
Article:
Delgadillo, J. orcid.org/0000-0001-5349-230X, Huey, D., Bennett, H. et al. (1 more author) (2017) Case complexity as a guide for psychological treatment selection. Journal of Consulting and Clinical Psychology, 85 (9). pp. 835-853. ISSN 0022-006X

https://doi.org/10.1037/ccp0000231

Citation: Delgadillo, J., Huey, D., Bennett, H., McMillan, D. (in press). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology*.

Case complexity as a guide for psychological treatment selection

Jaime Delgadillo [1*], Dale Huey [2], Hazel Bennett [2], Dean McMillan [3]

1. Clinical Psychology Unit, Department of Psychology, University of Sheffield, Sheffield, United Kingdom
2. Primary Care Psychological Therapies Service, Greater Manchester West Mental Health NHS Foundation Trust, Salford, United Kingdom
3. Hull Medical School and Department of Health Sciences, University of York, York, United Kingdom

Declarations of interest: None.

---

* Correspondence: Dr Jaime Delgadillo, Clinical Psychology Unit, University of Sheffield, Sheffield, United Kingdom. Email: jaime.delgadillo@nhs.net

Abstract

Objective: Some cases are thought to be more complex and difficult to treat, although there is little consensus on how to define complexity in psychological care. This study proposes an actuarial, data-driven method of identifying complex cases based on their individual characteristics.

Method: Clinical records for 1512 patients accessing low and high intensity psychological treatments were partitioned in 2 random subsamples. Prognostic indices (PI) predicting post-treatment reliable and clinically significant improvement (RCSI) in depression (PHQ-9) and anxiety (GAD-7) symptoms were estimated in one subsample using penalized (Lasso) regressions with optimal scaling. A PI-based algorithm was used to classify patients as standard (*St*) or complex (*Cx*) cases in the second (cross-validation) subsample. RCSI rates were compared between *Cx* cases that accessed treatments of different intensities using logistic regression.

Results: *St* cases had significantly higher RCSI rates compared to *Cx* cases (OR = 1.81 to 2.81). *Cx* cases tended to attain better depression outcomes if they were initially assigned to high intensity (vs. low intensity) interventions (OR = 2.23); a similar pattern was observed for anxiety but the odds ratio (1.74) was not statistically significant.

Conclusions: Complex cases could be detected early and matched to high intensity interventions to improve outcomes.

What is the public health significance of this article?

Complex cases tend to have a poor prognosis after psychological treatment for depression and anxiety problems. An evidence-based model of defining complexity is proposed to guide therapists in matching patients to treatments of differing intensity. The findings indicate that this personalized method of treatment selection could lead to better outcomes for complex cases and could improve upon decisions that are informed by clinical judgment alone.

Key words: stratified medicine; mental health; case complexity; psychotherapy

A commonly held view in clinical psychology is that complex cases require suitably intensive interventions guided by formulations that account for obstacles to improvement (Tarrier, 2006). Clinical wisdom reflected in treatment textbooks suggests that a variety of factors can complicate treatment, such as chronic symptoms, comorbidity, personality disorders, physical illnesses, etc. (Beck, 1998; Hawton, Salkovskis, Kirk, & Clark, 1989; Tarrier, Wells, & Haddock, 1998). Along these lines, Ruscio and Holohan (2006) proposed a list of more than forty factors that characterise complex cases, clustered around several themes including symptoms, safety, physical, intellectual, personality and other features. Evidently, case complexity is a heterogeneous concept and there is little consensus about the features that define such cases.

