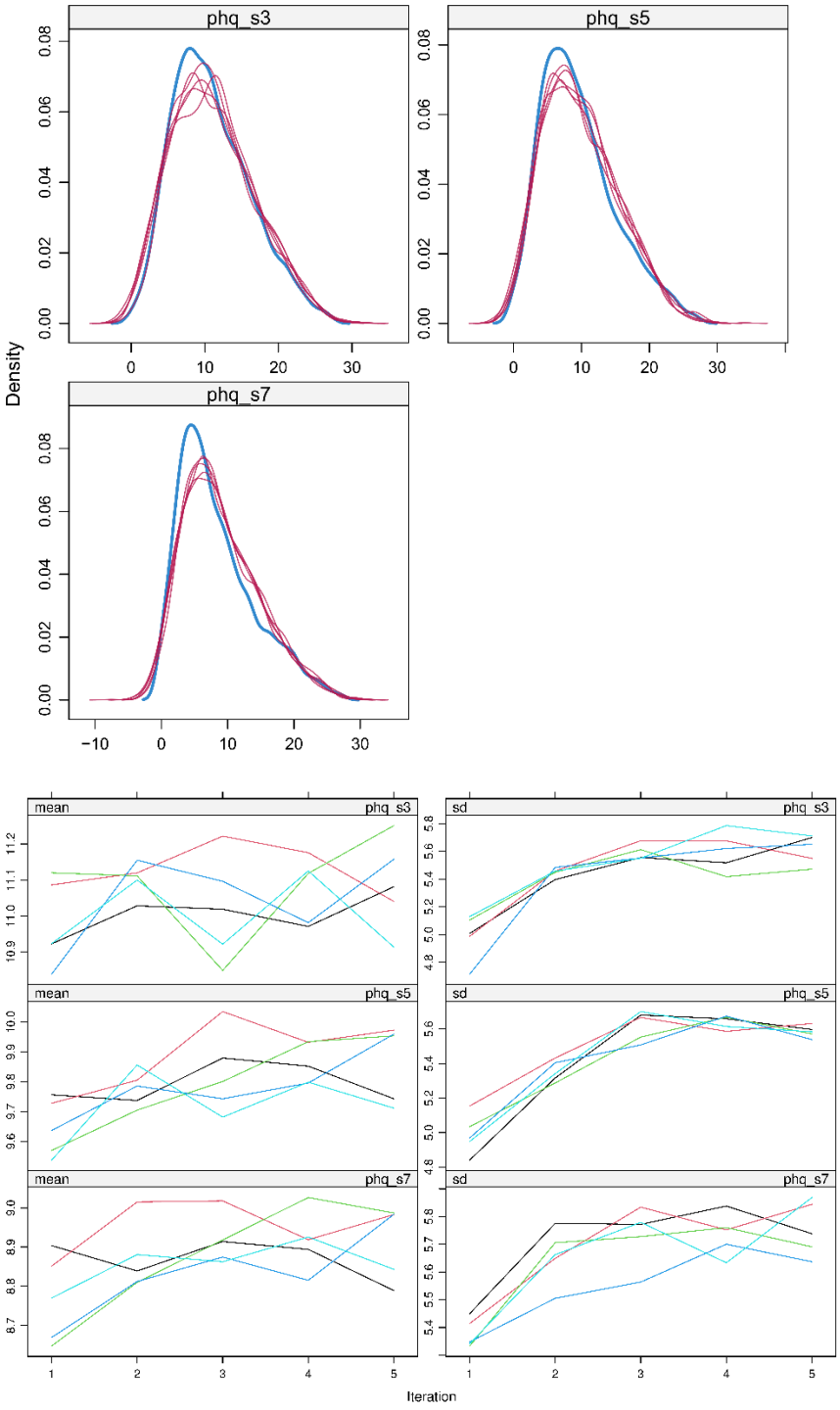
Variable	*M*	*SD*	*Mdn*	*skewness*	*kurtosis*
1. ΔPHQ-9 (symptom change score)	-3.58	5.02	-3	-0.44	0.50
2. OASIS at 1st assessment (symptom score)	10.36	3.68	11	-0.37	0.10
3. Distinct past F diagnoses (count)	1.74	2.61	1	2.96	12.97
4. Distinct past non-F diagnoses (count)	10.21	9.55	7	2.17	7.98
5. Number of relatives (count of parents and siblings)	4.08	1.91	4	2.10	9.17
6. Maternal F diagnoses (count)	0.76	2.29	0	5.71	45.90
7. Paternal F diagnoses (count)	0.62	2.10	0	5.95	49.45
8. Siblings' F diagnoses per sibling (count)	0.58	1.31	0	4.36	26.59
9. Distinct purchased medicine (ATC-N) prescriptions (count)	4.53	4.20	3	1.81	4.55
10. Distinct ATC-N medicine purchase events (count)	26.77	53.92	8	5.15	38.74
11. Insurance-reimbursed sickness absence days (count)	113.57	187.58	21	2.67	10.30
12. Insurance-reimbursed sickness absence days with F diag. (count)	71.17	126.89	0	2.49	7.84
13. Previous psychiatry visits (count)	16.29	41.45	1	5.53	45.09
14. Age (years)	34.49	11.89	31	0.83	-0.05
