The last half century of research has provided a wealth of evidence that cognitive behavior therapy is effective for depression and anxiety disorders (Cuijpers et al., 2016), and the last 20 years have shown that these treatments can be delivered as guided self-help via the internet, often called internet-delivered cognitive behavior therapy (ICBT; Carlbring et al., 2018; Cuijpers et al., 2010; Hedman et al., 2012; Karyotaki et al., 2017). Despite large group-level effects, not all patients get better. In fact, 5%–10% of patients both in psychological treatments in general and in ICBT specifically leave treatment worse off (Rozental et al., 2014; Slade et al., 2008) and, depending on how outcomes are defined, 25%–65% of patients do not achieve a satisfactory response. A recent set of individual patient data meta-analyses of 29 randomized controlled trials of ICBT found that a quarter of patients exhibit no reliably measurable change at all during treatment (Rozental et al., 2019), approximately half of patients respond positively to treatment, whereas only 35% reach remission (Andersson et al., 2019). On top of this, 6% actually get reliably worse (Rozental et al., 2017). This is consistent with earlier research in face-to-face psychotherapy (Lambert, 2015).

A high-level commission statement on the future of psychotherapy research (Holmes et al., 2018) calls for research to, among other things, use personalization of treatments that could possibly help in addressing the issue of treatments not working for everyone. One way to achieve this could be to identify patients at risk of not benefitting and target them specifically with personalized support and content. Predicting whether a patient will benefit from treatment based on intake data alone is very difficult (Andersson, 2018), but using early developments in treatment, for example, by monitoring patients' self-rated symptoms, has more potential (Forsell et al., 2020). Early identification of probable failing treatments has been shown to help prevent failures for those at-risk in both ICBT (Forsell et al., 2019) and traditional psychological treatment (Lambert, 2017; Shimokawa et al., 2010; Slade et al., 2008).

Patients, health care providers, and therapists themselves might be under the assumption that therapists have a good sense of whether a specific patient will improve. However, previous research suggests that therapists in traditional face-to-face psychotherapy are bad at such predictions (Hannan et al., 2005; Harmon et al., 2005; Lambert, 2015; Shimokawa et al., 2010; Slade et al., 2008; Whipple et al., 2003; White et al., 2015). Recently, various statistical and machine learning approaches outperformed psychotherapists in predicting outcome for face-to-face alcohol abstinence programs (Symons et al., 2020). This failure in therapist accuracy has been suggested to be driven by optimism, where therapists predict good outcomes among their own cases far more often than they occur (Lambert, 2017; Walfish et al., 2012).

However, one reason for therapists' difficulties in making realistic predictions could be that traditional psychological treatments are quite heterogeneous in how they are executed and that structured ways of monitoring progress are lacking. This could make it difficult for therapists to compare the progress of historical, and current, patients against each other. It has, for example, been shown that in traditional psychological treatment, measuring and showing the therapist the current status of the patient in a standardized way can help them personalize treatment (Lambert, 2017). This makes it relevant to explore therapists' predictive abilities when they work with more structured and standardized treatments and settings, as in the case of ICBT. In this treatment format, therapists very often have access to structured data on patients' progression, for instance via graphs based on continuous measures showing symptom change (Titov et al., 2018). ICBT therapists also have information on patient's treatment activity via text messages and homework reports, and the expected path of the treatment is highly standardized. These factors could make it easier to realize when patients deviate from typical progress patterns and enhance therapists' predictive accuracy. On the other hand, a lot of patient information is not available since ICBT-therapists rarely meet or speak to their clients. They might thus miss nonverbal cues, have difficulties determining the quality of the working alliance, and cannot use intense and direct interaction to get information about patients' state and progress. So far, no study has examined the predictive accuracy of ICBT-therapists.

The emergence of statistical prediction modeling and machine learning approaches for prediction in mental health care is a promising step toward realizing the potential of personalized and precision care. However, a criticism raised by data scientists is that models tend to be evaluated either only against chance or against iterations of the same model, and not against any clinically valid or relevant benchmark, making the achieved accuracy levels difficult to interpret and creating false optimism around machine learning approaches (DeMasi et al., 2017). To demonstrate clinical utility, a data-driven model should outperform the simplest and most readily available method for completing the same task (Scott et al., 2021). Assessing the predictive accuracy of therapists in a naturalistic setting should be the first such benchmark to compare to, when deciding if the creation and implementation of a complex computational model is warranted. It is so far unknown if therapists working with more standardized treatments and are provided structured data on patients' progress might increase their predictive accuracy enough to render statistical predictive models obsolete, or at least not very useful and costly accessories.

Some other aspects also remain largely unexplored. Prediction accuracy and therapists' optimism may, for example, vary depending on what disorder is being treated. The relevance of what, more specifically, the therapists are asked to predict is also not explored. For example, are they more skilled in predicting quantitative changes in symptom levels as measured by well-known scales, rather than more qualitative, intuitive, and general outcome categories like remission, response, or "being cured" (e.g., Salomonsson et al., 2019, asked only if the client would *improve*, without further specifying or operationalizing what that means). Furthermore, with the exception for predicting deterioration (Hannan et al., 2005; Lambert, 2017), no comparisons between the accuracy of therapists and standardized algorithms, or pure chance, have been made. Finally, little is known of therapists' confidence in their

Nils Hentati Isacsson played a supporting role in data curation, formal analysis, methodology, software, and validation and an equal role in writing–review and editing. Viktor Kaldo played a lead role in funding acquisition and supervision, a supporting role in project administration, writing–original draft, and writing–review and editing, and an equal role in conceptualization, formal analysis, and methodology.

Correspondence concerning this article should be addressed to Erik Forsell, Department of Clinical Neuroscience, Centre for Psychiatry Research, Karolinska Institutet, Kaldo, M58, Huddinge Hospital, 141 58 Huddinge, Sweden. Email: erik.forsell@ki.se

own predictions, and if that confidence relates to whether they are correct. A previous meta-analysis showed that confidence was a very weak indicator of whether clinicians were correct in their decision making, but that these things were a little bit better aligned when decisions were made with written materials (Miller et al., 2015).

Aim

The aim of the study was to examine the predictive accuracy and the confidence in predictions of ICBT-therapists when using methods that are feasible to use in regular care to predict treatment outcomes for their patients and to compare this to a statistical benchmark that uses linear regression with only patient symptom scores from the same timepoints in treatment. While this study is not a clinical trial, we do use several benchmarks to support the interpretability of the results and have some preliminary hypotheses. Specifically, based on previous research we will address the following:

- We hypothesize that ICBT-therapists make predictions of outcomes that are better than chance, but still below a benchmark based on a statistical model using weekly symptom data to predict outcome. This is expected regardless of whether therapists predict a qualitative, intuitive, and clinical categorical outcome or make a quantitative prediction of how many points the patient will change on the main symptom measure.
- We hypothesize that ICBT-therapists are optimistic in their predictions (i.e., they will predict positive categorical outcome more often, and larger overall symptom reduction, than occurs).
- We will explore if confidence in one's predictions differs between therapists, and if confidence is associated with correctness. We will also explore if therapists' confidence differs depending on what outcome the therapist thinks will occur for a patient (e.g., responder vs. deteriorater).

Method

This is a prospective study in regular care where ICBT-therapists' predictions of several key clinical outcomes for their patients, made during the fourth week of treatment, were compared to observed outcomes. The present study was registered with and approved by the regional ethical review board in Stockholm (2011/2091-31/3 with amendments 2016/21-32, 2017/2320-32, and 2018/2550-32). The analyses included in this report have not been preregistered.

Participants, Treatments, and Therapists

The participants included in this study were 867 patients who underwent ICBT for major depressive disorder (MDD, $n = 373$), social anxiety disorder (SAD, $n = 273$), or panic disorder (PD, $n = 251$) at the Internet Psychiatry Clinic in Stockholm, Sweden, between January 2017 and November 2018, were still in treatment after 4 weeks, and had their therapist complete the prediction form. Participants were self-referred and assessed live at the clinic by a psychiatrist or resident physician under psychiatrist supervision. All patients at the clinic during this time period were included. Patients are included in the regular care treatment if they are 18 years old, fulfill diagnostic criteria for the current diagnosis based on a clinical assessment using the Mini International Neuropsychiatric Interview, can read and write Swedish and have no direct contraindications for ICBT (i.e., any severe other psychiatric or somatic illness that needs to take precedence or are unwilling or unable to perform online treatment). The treatments were 12 weeks long, mainly text-based, and consisted of self-help materials, interactive worksheets and a personal therapist, providing feedback via a secure email-like system within the platform. Therapists were licensed psychologists or resident psychologists who work full time as ICBT therapists at the clinic. They have a minimum of 5 years of university education to become a clinical psychologist, followed by a 1-year residency program before getting their license. The typical therapist at the clinic during the time this study was completed had worked with ICBT for 5 years. Fourteen different therapists worked at the clinic during the study and provided predictions.