Moreover, the empirical literature casts doubt over the predictive value of many variables presumed to hinder the effectiveness of therapy (Garfield, 1994). A case in point is found in the study by Myhr et al. (2007), in which only five out of ten variables thought to be indicative of poor suitability for cognitive therapy were (weakly) correlated with post-treatment outcomes. It is also well documented that clinicians' prognostic assessment of patients tends to be inaccurate (Ægisdóttir et al., 2006; Grove & Meehl, 1996), often failing to identify complex cases at risk of poor treatment outcomes (Hannan et al., 2005). In another study, patients randomly assigned to brief manualized interventions offered in a stepped care model had comparable outcomes to patients whose treatments were selected and informed by clinical judgment (Van

Straten, Tiemens, Hakkaart, Nolen, & Donker, 2006). Such evidence calls into question clinicians' ability to match patients to treatments and supports current guidelines to apply a stepped care approach (National Institute for Health and Care Excellence [NICE], 2011). Overall, three key problems are apparent: a lack of conceptual clarity about complex cases, a gap between clinical judgement and research evidence, and limitations in clinicians' ability to identify and select optimal treatments for complex cases.

Concerns regarding complexity are not exclusive to the practice of psychotherapy. The simultaneous growth and ageing of the general population have confronted many other areas of healthcare with the challenges of treating patients who present with multiple chronic conditions (Smith & O'Dowd, 2007), leading some to question the usefulness of evidence-based guidelines that are formulated for 'prototypical' patients (Boyd et al., 2005; Tinetti, Bogardus, & Agostini, 2004). Consequently, theoretical models to account for case complexity in medicine have been proposed in the last decade. Some of these models conceptualise complexity as arising from a combination of clinical (e.g., diagnostic), biological, socioeconomic, cultural, environmental and behavioural factors that are statistically associated with clinical prognosis (Safford, Allison, & Kiefe, 2007; Schaink et al., 2012). Individual patients may have protective or risk factors across these domains, and their overall complexity level results from the sum of risks. In an attempt to move beyond mere description, Shippee et al. (2012) proposed a *cumulative complexity* model which attempts to

explain how risk factors accumulate and interact to influence healthcare outcomes. They proposed that (clinical, socioeconomic, cultural) risk factors complicate healthcare outcomes by disrupting the balance between *patient workload* (i.e., number and difficulty of daily life demands including self-care) and *patient capacity* (i.e., resources and limitations affecting ability to meet demands). From this perspective, effective healthcare for complex cases would not only require intensive treatment of acute symptoms and specific disease mechanisms, but also attending to wider biopsychosocial aspects that may redress the balance between demands and capacity. Common to these models are the focus on empirically-supported prognostic factors, the consideration of factors across multiple domains, and the conceptual understanding of case complexity as resulting from the accumulation of risks and challenges to self-management.

Informed by these theoretical models emerging from the biomedical sciences, this study investigated the impact of case complexity in routine psychological care. Considering the problems outlined above, we sought to assess the merits of an actuarial, data-driven, cumulative model of defining case complexity. Specific objectives were: (1) to identify prognostic variables associated with psychological treatment outcomes; (2) to develop an algorithm that could aid clinicians in identifying complex cases at risk of poor outcomes; (3) to determine whether or not complex cases respond differentially to treatments of differing levels of intensity; (4) to ascertain

the extent to which patients are adequately matched to available stepped care interventions.

Method

Setting and Interventions

This study was based on the analysis of clinical data routinely collected by a primary care psychological therapy service in Northern England. The study was approved as a service evaluation by the local National Health Service (NHS) Trust, which did not require formal ethical approval. The service offered low and high intensity interventions for depression and anxiety problems, as part of the *Improving Access to Psychological Therapies* (IAPT) programme (Clark et al., 2009). Low intensity treatments (LIT) consisted of brief (<8 sessions lasting 30 minutes) psychoeducational interventions based on principles of cognitive behavioural therapy (CBT). These were highly structured interventions, supported by didactic materials and delivered by a workforce of psychological wellbeing practitioners trained to a standard national curriculum (Bennett-Levy et al., 2010). High intensity treatments (HIT) were lengthier (up to 20 sessions lasting around 60 minutes) interventions including CBT and counselling for depression. These interventions were also protocol-driven, delivered by post-graduate level counsellors and psychotherapists, following national treatment guidelines (NICE, 2010) and competency frameworks (e.g., Roth & Pilling, 2008). All therapists practiced under regular clinical supervision (weekly or fortnightly) to ensure ethical

practice and treatment fidelity. These interventions were organised in a stepped care model (NICE, 2011), where most patients initially accessed a LIT and those with persistent and/or severe symptoms accessed HIT. Initial treatment assignment was determined by therapists who carried out standardised intake assessments.