Variable	*M*	*SD*	*Mdn*	*skewness*	*kurtosis*
1. ΔPHQ-9 (symptom change score)	-2.46	4.80	-1	-0.27	1.09
2. OASIS at 1st assessment (symptom score)	9.97	3.81	10	-0.14	-0.22
3. Distinct past F diagnoses (count)	3.34	3.92	2	1.51	2.12
4. Distinct past non-F diagnoses (count)	15.57	12.55	12	1.52	2.52
5. Number of relatives (count of parents and siblings)	3.78	1.85	4	1.38	6.27
6. Maternal F diagnoses (count)	0.94	2.83	0	5.21	33.66
7. Paternal F diagnoses (count)	0.50	1.94	0	8.09	87.40
8. Siblings' F diagnoses per sibling (count)	0.50	1.28	0	6.12	49.57
9. Distinct purchased medicine (ATC-N) prescriptions (count)	5.77	5.52	4	1.13	0.93
10. Distinct ATC-N medicine purchase events (count)	45.00	83.54	12	4.37	30.62
11. Insurance-reimbursed sickness absence days (count)	135.64	196.74	23.5	1.65	2.40
12. Insurance-reimbursed sickness absence days with F diag. (count)	72.80	137.70	0	2.21	4.72
13. Previous psychiatry visits (count)	43.89	81.63	7.5	2.95	10.58
14. Age (years)	40.37	15.57	36	0.75	-0.29