During the period, 1,058 patients were eligible for therapist prediction (i.e., started treatment at the clinic and were still in treatment 22 days). Out of these, the prediction was completed for 897 patients (85%) by 14 different therapists. Some patients did not complete posttreatment symptom measures and ultimately, 775 (86%) of patients with predictions had outcome data. Figure 1 describes the flow of patients during the period in more detail, divided by treatment, and Table 1 describes the sample characteristics such as demographic data and symptom scores.

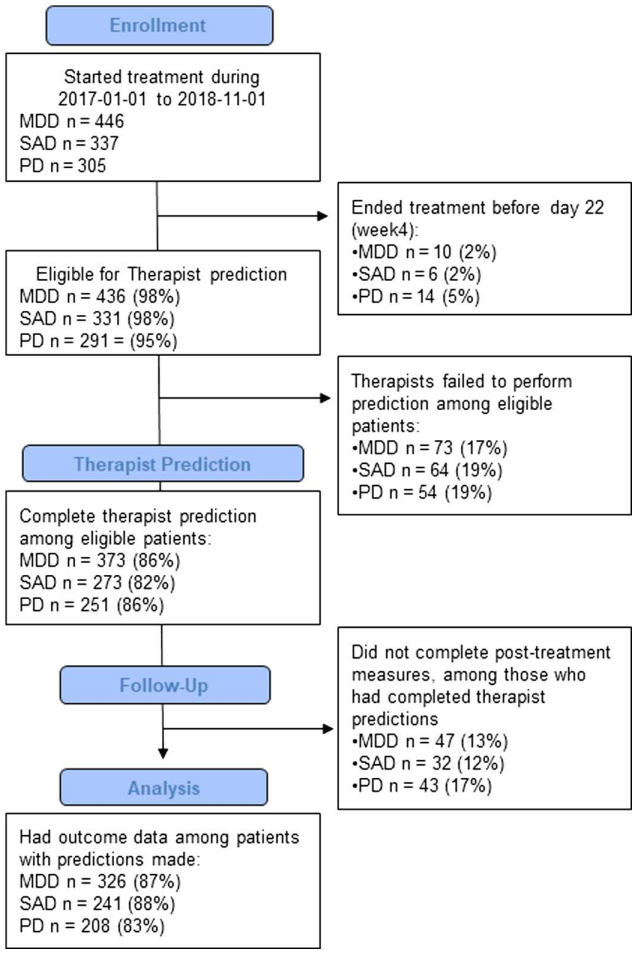
Clinician Predictions

Therapists were asked to make an assortment of predictions about each of their patients. The prediction was made for each patient in a clinician-rated questionnaire (see Appendix A), and was supposed to be completed sometime during the fourth week of treatment (Days 22–28) when the therapist also had other reasons to tend to the patient (e.g., replying to a message from the patient). The timing of the prediction was informed by two of our previous studies; Schibbye et al. (2014), where we found that Week 4 was preferable when predicting outcome in a similar sample from the clinic sample using only symptom scores, and Forsell et al. (2019), where we found that predicting in Week 4 and that intervening when failure was predicted could help avoid the failure. Before filling out the questionnaire, therapists were instructed to have a quick look at the pretreatment score of the patient as well as the graph within the platform showing their scores week by week. They were allowed to look at other things that they thought would be helpful to make the prediction but were instructed to spend a maximum of five extra minutes collecting this information, above the time they would spend to gather information needed to perform their regular clinical tasks, for example, reading a homework report and a worksheet before giving the patient feedback. These instructions were designed to approximate a feasible and not too time-consuming prediction task for therapists in a naturalistic setting, rather than to optimize the clinicians' predictive skills and accuracy.

Primary Outcome Measures for Patients

The RCI is calculated with sample variances as no external validation manuals with normative data exist for the scales.

Figure 1
Flowchart



Note. MDD = major depressive disorder; SAD = social anxiety disorder; PD = panic disorder. See the online article for the color version of this figure.

For MDD, the outcome measure was the Montgomery–Åsberg Depression Rating Scale–Self-report version (MADRS-S; Montgomery & Asberg, 1979; Svanborg & Asberg, 1994, 2001). MADRS-S is a widely used, unidimensional measure for depression designed to be sensitive to change. Scores range from 0 to 54 points. Test–retest reliability is high (intraclass correlation coefficient = .78; Fantino & Moore, 2009) and the standard deviation in the sample is relatively small, giving it a relatively narrow RCI of 8 or more points change. The cutoff for remission is a score of 10 or less at posttreatment. In the current sample, Cronbach's α was calculated to be .78 indicating good internal consistency.

For SAD, the outcome measure was the Leibowitz Social Anxiety Scale–Self-report version (Baker et al., 2002; Fresco et al., 2001). Leibowitz Social Anxiety Scale–Self-Report version is a widely used measure for social anxiety that includes not only subscales for anxiety and avoidance but also has a valid total score. Scores range from 0 to 144 points. Test–retest reliability is high (r = .83), but there is a lot of variation in the sample and the range of the measure is large, giving it an RCI of 28 or more points change. The cutoff for remission is a score of 34 or less at posttreatment (von Glischinski et al., 2018). In the current sample, Cronbach's α was calculated to be .93 indicating excellent internal consistency.

For PD, the outcome measure was the Panic Disorder Symptom Scale–Self-Report (PDSS-SR). PDSS-SR has seven items and scores range from 0 to 28 points. PDSS-SR is a widely used, unidimensional measure for PD symptoms. Test–retest reliability is high (intraclass correlation coefficient = .81; Houck et al., 2002), and the RCI was calculated to be 6 or more points. A score of 6 or less indicates subclinical symptoms on the clinician-rated version (Monkul et al., 2004) and will be used as the cutoff for remission, since no validated cutoffs exist for the patient-rated version. In the current sample, Cronbach's α was calculated to be .80 indicating good internal consistency.

Definitions of Categorical Outcomes and Therapists Qualitative and Quantitative Predictions

Since it is not known how the prediction task should be designed to optimize therapists' accuracy, we evaluated two types of predictions. A "qualitative prediction" where therapists read a short description of different outcomes and decided which best fit the patient. These are defined below. To make a more "quantitative prediction," therapists were asked to indicate exactly how many points on the main symptom outcome (MADRS-S, PDSS-SR, or Leibowitz Social Anxiety Scale–Self-Report version) they predicted the patient to change. Registering and monitoring raw scores from symptom rating scales is routine work at the clinic, so it was assumed that therapists were well versed in thinking about the scores on these measures. This change was then used, sometimes together with the observed pretreatment value, to calculate the predicted categorical outcome in the same way as the observed outcome was calculated (described below).

Remitter and "Being Cured"

Remitter was defined as having a posttreatment score on the primary symptom measure for the disorder being treated that is below the clinical cutoff for the disorder. Therapists' qualitative prediction was made by answering the following question:

> Will the patient be "cured" from his/her [depression, social phobia, panic disorder]—that is to say do you think that the patient's symptoms post treatment will be at a level comparable to a person without [depression, social phobia, panic disorder]?
>
> • Yes
>
> • No

Responder, Deteriorater, and "Meaningful Change"

Responder was measured in two ways commonly found in the literature. RCI (Andersson et al., 2019; Jacobson & Truax, 1991) was used to calculate a minimal score difference that is statistically significant. This RCI however does not guarantee a clinically meaningful difference and does not consider individual pretreatment symptoms where high pretreatment scores makes it easier to "respond" due to regression to the mean. Karin et al. (2018) argued that a 50% decrease in symptoms handles this issue better and also can be an indicator of a clinically meaningful reduction.

Table 1
Baseline Characteristics

Characteristic	MDD (n = 363)	SAD (n = 267)	PD (n = 237)
Age in years (*SD*)	37 (12)	31 (11)	34 (11)
Female	241 (66%)	164 (61%)	146 (62%)
Married or defacto	198 (57%)	127 (51%)	157 (68%)
With children	146 (42%)	59 (24%)	87 (38%)
Education			
Primary	15 (4%)	23 (9%)	25 (11%)
Secondary	156 (43%)	117 (44%)	126 (53%)
Postsecondary	179 (49%)	109 (41%)	81 (34%)
Current sick leave	26 (8%)	5 (2%)	13 (6%)
Mean pretreatment score on primary symptom measure (*SD*)	23.14 (6.32)	73.88 (21.84)	11.54 (4.27)
Mean posttreatment score on primary symptom measure (*SD*)	14.34 (8.46)	52.74 (24.28)	5.92 (4.78)
Pre–post effect size (Cohen's *d*)	1.18	0.92	1.24

Note. MDD = major depressive disorder; SAD = social anxiety disorder; PD = panic disorder.