Measures and Data Sources

Primary outcome measures. Patients accessing IAPT services self-complete standardised outcome measures on a session-to-session basis to monitor response to treatment. The PHQ-9 is a nine-item screening tool for major depression, where each item is rated on a 0 to 3 Likert scale, yielding a total depression severity score between 0–27 (Kroenke, Spitzer, & Williams, 2001). A cut-off ≥ 10 has been recommended to detect clinically significant depression symptoms (Kroenke, Spitzer, & Williams, 2001), and a difference of ≥6 points between assessments is indicative of reliable change (Richards & Borglin, 2011).

The GAD-7 is a seven-item measure developed to screen for anxiety disorders (Spitzer, Kroenke, Williams, & Löwe, 2006). It is also rated using Likert scales, yielding a total anxiety severity score between 0–21. A cut-off score ≥8 is recommended to identify the likely presence of a diagnosable anxiety disorder (Kroenke, Spitzer, Williams, Monahan, & Löwe, 2007), and a difference of ≥5 points is indicative of reliable change (Richards & Borglin, 2011). Pre-treatment and last observed PHQ-9 and GAD-7 scores were available for analysis.

Other measures. The Work and Social Adjustment Scale (WSAS) is a measure of functioning across five domains: work, home management, social leisure activities, private leisure activities, family and close relationships (Mundt, Marks, Shear, & Greist, 2002). Each item is rated on a scale of 0 (no impairment) to 8 (very severe impairment), rendering a total functional impairment score between 0–40.

The Standardised Assessment of Personality – Abbreviated Scale (SAPAS) is an eight-item questionnaire developed to screen for the likely presence of a personality disorder (Moran et al., 2003). Each question prompts respondents to endorse specific personality traits (yes/no), yielding a total score between 0–8 where a cut-off >3 is indicative of cases with a high probability of diagnosable personality disorders. The WSAS and SAPAS were gathered at the time of initial assessments.

De-identified treatment and demographic data were also available, including information on referral sources, the intensity and sequence of treatments received (LIT and/or HIT along the stepped care pathway), age, gender, ethnicity and employment status. Formal diagnostic assessments were not carried out in routine care, but primary presenting problems noted in clinical records were available in summary form as group-level percentages.

Sample Characteristics

The study included case records for a total of 2202 patients who had been discharged from the service at the time of data collection.

Complete data (described above) were available for 1512 (68.7%) cases. More than half were females (63.9%), with a mean age of 41.99 (SD = 14.54; range: 16 – 87) and of white British ethnic background (88.2%). A quarter (24.9%) of all cases were unemployed and/or in receipt of incapacity benefits. Approximately 59.9% were referred to treatment by general medical practitioners; the remainder self-referred (24.3%) or were referred by other social and healthcare providers (15.8%). The presenting problems noted in clinical records were depression (21.0%), recurrent depression (6.6%), obsessive-compulsive disorder (4.4%), adjustment disorders (5.7%), somatoform disorders (0.4%), eating disorders (0.4%), phobic disorders (5.7%), other anxiety disorders (42.4%), and unspecified mental health problems (13.4%). Mean baseline severity scores for the whole cohort were PHQ-9 = 14.86 (SD = 6.33), GAD-7 = 13.27 (SD = 5.07), WSAS = 18.39 (SD = 9.46), SAPAS = 3.82 (SD = 1.89; cases with SAPAS > 3 = 54.2%). Many patients had comorbid presentations, where 71.4% of cases had case-level symptoms in both PHQ-9 and GAD-7. Approximately 76.6% of patients were initially assigned to LIT and 23.4% were initially assigned to HIT. Overall, 40.6% only accessed LIT, 36.0% accessed LIT + HIT, and 23.4% only accessed HIT. Overall, 31.3% dropped out of treatment (32.2% of those initially assigned to LIT; 28.5% of those initially assigned to HIT).