Supplementary Figure S1. Variable importance. **Variable importance for the super learners for (A) treatment outcome and (B) treatment assignment.**



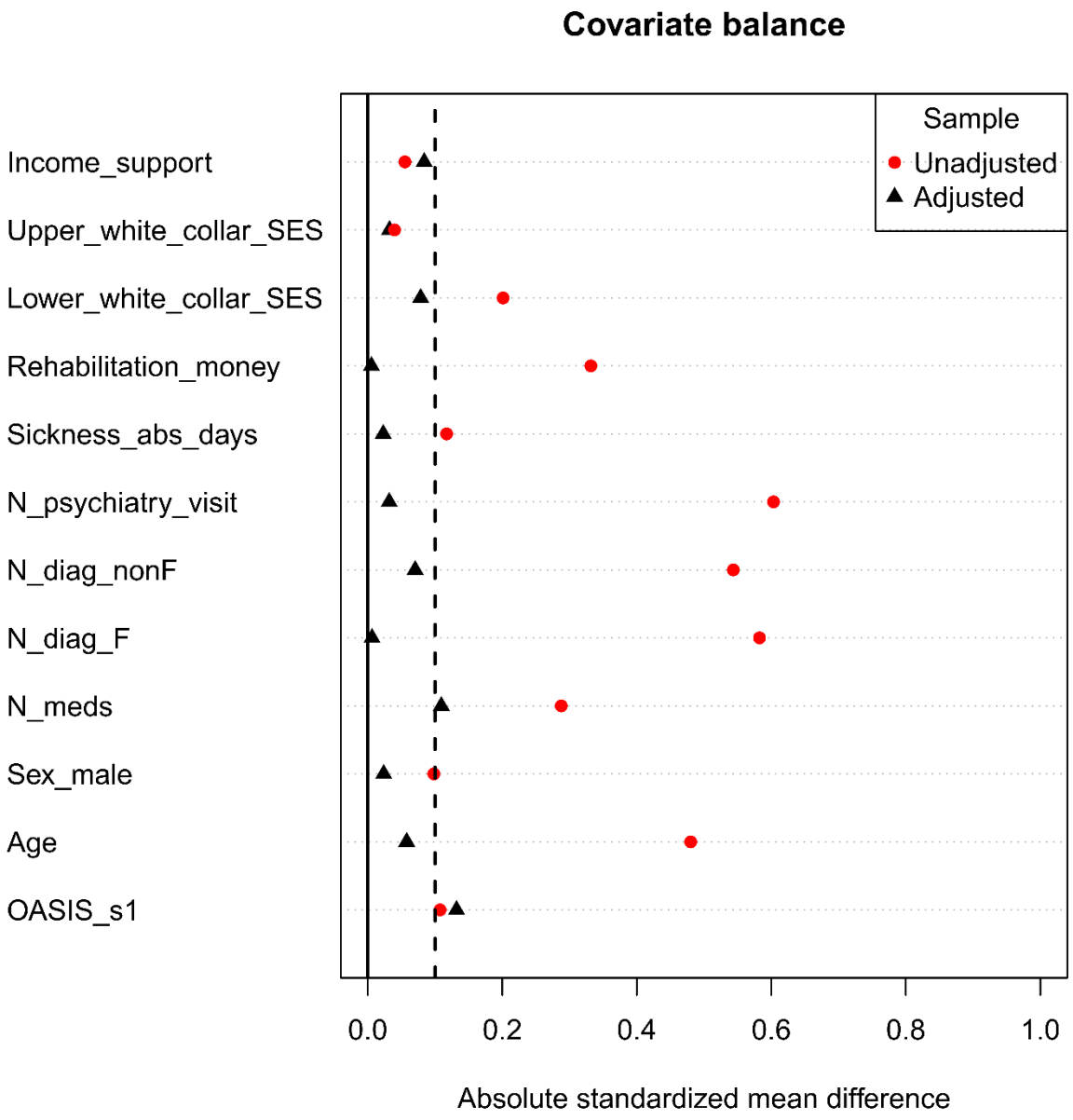
Supplementary Figure S1. Importance plot for treatment outcome (A) and assignment (B) predictors in the super learners. Outcome-prediction performance was measured by mean squared error (MSE) improvement relative to a model not including the variable, whereas (binary) treatment-assignment-prediction performance was measured by analogous negative log-likelihood (NLL) difference. The variable names are abbreviated from Table 1. Specifically, "not including" variable means here that its values were permuted to destroy any systematic relationship to the outcome or the other covariates. The values are differences in 10-fold cross-validated loss when including vs. not including the true variable. Note, baseline PHQ-9 was not included in the prediction models to avoid regression-to-mean bias (but see sensitivity analyses.