Deterioration was defined using the RCI, but with an increase of .84 rather than 1.96 as the cutoff in accordance with Wise (2004), as it avoids underestimating the already very small group of patients who deteriorate. Patients neither being observed as a deteriorater or a responder are defined as nonresponders.

Therapists' quantitative prediction was determined by the following question:

> How do you think the patient will have changed in their [depression/social anxiety/panic disorder] measured with [measure] from pre-measurement to post-measurement?
>
> • Deteriorated in a clinically meaningful manner
>
> • No change so large that it can be considered clinically meaningful
>
> • Improved in a clinically meaningful manner

Evaluation and Comparison of Predictive Accuracy Against a Benchmark Regression Model

The most meaningful way to understand predictive accuracy is to compare it to benchmarks. The first and most obvious is to compare it to pure chance. However, this is usually not very helpful, as the clinical utility of a classification cannot be decided on simply being better than a coin flip, even if it is a minimal requirement. Another relevant benchmark one could use is a predictive model trained in the same context or sample. We have recently trained a predictive model based on linear regression using weekly patient symptom ratings to predict final symptom levels in a larger sample of patients from the same clinic (Forsell et al., 2020). In that study, weekly symptom measures (the same measures and timepoints as in this study) from 4,310 patients were used in a linear regression model to predict outcome and were evaluated on a holdout test set. This same model was used again in the present study to predict the continuous outcome for the current sample, which is a completely new test set for this already trained model. This prediction was made for the current sample in week 4 (i.e., the same week as the clinician's predictions in this study were made), making the predictions comparable. The continuous predictions are subsequently converted to categorical outcomes (remission and response) in order to perform analysis with a balanced accuracy for the classification task. This benchmark can show what a simple statistical model could potentially achieve at the same point in time in the same context and treatments.

Comparing pure chance or competing predictive strategies gives relevant information, but still does not determine if the level of accuracy is clinically useful. A recent study, we demonstrated the clinical usefulness of a predictive model used in an adaptive treatment strategy in ICBT for Insomnia (Forsell et al., 2019). The balanced accuracy of this model was 67% (Forsell et al., 2022) and although it was used to enhance ICBT for a different disorder to the ones studied here it can be used as a benchmark where it has been shown that the accuracy is high enough to be useful to clinicians.

Finally, we have identified a study where clinicians were asked to make clinical decisions based on various test results that had accompanying accuracies (Eisenberg & Hershey, 1983). They found that 65% accuracy was the key level where clinicians tended to choose to act on predictions, indicating a preliminary level of acceptance for clinicians.

Defining Therapist Confidence and Correctness

Therapists rated their confidence in the predictions for each patient, on a simple visual analog scale from 0% to 100% confident. The confidence rating made by the therapist pertains to all types of prediction (categorical and continuous) made for each patient. Comparing this directly to a balanced accuracy for each therapist would be preferable but is not feasible since many therapists made too few ratings. Instead, we created a simple correctness variable for the predictions in order to explore the relationship between confidence and correctness. Defining correctness will be done separately for the continuous outcomes and the categorical outcomes. For the continuous outcome, the absolute value of the difference between the predicted and the observed symptom change scores will be calculated and used as the correctness indicator. The outcome measures will be centered and standardized for these correctness analyses so that all treatment groups can be analyzed together, thereby increasing the sample size. For the categorical outcomes, correctness will be defined as predicted outcome = observed outcome. Therapists predict more than one categorical outcome, but

for this value, being correct on at least one of them counts as correct for that patient.

Missing Data

Missing data are considered those with a prediction made but without outcome data for comparison. Missing data were between 12% and 17% and no imputations were done, since more than 80% of true observed outcomes were available which is high for any clinical setting and statistical imputation could, if anything, artificially make relationships more linear since imputation is statistical predictions of outcomes. This is still an intent-to-treat analysis and not a complete analysis, as patients were included regardless of treatment adherence and patients who missed some measurements were also included if they had any measurement from the last 4 weeks of treatment. However, it is possible that early dropout is more likely among patients who are predicted to have poor outcomes by the therapists. This would potentially inflate how optimistic therapists seem if many of their negative predictions are followed by a dropout and therefore no comparator. To examine this, a sensitivity analysis will be done where dropout is equated with observed nonresponse.

Statistical Analysis

Categorical outcomes versus categorical predictions were analyzed using confusion matrix statistics. For categorical outcomes, balanced accuracy is the primary outcome in all cases but with deterioration, where we used the F1-score because deterioration is very rare (about 5%) and the F1-score is superior to balanced accuracy when classes are very uneven (i.e., predicting a very rare event). Balanced accuracy is a single value and combines true positives and true negatives and, as opposed to standard Accuracy, takes the base rate of the outcome into account. Accuracy is overused and often misleading as it does not take the base rate into account which means that, if classes are not perfectly even, an agnostic prediction model that always guesses the base rate of the most common outcome will have an accuracy that is equal to that base rate. Balanced accuracy solves this problem and is defined as balanced accuracy (BACC) = (true positives/positives + true negatives/negatives)/2. F1-score is defined as $2 \times ((\text{precision} \times \text{recall})/(\text{precision} + \text{recall}))$ where precision = true positives/positives (i.e., positive predictive value) and recall = true positives/condition positive (i.e., sensitivity). F1-score clearly emphasizes true positives and ignores true negatives altogether. Both measures range from 0 to 1 where .50 is completely random, and 1 is perfect classification. As there are many ways of quantifying classification accuracy and researchers and clinicians might want to compare to other studies using other metrics we will add Supplemental Files A and B for the qualitative and quantitative predictions respectively, that include several other commonly used such metrics (such as Sensitivity and Specificity, Positive and Negative predictive value).

For continuous outcome prediction, the observed change scores for each treatment will be compared to the therapist predicted change scores both in terms of their respective means with a simple *t*-test and in terms of their linear relationship using both Pearson's *r* and a simple regression, predicting observed change with therapist predicted change as the predictor. A sensitivity analysis will be done where the absolute difference between predicted and observed change scores are nested within therapists to account for variations in accuracy between therapists. This will be compared to a model without therapist nesting.

For assessing the associations between correctness and confidence for continuous outcomes, confidence will be correlated with the difference between predicted and observed change scores using Pearson's *r*. For categorical outcomes, confidence will be used to predict correctness using receiver operator characteristics curve analysis where a 95% confidence interval (CI) for the area under the curve that does not include .50 indicates a significant association.

Data and Study Materials Availability

Data used in this study are protected patient medical records and therefore not legal to share. Specific study materials such as the prediction questionnaire are included in the Supplemental Materials and are freely available for use for noncommercial purposes.

Results

Classification Accuracy of Therapist Predictions for Categorical Outcomes

Table 2 summarizes the predictions of categorical outcomes. Therapists predicted remission better than chance in all treatments. For the responder variables, therapists did better than chance in MDD and SAD but not in PD. Deterioration was never correctly identified, making an F1-score impossible to compute. A post hoc test examining the balanced accuracy for deterioration was no better than chance (overall 95% CI [.42, .56]). Only in two cases were the preliminary level of acceptance for clinicians of 65% met, and only in one of those (remitters in SAD) was the BACC significantly higher than 65%. The benchmark for clinical usefulness in an adaptive treatment strategy (67%) was only met for remission in SAD, though the CI overlaps the 67% benchmark also for remission in MDD. The benchmark with accuracies from the statistical model (linear regression) using weekly symptom measures to predict outcome (based on Forsell et al., 2019) indicated on average nine percentage points higher BACC for the statistical models in all cases, though often with overlapping CIs. BACC was statistically significantly higher for MDD for the responder and nonresponder outcomes but not for remission, and for the RCI-based responder outcome the statistical model statistically significantly outperformed the qualitative therapist predicted outcomes for PD and SAD as well.

Sensitivity analysis on prediction prevalence (i.e., how often certain outcomes were predicted) including all patients who dropped out did not indicate a large inflation of the prediction prevalence due to dropout (0%–2% differences on average and no differences on deterioration).

Associations Between Therapist-Predicted Change Scores and Observed Change Scores

Therapist-predicted change had a weak to moderate, but highly significant, positive correlation with observed change and could explain 13%–18% of the variation in observed change as presented in Table 3. This is further illustrated in Appendix B where scatterplots of predicted versus actual change scores are presented.