Statistical Analysis

Consistent with the objectives of the study, data analyses were performed in 4 stages aiming to develop, validate and assess the clinical

utility of a cumulative complexity model. The primary analyses were carried out in the dataset of cases with complete data (N = 1512). Following a cross-validation approach, we partitioned this dataset into two random halves which were treated as estimation (N = 755) and validation (N = 757) samples. In order to assess the potential influence of missing data, a single imputed estimation sample (N = 1108) was derived using an expectation-maximization method (Schafer & Olsden, 1998) and was used for sensitivity analyses described below.

Stage 1 involved the development of a prognostic index and classification method to identify complex cases in routine care. The dependent variable in all models was a binary indicator of post-treatment reliable and clinically significant improvement (RCSI), with separate models for depression (PHQ-9) and anxiety (GAD-7) measures. RCSI was determined using the criteria proposed by Jacobson and Truax (1991), based on combining reliable change indices for PHQ-9 (≥6) and GAD-7 (≥5) described by Richards and Borglin (2011) and diagnostic cut-offs for each measure (PHQ-9 <10; GAD-7 <8). The dependent variable was coded as follows: 0 = RCSI; 1 = no RCSI, such that the prognostic models would be constructed to identify (more complex) cases with increased probability of poor outcomes.

As an initial variable screening step, we used univariate logistic regressions to examine the *goodness-of-fit* (based on -2 log likelihood test and magnitude of AIC and BIC statistics) of linear and non-linear trends for continuous variables, as well as alternative ways to model the SAPAS questionnaire (as a total score, dichotomized based on a cut-

off >3, or entered as a series of 8 binary items). Entering all 8 SAPAS binary items yielded the best fitting models in preliminary tests (i.e., lowest AIC and BIC, significant -2 log likelihood tests) and confirmed that only 5 items were significant ($p < 0.05$) predictors of outcome. Furthermore, baseline severity (PHQ-9, GAD-7), impairment (WSAS) and age variables were optimally modelled using non-linear trends. Age was rescaled to ordinal decade groups (e.g., teens, twenties, thirties, etc.) and reverse scored (oldest group coded 0, youngest group coded 6) based on the observed trend of correlations between age and RCSI.

Informed by these preliminary tests, we applied penalized categorical regressions with optimal scaling (CATREG-Lasso) in the main analysis. CATREG applies classical linear regression to predictor variables that are transformed to categorical quantifications which are optimally suited to explore nonlinear relations in the data (Gifi, 1990). Continuous variables were thus transformed using a monotonic spline scaling level to examine non-linear associations with the dependent variable. Variable selection and regularization were performed combining the Lasso (Least Absolute Shrinkage and Selection Operator; Tibshirani, 1996) and the .632 bootstrap resampling method (Efron & Tibshirani, 1997). The Lasso imposes a penalty term that shrinks coefficients towards zero, penalizing the sum of the squared regression coefficients. This yields more generalizable prediction equations compared to conventional regression models which are prone to overfitting and are less reliable in the presence of multicollinearity. Since using different penalty terms results in different shrunken

coefficients, resampling techniques are often used to determine an optimal penalty. The .632 bootstrap resampling method is a smoothed version of the *leave-one-out* cross-validation strategy, which permits the estimation of a model's expected prediction error. This resampling method was applied 1000 times to each Lasso model, iteratively increasing the penalty term in 0.01 units, until all coefficients were shrunk to zero. The *one-standard-error rule* was applied to select the most parsimonious Lasso model within one standard error of the model with minimum expected prediction error. The predictors entered into CATREG-Lasso models included clinical (baseline PHQ-9, GAD-7, WSAS), personality (SAPAS items 1, 2, 3, 5, 7) and demographic variables (age groups, gender, ethnicity, employment status). Shrunken coefficients from the optimal models were used to calculate a prognostic index (PI) for each patient, where a higher PI denotes poorer prognosis. PI's were retained in the CATREG quantifications scale, with signed and continuous scores centred at zero.