Supplementary Figure S2. Imputation diagnostics



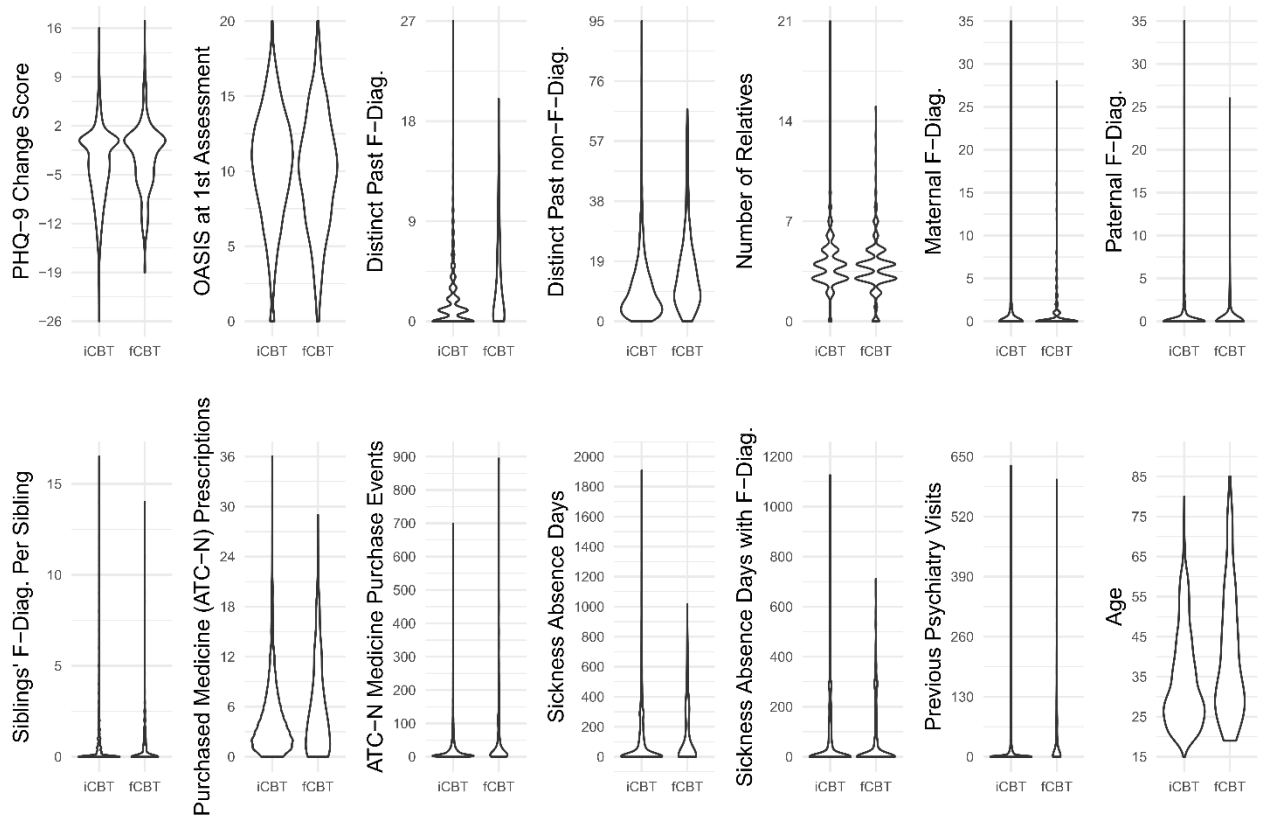
Supplementary Figure S2. Imputation diagnostics. Three upper panels show densities for observed values (blue lines) and for missing but imputed values (red lines) from the five independent imputation chains (several are needed to model uncertainty in imputed values). Different panels are for PHQ-9 at different sessions (s). Imputed values are in a sensible range but, as expected, the model suggests dropouts would have had more symptoms at the last session on average. The six lower panels show how imputed-value means and standard deviations (sd) per variable per imputation chain evolved by the iteration of the imputation algorithm. The chains mixed and converged properly. [note, imputing the procedure is computationally, which is why we used minimum recommendable chains and iterations]

Supplementary Figure S3. Covariate balance. **Covariate balance for classic inverse-probability weights**