Table 2

Categorical Outcomes—Qualitative and Quantitative Therapist Predictions and a Benchmark Using a Statistical Prediction Model Trained on Data From the Same Treatments

Outcome	Observed proportion	Predicted proportion		BACC [95% CI]		Statistical prediction benchmark BACC [95% CI]
		Qual	Quant	Qual	Quant	
Remitter						
MDD	.37	.51	.44	.65 [.59, .70]*	.69 [.63, .74]*	.72 [.66, .77]*
PD	.71	.80	.90	.61 [.54, .67]*	.60 [.54, .67]*	.67 [.60, .75]*
SAD	.24	.26	.18	.73 [.67, .79]*	.71 [.65, .77]*	.76 [.69, .82]*
Responder (RCI)						
MDD	.54	.87	.77	.57 [.52, .63]*	.63 [.57, .68]*	.75 [.69, .80]*
PD	.52	.90	.71	.53 [.46, .60]	.65 [.58, .72]*	.71 [.65, .78]*
SAD	.33	.73	.24	.58 [.52, .65]*	.69 [.62, .75]*	.75 [.68, .81]*
Responder (50%)						
MDD	.32	.82	.50	.59 [.54, .64]*	.64 [.59, .69]*	.74 [.69, .79]*
PD	.49	.88	.84	.53 [.47, .59]	.59 [.52, .65]*	.63 [.56, .70]*
SAD	.17	.69	.13	.58 [.52, .64]*	.61 [.55, .67]*	.69 [.63, .76]*
Nonresponder						
MDD	.54	.18	.03	.60 [.55, .65]*	.52 [.47, .57]	.70 [.65, .76]*
PD	.37	.12	.03	.54 [.48, .61]	.53 [.47, .59]	.59 [.52, .66]*
SAD	.55	.31	.07	.62 [.56, .68]*	.53 [.47, .59]	.64 [.58, .71]*
Deteriorated					F1[a]	
MDD	.07	.01	<.00	NA	NA	NA
PD	.06	.02	<.00	NA	NA	NA
SAD	.02	.03	<.00	NA	NA	NA

Note. MDD = major depressive disorder; PD = panic disorder; SAD = social anxiety disorder; RCI = Reliable Change Index; Observed = prevalence in sample (proportion); Predicted = how often predicted by therapists (proportion); BACC = balanced accuracy; CI = confidence interval; Statistical Benchmark = using the model trained in Forsell et al. (2019) where outcomes were predicted using symptom scores and linear regression applied to this sample as a new test sample; Nonresponder = neither of the responder definitions and includes deterioraters as nonresponders; NA = not available.
[a] F1 is used instead of BACC for rare outcomes, but could not be calculated since there were zero true positives.
* $p < .05$ significantly better than chance.

To provide some insight to the potential effect of variations in prediction accuracy between therapists, a sensitivity analysis was done on the absolute differences between predicted and observed continuous outcome change on the primary symptom measure (reported in Table 3). Creating a hierarchical regression model where predictions are nested within therapists indicated that therapists did not differ in how accurate they were at predicting continuous outcomes.

Tendency of Therapists to Be Optimistic in Their Predictions

As can be seen in Table 2, therapists predicted positive categorical outcomes substantially more often than they occurred in all cases, with the exception of remission in SAD, where estimates are very similar. Table 4 reports the means of the predicted and observed change scores across therapies. The therapist predicted change scores were significantly larger than the observed change scores for MDD and PD but not for SAD, again indicating some degree of optimism.

Table 3

Therapist Versus Statistically Predicted Change as a Predictor for Observed Change

Treatment	Therapist predictions		Statistical benchmark	
	Pearson's r	Adjusted R^2	Pearson's r	Adjusted R^2
MDD	.37***	.13***	.67***	.45***
SAD	.43***	.18***	.72***	.51***
PD	.40***	.16***	.63***	.39***

Note. MDD = major depressive disorder; PD = panic disorder; SAD = social anxiety disorder.
*** $p < .001$.

Therapist's Confidence in Their Predictions and Its Association With Correctness

Figure 2 illustrates the differences in confidence in predictions across the 14 different therapists. There are statistically significant differences in how confident different therapists are in their predictions, as indicated in Figure 2 by the nonoverlapping interquartile ranges between many therapists. There were, however, no statistically significant differences between therapists in terms of mean correctness of categorical or continuous predictions. Figure 3 presents the average confidence therapists rated depending on which categorical outcome they were predicting; there were no differences in confidence depending on which of the categorical outcomes they predict will occur.

Higher therapist confidence was weakly related to correctness in continuous outcome, since confidence had a small but significant negative correlation with the absolute deviation between predicted and observed change score ($r = -.11$, $t = 3.008$, $df = 743$, $p = .003$), and was a significant predictor in the linear regression, explaining 1% of the variance in the deviation (adjusted $r^2 = .01$, $p = .02$). For the categorical outcome, there was a significant but very

Table 4
Mean Therapist-Predicted and Observed Change Scores

Treatment	Predicted change score *M* (*SD*)	Observed change score *M* (*SD*)	*t*
MDD (*n* = 316)	−10.9 (5.5)	−6.5 (7.4)	11.621*
SAD (*n* = 235)	−18.8 (13.9)	−19.3 (20.1)	0.311
PD (*n* = 194)	−7.9 (4.1)	−5.6 (4.8)	5.257*

Note. MDD = major depressive disorder; PD = panic disorder; SAD = social anxiety disorder.
* $p < .05$ for mean difference between observed and predicted.

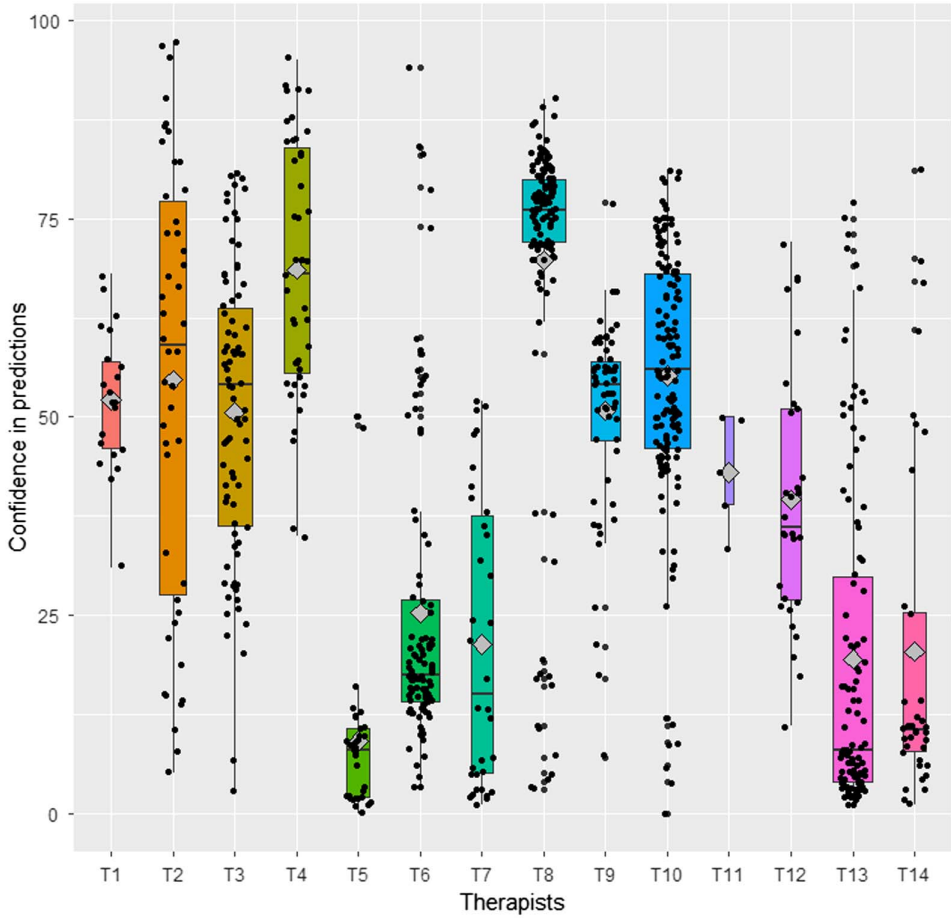
weak association between confidence and correctness according to the receiver operator characteristics curve analysis (area under the curve = .55, 95% CI [.51, .58]).

Discussion

We found overall support for the hypothesis that ICBT therapists make predictions that are better than chance, both when predicting quantitative change in symptom scores and qualitative categorical outcomes, although this was not always the case. PD seemed to be more difficult to predict, especially with the qualitative method, and deteriorators were never predicted correctly. Overall, therapists were less optimistic and more accurate when making the quantitative predictions as opposed to the qualitative ones.