The above procedure was conducted in the estimation samples with complete and imputed data, allowing us to compare the area under the curve (AUC) for the PI's derived from each dataset as an indicator of predictive accuracy. PI's derived using complete and imputed samples had comparable AUC statistics albeit with some shrinkage observed in the imputed dataset (PHQ-9: 0.67 ± 0.04 vs. 0.63 ± 0.05; GAD-7: 0.74 ± 0.04 vs. 0.66 ± 0.04). Therefore, subsequent analyses were applied in the dataset with complete data.

In stage 2, we applied receiver operating characteristic (ROC) curve analysis (Altman & Bland, 1994) in the estimation sample to determine empirical cut-offs that optimally balanced sensitivity and specificity on each PI. Consistent with our assumptions about clinical complexity, cases where both (PHQ-9 and GAD-7) PI's were above empirical cut-offs were classed as complex (*Cx*), and others (including all those with sub-clinical symptoms) were classed as standard (*St*) cases. The agreement of both PI's was taken as a conservative means of minimising 'false positive' classifications, and limiting the *Cx* classification to cases with the poorest prognoses across both outcome domains. We then tested our assumptions about prognosis and cumulative complexity in the validation subsample, with cases whose symptoms were above diagnostic cut-offs for each outcome measure (PHQ-9: N = 675; GAD-7: N = 755). ROC curve analyses were used to assess how well the PI's (using Lasso-based shrunken coefficients from the estimation sample) performed out-of-sample (in a statistically independent validation sample). In addition, separate logistic regression models were applied for each outcome (PHQ-9, GAD-7), where the dependent variable was post-treatment RCSI status (0 = no RCSI; 1 = RCSI) and the predictors included case complexity (0 = *Cx*, 1 = *St*) controlling for baseline severity of symptoms (PHQ-9 or GAD-7 respectively).

Stage 3 analyses were also conducted in the validation sample. A logistic regression model predicting (HIT vs. LIT) group membership based on all clinical and demographic characteristics was performed to

estimate propensity scores, denoting the predicted probability of completing a treatment episode at HIT. Propensity scores were entered as a covariate in subsequent analyses to control for confounding by indication. Next, logistic regression models were applied with RCSI status as a dependent variable, entering baseline severity (PHQ-9 or GAD-7 respectively), propensity scores, and treatment pathway (LIT or HIT only vs. LIT + HIT) as predictors. The models were performed separately in the subgroups of *Cx* (N = 269) and *St* (N = 425) cases (with available data to estimate propensity scores), to minimise multicollinearity between propensity scores and case complexity dummy variables in the same model.

In stage 4, we assessed the extent to which initial treatment assignment (LIT or HIT) determined by clinical judgement was consistent with the assignment that would be indicated by the prognostic method described above. A prognostic treatment assignment was coded for all patients, where starting at HIT was recommended for *Cx* cases and starting at LIT was recommended for all other cases. Next, agreement codes were noted for each case in the full sample, where '1' indicated agreement between clinical judgment and prognosis, and '0' indicated disagreement. Agreement codes were aggregated across the entire sample to estimate a 'hit rate', denoting the percentage of cases where clinicians' decisions converged with a prognostic strategy for treatment assignment. Next, we applied Cohen's Kappa across agreement codes to derive a *Treatment Matching precision* (TMaP) score, which takes into account the probability that 'hit rates' may be due to

chance. The TMap score is therefore a robust measure of convergence between clinical and empirical decision-making strategies, ranging between 1 (perfect agreement) and –1 (complete disagreement), where 0 is indicative of agreement by chance. TMaP scores were estimated for the full sample and for individual clinicians that undertook initial assessments and made decisions about treatment assignment for at least 20 patients (to eschew extreme scores in caseloads with small base rates).