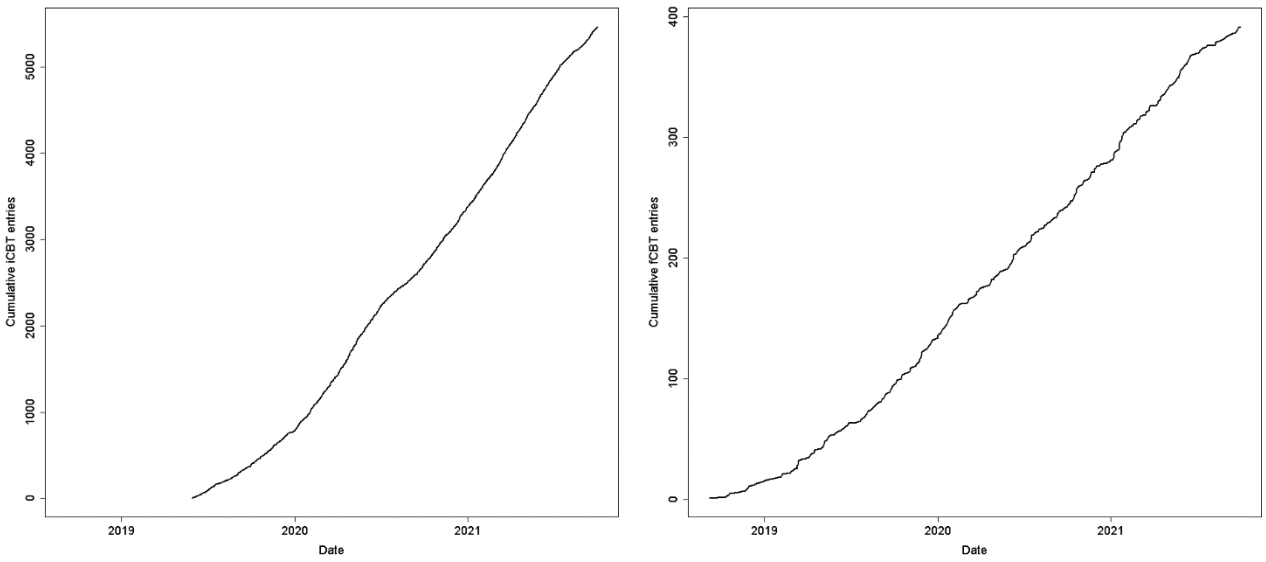


Supplementary Figure S3. Absolute standardized mean difference between fCBT and iCBT treatment groups in unadjusted and adjusted samples. Adjusting refers to inverse-probability weighting. Dashed line shows 10% of standard deviation for the respective variable. Differences < 0.1, or < 10%, indicate good covariate balance. Adjusted sample (black triangles) achieved a good balance and unadjusted sample (red discs) clear failed to achieve it. The variable names are abbreviated from Supplementary Table S3.

Supplementary Figure S4. Violin plots of important predictors. **Continuous-valued variables illustrated (see Table 1 for complete characterization of binary variables)**



Supplementary Figure S5. Time accumulation of included psychotherapies.



Supplementary Figure S5. Time-accumulation of initiated iCBTs (left) and fCBTs (right). Note, while the x-axes are comparable, the y-axes had to be given in different scales for the growth of the slower-growing fCBT data to be discernible. Furthermore, these figures do not represent the growth rates of the entire HUS-iCBT and FPQR registries but only the parts included in this study (see, e.g., ref 16 of the main text for a more general characterization of FPQR).