We found that the overall balanced accuracies achieved by therapists were between 2 and 18 and 3 and 18 percentage

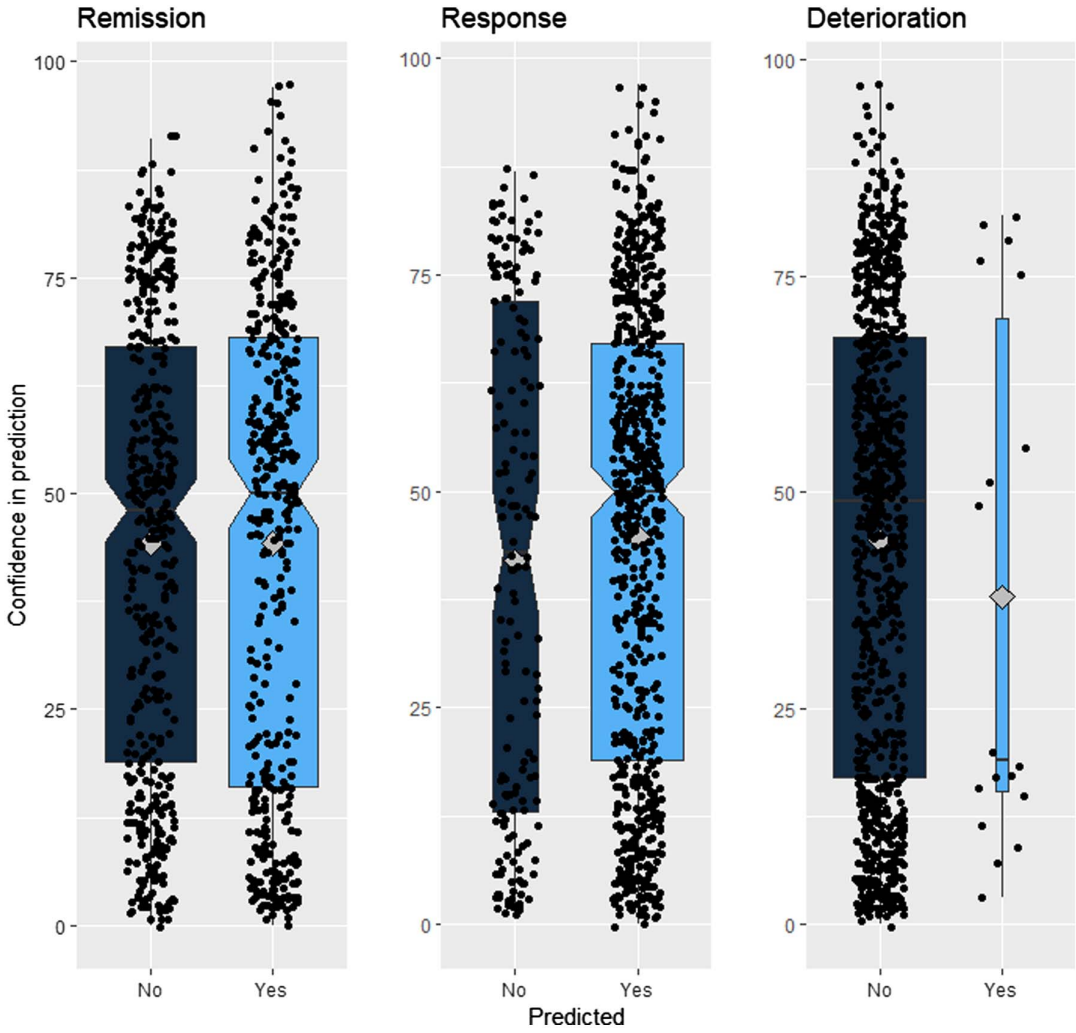
Figure 2
Confidence in Predictions Across Therapists



Note. The lower and upper hinges correspond to the 25th and 75th percentiles. The upper whisker extends from the hinge to the largest value no further than 1.5 times the interquartile range from the hinge. The horizontal lines correspond to the median and the gray diamonds corresponds to the mean. See the online article for the color version of this figure.

Figure 3

Confidence in Therapist Predictions Across Predicted Categorical Outcomes



Note. The lower and upper hinges correspond to the 25th and 75th percentiles. The upper whisker extends from the hinge to the largest value no further than 1.5 times the interquartile range from the hinge. The horizontal lines corresponds to the median and the gray diamonds corresponds to the mean. The notch corresponds to the 95% confidence interval around the median. See the online article for the color version of this figure.

points lower than the statistical benchmark for the qualitative and quantitative predictions, respectively. The smallest difference between the therapist predictions and the statistical benchmark was seen for the qualitative prediction of nonresponse in patients with Social Anxiety Disorder where the statistical benchmark was only two percentage points better. The largest difference was for the qualitative prediction of responders based on the RCI, where the statistical benchmark was 17–18 percentage points higher. The average advantage of the statistical benchmark compared to the qualitative predictions was 10.2 percentage points, whereas for the quantitative it was eight percentage points, indicating that therapists get closer if they predict the actual change scores on the relevant symptom measure rather than more qualitative clinical outcomes such as "improved" or "cured."

The fact that therapists performed better with the quantitative predictions could be for several reasons. One is that the therapists were then making predictions that are much closer to the definition of the ground truth (i.e., how we later defined the actual outcomes of the patients). It might also be that therapists, in this type of context, where symptom scores are collected weekly and presented as a graph, have a clearer picture of how the patient is doing in terms of those scores, but might think quite differently when asked if the treatment will cure the patient or significantly help them.

Of note is that the questions posed to the therapists do not correspond exactly to the computational definitions used to evaluate the outcome. We use definitions such as the RCI and the 50% reduction in symptoms, but that is not presented a priori to the therapist making the prediction. It is therefore entirely possible that a therapist faced with the observed outcome of a

patient would still disagree with our assessment of whether the therapist was right. Thus, there is a difference between finding that a prediction did not correspond with the outcome and that the therapists are wrong.

Similarly, the clinical outcomes themselves are imperfect operationalizations of a very complex phenomenon (i.e., being well, or improved, or ideally getting the kind of effect from the treatment that we can reasonably hope for). Psychotherapy outcome is a difficult thing to define, and any prediction effort is only as good as the definition of the outcome. We use definitions that are common in the field of treatment evaluation and use several different versions of these outcomes to minimize this problem. However, if blinded therapists had reviewed every patient and defined what the outcomes were (i.e., labeled the patients as for instance responders based on their holistic assessment of the treatment rather than only a symptom score) after the fact, perhaps that might have aligned better with the predictions, although there is no guarantee that it would.

The finding that statistical models outperform humans has been consistently reported in earlier studies (Ægisdóttir et al., 2006; Lambert, 2015, 2017). Some of the differences we found were relatively small, but it should also be taken into account that the statistical benchmark was a very basic statistical model that used only weekly symptom scores as predictors and does not represent the ceiling of what computational models can achieve. For example, a recent publication using data from the same clinic, ICBT-programs, and treatment week in various machine learning pipelines achieved balanced accuracies of about 74% for each separate treatment and 78% for a model where all treatments were combined (Kaldo et al., 2023) which itself is another 4–8 percentage points more than the benchmark used here. More advanced models have a great potential in using even more predictors, such as registry or genetics data (Boberg et al., 2023) to improve the accuracy even further. On the other hand, it might be possible to train therapists to make better predictions than in the present study where we prioritized feasibility in a naturalistic setting. Training therapists would be associated with a higher marginal cost for training and supervision and the extra time therapists would need to make the predictions. In contrast, a computational model is costly to create, but has little or no marginal cost once implemented, at least in a context such as this where all data are already being collected digitally and stored in an accessible database.

The very rare use of statistical prediction and monitoring of psychotherapy patients in routine care is problematic. Continuously measuring symptom or general distress levels could be implemented relatively easily and should be common practice, since it would not only allow for statistical predictions of outcome but also constitute a relevant quality assurance of psychological treatments in regular care.

When comparing predictions of continuous change scores directly to the continuous outcome measures, the statistical benchmark explained on average 45% of the variance (Schibbye et al., 2014), compared to 16% by the therapists. Explained variance in a regression cannot be directly compared to classification accuracy, but it is interesting to note that confidence intervals for the balanced accuracies mostly overlap with the statistical benchmark, whereas explained variance in this study was just over one third of what was achieved with the regression model (Schibbye et al., 2014).

The preliminary clinical acceptance criteria based on Eisenberg and Hershey (Eisenberg & Hershey, 1983), where the lower bound of the accuracy should be 65% or above, as well as the higher standard of 67% from our previous RCT showing a clinical benefit of a prediction within an adaptive treatment strategy (Forsell et al., 2019, 2022) was only met when predicting remission in Social Anxiety Disorder. This means that almost all therapist-made predictions were so uncertain that they themselves probably would be unwilling to act on them and that they probably would not be clinically beneficial if combined with an adaptive treatment strategy.

The predictive accuracy of the current ICBT-therapists cannot be compared to previous research because classification accuracy has not previously been reported. We can, however, compare the relative optimism of ICBT-therapists to face-to-face psychotherapists. Walfish et al. (2012) found that clinicians predicted positive outcomes on average 85% of the time. Compared to this, we find that ICBT-therapist were quite similar, predicting good outcomes 81% and 86% of the time in Depression and Panic Disorder, respectively, and a bit less optimistic in Social Anxiety Disorder where they predict good outcomes in 66% of cases. ICBT-therapists predicted positive categorical outcomes (remitter, responder, and reversed nonresponder) more than twice as often as they occurred, and mean predicted change was 1.68 and 1.41 times as large as observed change for Depression and Panic Disorder, though actually only .97 times as large as observed change for Social Anxiety Disorder.

This supports the hypothesis that therapists are generally optimistic about how their own patients will fare in treatment, though perhaps less so in social anxiety disorder, and is well in line with previous research from face-to-face psychotherapy (Hannan et al., 2005). It is important to note that it might not be a bad thing in and of itself that therapists are optimistic about the potential progress of their patients since this is likely to induce hope and motivation in both patient and therapist. Thus, rather than training therapists to be more pessimistic in their appraisal of the progression of a patient, as that could interfere with therapeutic processes, this might be an argument for using an objective and external prediction model of treatment outcome that indicates when a patient will need more help.

Another factor regarding therapist optimism is that these patients have all been included in treatment partly based on an assessment of their capability to benefit from the proposed ICBT. As such, all patients in the data were patients that were at least initially predicted to be able to benefit from treatment and hence an optimistic view of the outcome is quite reasonable.

Therapists clearly differed from each other in how confident they were in their predictions overall and how much their confidence levels varied between patients, but the relation between confidence and correctness, although significant, was very weak. In short, the finding does not support the notion that if therapists are confident, their prediction should be trusted more than if they are not confident. Their confidence also did not vary depending on whether they believed the patient would remit, respond, or deteriorate.

Therapist predictions (and the statistical models) were consistently better for social anxiety disorder, which also shows the least favorable outcomes. Through quarterly reports at the Internet Psychiatry Clinic, therapists are quite aware of this. Social anxiety disorder also has a lower natural recovery rate than depression and panic disorder (Bruce et al., 2005). This means that change in social anxiety might be more closely related to treatment activity and early progress, perhaps making prediction easier. In addition, regardless of if the smaller change in the Social anxiety treatment is an artifact

of a more change-insensitive measure or reflects actual less change, both therapists and statistical prediction might benefit from the lower change between pretreatment and posttreatment.

Related to this, it is interesting that therapists seem to be better in predicting remission than response. This could indicate that therapists are more aware of the criteria for, and have a more reasonable assessment of the likelihood of, remission compared to response. Even more clinically important, therapists very rarely predicted deterioration and, surprisingly, were never correct when they did. This is problematic as deterioration is arguably the most important outcome to detect early, as it is associated with the most harm for patients. However, also with statistical models, it is very difficult to predict rare events, and future research is needed to improve predictions of deterioration and other adverse events.

Even though therapists were less accurate than a statistical model, they mostly fared better than chance and predicted 13%–18% of the outcome variance. Therapist predictions could thus be a valuable input into a more complex predictive algorithm, at least if they are not highly correlated to other predictors. A recent study used clinician ratings not of outcome but of progress, adherence, and onboarding at the same point in treatment for patients undergoing ICBT for insomnia as a part of a more comprehensive classification algorithm that also used symptom scores and statistical calculations. They found that the clinician ratings, while not as strong by themselves as the computational part, did provide unique variance to the full model when combined (Forsell et al., 2022). The possibly increased predictive accuracy must be weighed against the resources needed to collect predictions from therapists. However, it could be worth the effort to let therapists make a quick guess at Week 4 in treatment and include this in a more complex predictive algorithm, for instance based on machine learning (Boman et al., 2019).

Considering generalizability, it is important to remember that the current predictions had no consequences for the therapists, other than taking some time to complete. If they had somehow been incentivized to be correct, or given instructions that were more precise on what to consider before making a prediction, they might have performed better. Also, if the prediction itself would have had immediate practical consequences, for example, if therapists had to offer extra support to patients predicted to have poor outcomes, which could have influenced their predictions. We did not aim to capture the maximum capacity of therapists' predictions in this study, but rather to assess how good their general, everyday ability to make such predictions is. Furthermore, therapists were asked to complete the predictions relatively quickly, not spending more than a few minutes gathering additional information beyond what they would need for managing the patient that day. We did this to ensure that the predictions were made in a way that was likely implementable (i.e., could conceivably be done as part of routine care in this real-world setting) as well as increasing the therapists adherence to the study and actually did the predictions they were supposed to do. This however limits the extent to which our findings could be viewed as a comparison between a therapist's maximum capacity to predict and a statistical model, and rather shows what a clinician could be assumed to achieve within routine care.

Another important aspect of generalizability is that the context of ICBT for these specific diagnoses in a specialized clinic could influence the accuracy of therapist predictions. The treatments and clinical routines themselves are highly structured and therapists are continuously presented with structured symptom ratings and activity data from patients. They also have many patients undergoing exactly the same treatment at any one time to compare with. This could make ICBT-therapists extra good at predictions. On the other hand, the therapists do not meet, converse with or perhaps know the patients quite as intimately as face-to-face therapists and the similarities between patients and treatments could conversely blind the therapists, making them poorer than traditional therapists. Our findings that these therapists are better than chance might not mean that therapists in a more traditional face-to-face setting with a more eclectic mix of ongoing treatments would be as good or bad at prediction.

Limitations

While sensitivity analysis shows that missing outcome data did not meaningfully affect the accuracy of made predictions, some patients (18% of eligible patients) were excluded due to the therapist failing to make and log the prediction. When asked, therapists stated that the prediction was not systematically avoided depending on the patient, but that the overall time constraints at the time a prediction was to be made dictated if a prediction was skipped. However, it is still possible that the omission of the prediction was biased by such things as the therapist's belief in the likelihood of success or failure.

The sample size divided by the large number of therapists, all of whom treated a different number of patients during the period, prohibits in-depth analysis of the accuracy of each therapist. Future research should examine what, if anything, makes a therapist good or bad at predicting outcomes. Therapists may look at different things and value them differently. It could also be that some therapists have their own definitions of outcomes that align with research definitions.

Finally, therapists rated their overall confidence in all different types of predictions they made for a patient, which makes it difficult to know if the therapist was more confident in some predictions than in others. Having the therapist rate their confidence in each subquestion would remedy this but doubles the number of questions they must answer.

Conclusions

Therapists can often, but not always, predict treatment outcomes better than chance in ICBT for depression and anxiety, though generally not as well as statistical models using weekly patient-rated symptom data, and probably not accurately enough that they themselves would be willing to act on the predictions, or that the predictions would be clinically useful in an adaptive treatment strategy. They differ in how confident they are, but confidence does not relate strongly to being correct. Our previous findings suggest that patients would benefit if statistical monitoring and prediction tools were used routinely in clinical settings. Future research is needed to examine if therapist predictions could be incorporated as input into machine learning models and provide unique variance for predictions, and if such efforts would be worthwhile and improve predictive algorithms enough to justify spending clinician time on making predictions rather than only using data that are already being collected.