Supplementary R script on TMLE estimation

```
# example_script_TMLE_analysis.R
# Tom Rosenstrom, 2nd Sep 2024

library(tlverse)
library(sl3)
library(tmle3)
library(tmle3mopttx)
library(data.table)

# Effective sample size
n_eff <- min(c(5*nrow(dd)*mean(dd$ftft), nrow(dd)))

# Learners
lrn_mean <- Lrnr_mean$new()
lrn_glm <- Lrnr_glm_fast$new()
lrn_bayesglm <- Lrnr_bayesglm$new()
lrn_gam <- Lrnr_gam$new()
lrn_lasso <- Lrnr_glmnet$new(alpha = 1)
lrn_enet.5 <- Lrnr_glmnet$new(alpha = 0.5)
lrn_ridge <- Lrnr_glmnet$new(alpha = 0)
lrn_polymars <- Lrnr_polspline$new()
lrn_mars <- Lrnr_earth$new()
lrn_xgboost_autotune <- Lrnr_caret$new(algorithm = "xgbTree", metric = "RMSE")
lrn_bart <- Lrnr_dbarts$new(ndpost = 1000, verbose = FALSE)
lrn_hal <- Lrnr_hal9001$new(max_degree = 2, num_knots = 3)

# Stack learners to an ensemble
stack <- Stack$new(
  lrn_mean, lrn_glm, lrn_bayesglm, lrn_gam, lrn_lasso, lrn_enet.5, lrn_ridge,
  lrn_polymars, lrn_mars, lrn_xgboost_autotune, lrn_bart, lrn_hal
)

# Super learner
sl <- Lrnr_sl$new(learners = stack, metalearner = Lrnr_nnls$new())

### Treatment-group prediction with stratified cross-validation ###
task_g <- make_sl3_Task(data = dd[, c(4:ncol(dd))],
                        outcome = "ftft", # (ftft refers to fCBT)
                        outcome_type = "binomial",
                        folds = origami::make_folds(n=nrow(dd),
                                                    origami::folds_vfold,
                                                    strata_ids = dd$ftft),
                        covariates = names(dd)[c(4,6:ncol(dd))]) # drop group!

# Fit treatment-group prediction model
t0 <- proc.time()
sl_fit_g <- sl$train(task = task_g)
(runtime_sl_fit_g <- proc.time() - t0)

# Cross-validate the super learner and make a table describing the model
cv_sl_fit_g <- CV_lrnr_sl(lrnr_sl = sl_fit_g, task = task_g,eval_fun =
loss_loglik_binomial)
slres_g <- cbind(c(sl_fit_g$coefficients,NA), cv_sl_fit_g[,3:7])
colnames(slres_g)[1] <- "weights"
slres_g <- cbind(Learner = levels(cv_sl_fit_g$learner), round(slres_g, 3))
write.csv(slres_g, file = "to_results_export/Table_S1b.csv", quote = TRUE)
```

```
# Make a variable-importance plot
system.time({
  g_varimp <- sl3::importance(sl_fit_g, eval_fun = loss_loglik_binomial,
                              type = "permute")
})
importance_plot(x=g_varimp)

### Outcome prediction model with cross-validation ###
task <- make_sl3_Task(data = dd[, c(3:ncol(dd))],
                      outcome = "y",
                      folds = origami::folds_vfold(n=nrow(dd), V = 5),
                      covariates = names(dd)[4:ncol(dd)]) # incl. delivery!

# Fit outcome prediction model
t0 <- proc.time()
sl_fit <- sl$train(task = task)
(runtime_sl_fit <- proc.time() - t0) # 25 min
save(sl_fit, runtime_sl_fit, file = "sl_fit.Rdata")

# Cross-validate the super learner and make a table describing the model
cv_sl_fit <- CV_lrnr_sl(lrnr_sl = sl_fit, task = task,eval_fun =
loss_squared_error)
slres <- cbind(c(sl_fit$coefficients,NA), cv_sl_fit[,3:7])
colnames(slres)[1] <- "weights"
slres <- cbind(Learner = levels(cv_sl_fit$learner), round(slres, 3))
write.csv(slres, file = "to_results_export/Table_S1a.csv", quote = TRUE)

# Make a variable importance plot
system.time({
  varimp <- sl3::importance(sl_fit, eval_fun = loss_squared_error,
                            type = "permute")
})
importance_plot(x=varimp)

node_list <- list(W = names(dd)[c(4,6:ncol(dd))],
                  A = "ftft",
                  Y = "y")

# Define Q learner
Q_learner <- Lrnr_sl$new(
  learners = list(lrn_mean, lrn_glm, lrn_bayesglm, lrn_gam, lrn_lasso,
                  lrn_enet.5, lrn_ridge, lrn_polymars, lrn_mars,
                  lrn_xgboost_autotune, lrn_bart, lrn_hal),
  metalearner = Lrnr_nnls$new()
)

# Define g learner
g_learner <- Lrnr_sl$new(
  learners = list(lrn_mean, lrn_glm, lrn_bayesglm, lrn_gam, lrn_lasso,
                  lrn_enet.5, lrn_ridge, lrn_polymars, lrn_mars,
                  lrn_xgboost_autotune, lrn_bart, lrn_hal),
  metalearner = Lrnr_nnls$new()
)

# Compute the TMLE estimate
set.seed(48321)
ate_spec <- tmle_ATE(treatment_level = 1, control_level = 0)
ate_learner_list <- list(Y = Q_learner, A = g_learner)
system.time({
  tmle_fit <- tmle3(ate_spec, data = dd[,2:ncol(dd)], node_list = node_list,
learner_list = ate_learner_list)
```

```
})

# Check the target parameter estimate with confidence intervals
tmle_fit
```