References

Andersson, G. (2018). Internet interventions: Past, present and future. *Internet Interventions*, *12*, 181–188. https://doi.org/10.1016/j.invent.2018.03.008

Andersson, G., Carlbring, P., & Rozental, A. (2019). Response and remission rates in internet-based cognitive behavior therapy: An individual patient data meta-analysis. *Frontiers in Psychiatry*, *10*(13), Article 749. https://doi.org/10.3389/fpsyt.2019.00749

Baker, S. L., Heinrichs, N., Kim, H.-J., & Hofmann, S. G. (2002). The Liebowitz social anxiety scale as a self-report instrument: A preliminary psychometric analysis. *Behaviour Research and Therapy*, *40*(6), 701–715. https://doi.org/10.1016/S0005-7967(01)00060-2

Boberg, J., Kaldo, V., Mataix-Cols, D., Crowley, J. J., Roelstraete, B., Halvorsen, M., Forsell, E., Isacsson, N. H., Sullivan, P. F., Svanborg, C., Andersson, E. H., Lindefors, N., Kravchenko, O., Mattheisen, M., Danielsdottir, H. B., Ivanova, E., Boman, M., Fernández de la Cruz, L., Wallert, J., … Rück, C. (2023). Swedish multimodal cohort of patients with anxiety or depression treated with internet-delivered psychotherapy (MULTI-PSYCH). *BMJ Open*, *13*(10), Article e069427. https://doi.org/10.1136/bmjopen-2022-069427

Boman, M., Ben Abdesslem, F., Forsell, E., Gillblad, D., Görnerup, O., Isacsson, N., Sahlgren, M., & Kaldo, V. (2019). Learning machines in Internet-delivered psychological treatment. *Progress in Artificial Intelligence*, *8*, 475–485. https://doi.org/10.1007/s13748-019-00192-0

Bruce, S. E., Yonkers, K. A., Otto, M. W., Eisen, J. L., Weisberg, R. B., Pagano, M., Shea, M. T., & Keller, M. B. (2005). Influence of psychiatric comorbidity on recovery and recurrence in generalized anxiety disorder, social phobia, and panic disorder: A 12-year prospective study. *The American Journal of Psychiatry*, *162*(6), 1179–1187. https://doi.org/10.1176/appi.ajp.162.6.1179

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

Cuijpers, P., Cristea, I. A., Karyotaki, E., Reijnders, M., & Huibers, M. J. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry*, *15*(3), 245–258. https://doi.org/10.1002/wps.20346

Cuijpers, P., Donker, T., van Straten, A., Li, J., & Andersson, G. (2010). Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. *Psychological Medicine*, *40*(12), 1943–1957. https://doi.org/10.1017/S0033291710000772

DeMasi, O., Kording, K., & Recht, B. (2017). Meaningless comparisons lead to false optimism in medical machine learning. *PLOS ONE*, *12*(9), Article e0184604. https://doi.org/10.1371/journal.pone.0184604

Eisenberg, J. M., & Hershey, J. C. (1983). Derived thresholds. Determining the diagnostic probabilities at which clinicians initiate testing and treatment. *Medical Decision Making*, *3*(2), 155–168. https://doi.org/10.1177/0272989X8300300203

Fantino, B., & Moore, N. (2009). The self-reported Montgomery-Åsberg depression rating scale is a useful evaluative tool in major depressive disorder. *BMC Psychiatry*, *9*(1), Article 26. https://doi.org/10.1186/1471-244X-9-26

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

Forsell, E., Jernelöv, S., Blom, K., & Kaldo, V. (2022). Clinically sufficient classification accuracy and key predictors of treatment failure in a randomized controlled trial of Internet-delivered Cognitive Behavior Therapy for Insomnia. *Internet Interventions*, *29*, Article 100554. https://doi.org/10.1016/j.invent.2022.100554

Forsell, E., Jernelöv, S., Blom, K., Kraepelien, M., Svanborg, C., Andersson, G., Lindefors, N., & Kaldo, V. (2019). Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: A single-blind randomized clinical trial with insomnia patients. *The American Journal of Psychiatry*, *176*(4), 315–323. https://doi.org/10.1176/appi.ajp.2018.18060699

Fresco, D. M., Coles, M. E., Heimberg, R. G., Liebowitz, M. R., Hami, S., Stein, M. B., & Goetz, D. (2001). The Liebowitz Social Anxiety Scale: A comparison of the psychometric properties of self-report and clinician-administered formats. *Psychological Medicine*, *31*(6), 1025–1035. https://doi.org/10.1017/S0033291701004056

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

Harmon, C., Hawkins, E. J., Lambert, M. J., Slade, K., & Whipple, J. S. (2005). Improving outcomes for poorly responding clients: The use of clinical support tools and feedback to clients. *Journal of Clinical Psychology*, *61*(2), 175–185. https://doi.org/10.1002/jclp.20109

Hedman, E., Ljótsson, B., & Lindefors, N. (2012). Cognitive behavior therapy via the Internet: A systematic review of applications, clinical efficacy and cost-effectiveness. *Expert Review of Pharmacoeconomics & Outcomes Research*, *12*(6), 745–764. https://doi.org/10.1586/erp.12.67

Holmes, E. A., Ghaderi, A., Harmer, C. J., Ramchandani, P. G., Cuijpers, P., Morrison, A. P., Roiser, J. P., Bockting, C. L. H., O'Connor, R. C., Shafran, R., Moulds, M. L., & Craske, M. G. (2018). The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. *The Lancet Psychiatry*, *5*(3), 237–286. https://doi.org/10.1016/S2215-0366(17)30513-8

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

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

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

Karin, E., Dear, B. F., Heller, G. Z., Gandy, M., & Titov, N. (2018). Measurement of symptom change following web-based psychotherapy: Statistical characteristics and analytical methods for measuring and interpreting change. *JMIR Mental Health*, *5*(3), Article e10200. https://doi.org/10.2196/10200

Karyotaki, E., Riper, H., Twisk, J., Hoogendoorn, A., Kleiboer, A., Mira, A., Mackinnon, A., Meyer, B., Botella, C., Littlewood, E., Andersson, G., Christensen, H., Klein, J. P., Schröder, J., Bretón-López, J., Scheider, J., Griffiths, K., Farrer, L., Huibers, M. J., … Cuijpers, P. (2017). Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: A meta-analysis of individual participant data. *JAMA Psychiatry*, *74*(4), 351–359. https://doi.org/10.1001/jamapsychiatry.2017.0044

Lambert, M. J. (2015). Progress feedback and the OQ-system: The past and the future. *Psychotherapy*, *52*(4), 381–390. https://doi.org/10.1037/pst0000027

Lambert, M. J. (2017). Maximizing psychotherapy outcome beyond evidence-based medicine. *Psychotherapy and Psychosomatics*, *86*(2), 80–89. https://doi.org/10.1159/000455170

Miller, D. J., Spengler, E. S., & Spengler, P. M. (2015). A meta-analysis of confidence and judgment accuracy in clinical decision making. *Journal*

of Counseling Psychology, *62*(4), 553–567. https://doi.org/10.1037/cou0000105

Monkul, E. S., Tural, U., Onur, E., Fidaner, H., Alkin, T., & Shear, M. K. (2004). Panic Disorder Severity Scale: Reliability and validity of the Turkish version. *Depression and Anxiety*, *20*(1), 8–16. https://doi.org/10.1002/da.20011

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*, *134*(4), 382–389. https://doi.org/10.1192/bjp.134.4.382

Rozental, A., Andersson, G., Boettcher, J., Ebert, D. D., Cuijpers, P., Knaevelsrud, C., Ljotsson, B., Kaldo, V., Titov, N., & Carlbring, P. (2014). Consensus statement on defining and measuring negative effects of Internet interventions. *Internet Interventions: The Application of Information Technology in Mental and Behavioural Health*, *1*(1), 12–19. https://doi.org/10.1016/j.invent.2014.02.001

Rozental, A., Andersson, G., & Carlbring, P. (2019). In the absence of effects: An individual patient data meta-analysis of non-response and its predictors in internet-based cognitive behavior therapy. *Frontiers in Psychology*, *10*(15), Article 589. https://doi.org/10.3389/fpsyg.2019.00589

Rozental, A., Magnusson, K., Boettcher, J., Andersson, G., & Carlbring, P. (2017). For better or worse: An individual patient data meta-analysis of deterioration among participants receiving Internet-based cognitive behavior therapy. *Journal of Consulting and Clinical Psychology*, *85*(2), 160–177. https://doi.org/10.1037/ccp0000158

Salomonsson, S., Santoft, F., Lindsäter, E., Ejeby, K., Ingvar, M., Öst, L.-G., Lekander, M., Ljótsson, B., & Hedman-Lagerlöf, E. (2019). Predictors of outcome in guided self-help cognitive behavioural therapy for common mental disorders in primary care. *Cognitive Behaviour Therapy*, *49*(6), 455–474. https://doi.org/10.1080/16506073.2019.1669701

Schibbye, P., Ghaderi, A., Ljótsson, B., Hedman, E., Lindefors, N., Rück, C., & Kaldo, V. (2014). Using early change to predict outcome in cognitive behaviour therapy: Exploring timeframe, calculation method, and differences of disorder-specific versus general measures. *PLOS ONE*, *9*(6), Article e100614. https://doi.org/10.1371/journal.pone.0100614

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

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology*, *78*(3), 298–311. https://doi.org/10.1037/a0019247