Supplementary R script on alternative (sensitivity) analyses on ATE estimation

```
# example_script_sensitivity_analysis.R
# Tom Rosenstrom, 2nd Sep 2024
# Following Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME.
# Targeted maximum likelihood estimation for a binary treatment: A tutorial.
# Stat Med. 2018;37(16):2530-2546. doi:10.1002/sim.7628

# Functions for printing coefficients
fit_to_OR <- function(fit, rf = 2){
  s <- summary(fit)
  res <- matrix("", nrow = nrow(s$coefficients), ncol = 3)
  for (i in 1:nrow(res)){
    res[i, ] <- c(round(exp(s$coefficients[i,1]), rf),
                  paste0(round(exp(s$coefficients[i,1] -
1.96*s$coefficients[i,2]), rf),
                         "--",paste0(round(exp(s$coefficients[i,1] +
1.96*s$coefficients[i,2]), rf))),
                  round(s$coefficients[i,4], 3))
  }
  colnames(res) <- c("OR", "CI", "p")
  rownames(res) <- rownames(s$coefficients)
  res
}
fit_to_beta <- function(fit, rf = 2){
  s <- summary(fit)
  res <- matrix("", nrow = nrow(s$coefficients), ncol = 3)
  for (i in 1:nrow(res)){
    res[i, ] <- c(round(s$coefficients[i,1], rf),
                  paste0("(", round(s$coefficients[i,1] -
1.96*s$coefficients[i,2], rf),
                         ", ",round(s$coefficients[i,1] +
1.96*s$coefficients[i,2], rf), ")"),
                  round(s$coefficients[i,4], 3))
  }
  colnames(res) <- c("B", "CI", "p")
  rownames(res) <- rownames(s$coefficients)
  res
}


### After having reduced/simplified the models continue analogously and ###
# use only "important" variables from previous analyses...
varvect <- c("oasis_s1", "age", "sex_male", "n_meds_atcn_pre", "n_diag_F",
             "n_diag_nonF", "n_psych_visit_pre",
             "d_sickabs_days_pre", "rehab_money_F_pre",
             "ses_lowerwc", "ses_upperwc", "income_supp_pre")
# ... for propensity score (ftft refers to fCBT):
fA <- formula("ftft ~ oasis_s1 + age + sex_male + n_meds_atcn_pre + n_diag_F +
              n_diag_nonF + n_psych_visit_pre + d_sickabs_days_pre +
rehab_money_F_pre +
ses_lowerwc + ses_upperwc + income_supp_pre")
# ... for outcome:
fY <- formula("y ~ ftft + oasis_s1 + age + sex_male + n_meds_atcn_pre + n_diag_F
+
              n_diag_nonF + n_psych_visit_pre + d_sickabs_days_pre +
rehab_money_F_pre +
              ses_lowerwc + ses_upperwc + income_supp_pre")
```

```r
# outcome model
dtmp <- dtmp[,c("y","ftft",varvect)]
m <- lm(fY, data = dtmp)
res_betas <- fit_to_beta(m)

# treatment model (propensity score)
psm <- glm(fA, family = binomial, data = dtmp)
gW <- predict(psm, type = "response")
pROC::auc((gW>0.5)*1, dtmp$ftft)
res_ORs <- fit_to_OR(psm)
summary(gW)

# G-computation
QAW <- predict(m) # outcome prediction
Q1W <- predict(m, newdata = data.frame(ftft = 1, dtmp)) # counterfactual fCBT
Q0W <- predict(m, newdata = data.frame(ftft = 0, dtmp)) # counterfactual iCBT
mean(Q1W - Q0W) # G-computed ATE

# AIPTW estimate
EY1_aiptw <- mean(dtmp$ftft*(dtmp$y - Q1W)/gW + Q1W)
EY0_aiptw <- mean((1 - dtmp$ftft)*(dtmp$y - Q0W)/(1-gW) + Q0W)
( AIPTW_ATE <- EY1_aiptw - EY0_aiptw ) # AIPTW-ATE

# AIPTW estimate with 95% CIs
D1 <- dtmp$ftft*(dtmp$y - Q1W)/gW + Q1W - EY1_aiptw
D0 <- (1 - dtmp$ftft)*(dtmp$y - Q0W)/(1-gW) + Q0W - EY0_aiptw
varHat_AIPTW <- var(D1 - D0)/nrow(dtmp)
( ATE_aiptw_CI <- c(AIPTW_ATE - 1.96*sqrt(varHat_AIPTW), AIPTW_ATE +
1.96*sqrt(varHat_AIPTW)) )

# TMLE estimate
H1W <- dtmp$ftft/gW # clever covariates
H0W <- (1 - dtmp$ftft)/(1 - gW)
epsilon <- coef(lm(dtmp$y ~ -1 + H0W + H1W + QAW)) # perturbation
Q0W_1 <- Q0W + epsilon[1]/(1 - gW) # targeted counterfactuals
Q1W_1 <- Q1W + epsilon[2]/gW
( ATE_tmle <- mean(Q1W_1 - Q0W_1) ) # final estimate

# TMLE estimate's 95% CIs
D1 <- dtmp$ftft*(dtmp$y - Q1W_1)/gW + Q1W_1 - ATE_tmle
D0 <- (1 - dtmp$ftft)*(dtmp$y - Q0W_1)/(1 - gW) + Q0W_1 - ATE_tmle
EIC <- D1 - D0
varHat_IC <- var(EIC)/nrow(dtmp)
( ATE_tmle_CI <- c(ATE_tmle - 1.96*sqrt(varHat_IC), ATE_tmle +
1.96*sqrt(varHat_IC)) )
```