Slade, K., Lambert, M. J., Harmon, S. C., Smart, D. W., & Bailey, R. (2008). Improving psychotherapy outcome: The use of immediate electronic feedback and revised clinical support tools. *Clinical Psychology & Psychotherapy*, *15*(5), 287–303. https://doi.org/10.1002/cpp.594

Svanborg, P., & Asberg, M. (1994). A new self-rating scale for depression and anxiety states based on the Comprehensive Psychopathological Rating Scale. *Acta Psychiatrica Scandinavica*, *89*(1), 21–28. https://doi.org/10.1111/j.1600-0447.1994.tb01480.x

Svanborg, P., & Asberg, M. (2001). A comparison between the Beck Depression Inventory (BDI) and the self-rating version of the Montgomery Asberg Depression Rating Scale (MADRS). *Journal of Affective Disorders*, *64*(2–3), 203–216. https://doi.org/10.1016/S0165-0327(00)00242-1

Symons, M., Feeney, G. F. X., Gallagher, M. R., Young, R. M., & Connor, J. P. (2020). Predicting alcohol dependence treatment outcomes: A prospective comparative study of clinical psychologists versus 'trained' machine learning models. *Addiction*, *115*(11), 2164–2175. https://doi.org/10.1111/add.15038

Titov, N., Dear, B., Nielssen, O., Staples, L., Hadjistavropoulos, H., Nugent, M., Adlam, K., Nordgreen, T., Bruvik, K. H., Hovland, A., Repål, A., Mathiasen, K., Kraepelien, M., Blom, K., Svanborg, C., Lindefors, N., & Kaldo, V. (2018). ICBT in routine care: A descriptive analysis of successful clinics in five countries. *Internet Interventions*, *13*, 108–115. https://doi.org/10.1016/j.invent.2018.07.006

von Glischinski, M., Willutzki, U., Stangier, U., Hiller, W., Hoyer, J., Leibing, E., Leichsenring, F., & Hirschfeld, G. (2018). Liebowitz Social Anxiety Scale (LSAS): Optimal cut points for remission and response in a German sample. *Clinical Psychology & Psychotherapy*, *25*(3), 465–473. https://doi.org/10.1002/cpp.2179

Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M. J. (2012). An investigation of self-assessment bias in mental health providers. *Psychological Reports*, *110*(2), 639–644. https://doi.org/10.2466/02.07.17.PR0.110.2.639-644

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment and problem-solving strategies in routine practice. *Journal of Counseling Psychology*, *50*(1), 59–68. https://doi.org/10.1037/0022-0167.50.1.59

White, M. M., Lambert, M. J., Ogles, B. M., Mclaughlin, S. B., Bailey, R. J., & Tingey, K. M. (2015). Using the assessment for signal clients as a feedback tool for reducing treatment failure. *Psychotherapy Research*, *25*(6), 724–734. https://doi.org/10.1080/10503307.2015.1009862

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment*, *82*(1), 50–59. https://doi.org/10.1207/s15327752jpa8201_10

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

(*Appendices follow*)

Appendix A

The Prediction Questionnaire

Clinician Prediction of Outcome

(This has been translated from Swedish, corresponding versions exist for social anxiety disorder and panic disorder)

The purpose of these questions is to see how well you as a therapist can guess the primary treatment outcome for this patient. This means that you only assess outcomes regarding depression symptoms (e.g., ignore level of functioning or quality of life in this assessment).

Answer during the fourth week of treatment, that is, Treatment Day 22–28. Before answering the questions, always take a look at the patient's points on MADRS-S and on the patient's change curve.

You may also weigh in other information about the patient in your guess. If you do, describe these factors during the last question, but do not spend more than 5 min to look for information.

1. How do you think the patient will have changed in their depression (measured with MADRS-S) from premeasurement to postmeasurement?

   Deteriorated in a clinically meaningful manner

   No change so large that it can be considered clinically meaningful

   Improved in a clinically meaningful manner

2. Will the patient become "cured" from his depression? That is to say; do you think that the patient's discomfort at end of treatment will be at a level comparable to a person without depression?

   Yes

   No

3. Exactly how many points on MADRS-S do you guess the patient will be improved or changed from premeasurement to postmeasurement?

   Improved, Number of Points:

   Deteriorated, number of points:

   When the treatment is over, a clinician will estimate the patient on Clinical Global Impression–Severity and Clinical Global Impression–Improvement. We now want you to guess how these estimates will look.

4. Clinical Global Impression–Severity: Global severity. Taking into consideration your clinical experience in this patient population, how ill are you guessing that this patient will be when the treatment is over? Only judge on depression symptoms.

   1. No disease
   2. Slight disease, doubtful, transient, no disability
   3. Mild symptoms, minor impairment
   4. Moderate symptoms work with effort
   5. Moderate–serious symptoms, limited function
   6. Serious symptoms, mostly work with the help of others
   7. Extremely serious symptoms, completely inoperative

5. Clinical Global Impression–Improvement: Global improvement. Guess the total improvement from the initial estimate to the end of treatment, regardless of whether, according to your assessment, it depends on treatment or not. Only judge depression symptoms.

   1. Very much improved
   2. Much improved
   3. Minimally improved
   4. No change
   5. Minimal deterioration
   6. Much worse
   7. Very much deteriorated

6. How confident are your guesses 1–5? Rate between 0% and 100%

   0%

   100%

7. Before you made your guesses, did you weight your decision of anything more than just the value and the change curve on MADRS-S? If so, describe what more you based guesses on!

   Have just looked at the starting value and the change curve before the guess

   Have also done or looked at the following:

   Submit

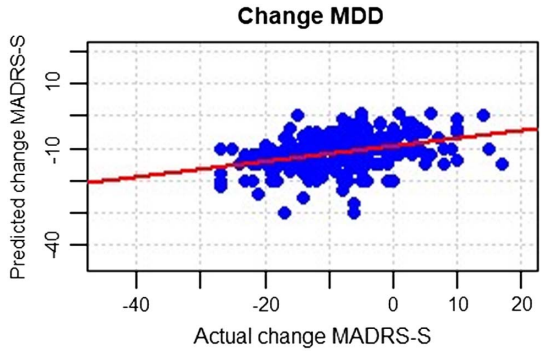
(*Appendices continue*)

Appendix B

Scatterplots of Predicted Versus Actual Change Scores Including Fitted Regression Lines

Figure B1

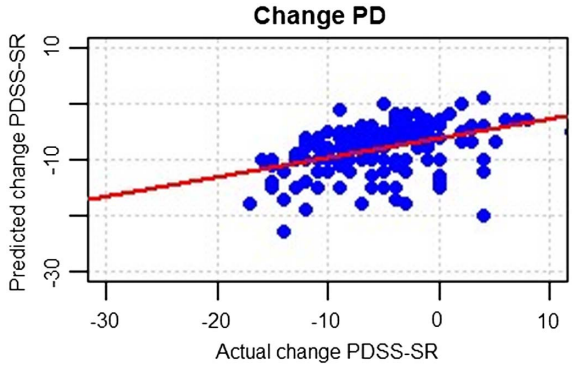
Correlation Between Predicted and Observed Change in Primary Symptom Measure for Depression



Note. MDD = major depressive disorder; MADRS-S = Montgomery–Åsberg Depression Rating Scale–Self-Report. See the online article for the color version of this figure.

Figure B2

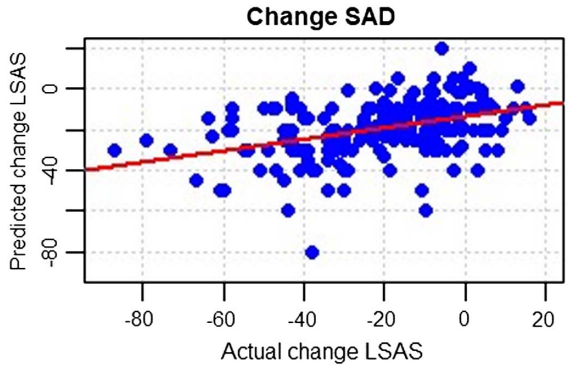
Correlation Between Predicted and Observed Change in Primary Symptom Measure for Panic Disorder



Note. PD = panic disorder; PDSS-SR = Panic Disorder Symptom Scale–Self-Report. See the online article for the color version of this figure.

Figure B3

Correlation Between Predicted and Observed Change in Primary Symptom Measure for Social Anxiety Disorder



Note. SAD = social anxiety disorder; LSAS = Liebowitz Social Anxiety Scale–Self-Report. See the online article for the color version of this figure.

Received March 17, 2023
Revision received November 22, 2024
Accepted November 23, 2024 ■