Supplementary R scripts on imputation and constructed variables

```
# Imputation (dd includes all covariates plus session-wise PHQ-9)
library(superMICE)
# Base-learner library (from superMICE package's options)
SL_lib <- c("SL.mean","SL.speedglm","SL.gam","SL.glmnet","SL.polymars",
            "SL.earth","SL.xgboost")
# Run it

system.time({

Imp_SL <- mice::mice(dd, maxit = 5, m = 5, method = "SuperLearner", print = T,
SL.library = SL_lib, kernel = "gaussian", bw = 0.5)

})



# A function to group the predictor F codes
fcode_to_disorder_predictor <- function(x){
  if (is.na(x)){return("None")}
  # Depressive episode, MDD, or dysthymia
  if
(grepl("F32",x)|grepl("F33",x)|grepl("F34.1",x)|grepl("F341",x)){return("Depress
ion")}
  # Social anxiety
  if (grepl("F40.1",x)|grepl("F401",x)){return("Social anx")}
  # Panic disorder
  if (grepl("F41.0",x)|grepl("F410",x)){return("Panic")}
  # Generalized anxiety
  if (grepl("F41.1",x)|grepl("F411",x)){return("GAD")}
  # Obsessive-compulsive disorder
  if (grepl("F42",x)){return("OCD")}
  # psychotic disorder
  if (grepl("F2",x)|grepl("F39",x)){return("Psychotic")}
  # bipolar disorder
  if
(grepl("F30",x)|grepl("F31",x)|grepl("F34.0",x)|grepl("F340",x)){return("Bipolar
")}
  # personality disorder
  if (grepl("F6",x)){return("Personality")}
  # other
  return("Other")
}

# A function to group the predictor F codes
atccode_to_meds_predictor <- function(x){
  if (is.na(x)){return("None")}
  # SSRIs
  if (grepl("N06AB",x)){return("SSRI")}
  # Other depression meds
  if (grepl("N06A",x)&(!grepl("N06AB",x))){return("nonSSRI_depr")}
  # Paracetamol
  if (grepl("N02BE01",x)){return("Paracetamol")}
  # Codeine opiate
  if (grepl("N02AA59",x)){return("Opiate_Codeine")}
  # Other opiate
  if (grepl("N02A",x)&(!grepl("N02AA59",x))){return("Opiate_other")}
  # Migraine
```

```
  if (grepl("N02C",x)){return("Migraine")}
  # Sleep meds
  if (grepl("N05C",x)){return("Sleep")}
  # Benzos
  if (grepl("N05BA",x)){return("Benzo")}
  # Other sedative neurosis
  if (grepl("N05B",x)&(!grepl("N05BA",x))){return("Sedative_nonBenzo")}
  # Psychosis
  if (grepl("N05A",x)){return("Psychosis")}
  # Epilepsy
  if (grepl("N03A",x)){return("Epilepsy")}
  return("Other")
}
```

Supplementary R scripts on full-analysis documentation (online repository)

Presumably only very rare occasions require examination of complete analysis code. For such rare cases, we provide a permanent online repository at

https://osf.io/m5uea/