Results

Estimation of Prognostic Equations

Using the CATREG-Lasso procedure in the estimation sample, we arrived at prognostic models that explained between 9% (PHQ-9: optimal scaling adjusted R^2 = 0.09) and 15% (GAD-7: adj R^2 = 0.15) of variance in post-treatment RCSI. Regression and ROC curve model estimates for each outcome measure are presented in Table 1 (with detailed outputs in supplementary appendix 1). Several predictors were selected into optimal Lasso models, including demographic (age, ethnicity, employment), personality (SAPAS items: 2 = “interpersonally avoidant”, 3 = “suspicious”, 5 = “impulsive”, 7 = “dependent”), and clinical features (baseline PHQ-9, GAD-7, WSAS).

The R^2 share statistic reflects the relative contribution of each predictor to the model’s overall adjusted R^2, after partialling out the specific and combined effects of the other variables. In the depression model, demographics had relatively greater explanatory influence

(22.5%) relative to personality (14.7%) and clinical features (15%), although the remaining R^2 variance was large (47.9%) and reflected the combined influence of all variables in the model. In the anxiety model, clinical features (55.9%) had two to three times greater explanatory power relative to personality (23.9%) and demographic features (15.2%), leaving only 5% of the remaining R^2 variance to combined effects. The *F* tests for specific variables in both models suggested that the removal of clinical factors (particularly PHQ-9) significantly deteriorated the predictive power of regression models. AUC statistics for the depression (0.67, SE = 0.02) and anxiety (0.74, SE = 0.02) prognostic indices applied to predict RCSI in the estimation sample were both statistically significant ($p < 0.001$); ROC curves are shown in appendix 2.

[Table 1]

Validation of Case Complexity Model

PI's using the shrunken coefficients derived from the estimation sample were applied in the validation sample, yielding stable and statistically significant ($p < 0.001$) AUC estimates for depression (0.64, SE = 0.02) and anxiety (0.70, SE = 0.02) measures (see ROC curves in appendix 2). Overall, 28.6% of all patients were classified as *Cx* by the prognostic classification rule derived using ROC curve analyses. The proportion of *Cx* cases was lower in the subsample of patients who only accessed LIT (15.9%) by comparison to those who accessed LIT+HIT

(37.3%) and those who only accessed HIT (36.7%); x^2 (2) = 97.05, p < 0.001.

[Table 2]

As illustrated in Figure 1, logistic regression models (Table 2) confirmed that *St* cases were significantly more likely to attain RCSI in depression (OR = 1.81) and anxiety (OR = 2.81) symptoms compared to *Cx* cases, after controlling for baseline severity.

[Figure 1]

Case Complexity and Treatment Selection

Logistic regression models presented in Table 3 indicated that *Cx* cases had a significantly greater probability of RCSI in depression symptoms if they directly accessed HIT, by comparison to a standard stepped care pathway LIT + HIT; OR = 2.23, p = 0.01. There was also a trend indicating the same advantage of HIT for *Cx* cases in the anxiety model, although this did not reach statistical significance; OR = 1.74, p = 0.08. No significant differences were found between treatment pathways in the regression models applied to *St* cases. These analyses controlled for baseline symptom severity and propensity scores (derived from logistic regression model in appendix 3). The results for the depression outcomes are illustrated in Figure 2; where *Cx* cases that were initially assigned to HIT (optimal prognostic treatment assignment)

had a 16.3% increased probability of RCSI by comparison to *Cx* who were assigned to a conventional stepped care pathway (LIT+HIT).

[Table 3]

[Figure 2]

Clinical judgment versus prognostic models

The aggregated hit rate in the full sample indicated that clinicians' treatment assignment decisions agreed with the prognostic strategy in 65.6% of cases. The TMaP score for the full sample, however, was low (k = 0.09, SE = 0.02, $p < 0.001$). A closer examination of individual therapists' treatment assignment decisions (N = 1247 nested within 26 therapists) revealed considerable variability in their hit rates (range = 36.5% to 84.7%, mean = 62.9, SD = 14.3) and TMaP scores (range = -0.27 to 0.44, mean = 0.05, SD = 0.20). As shown in Figure 3, hit rates and TMaP scores were moderately correlated ($r = 0.67$, $p < 0.001$), and approximately 48% of therapists had TMaP scores < 0.

[Figure 3]

Discussion

Main findings

This study set out to contribute to the understanding of case complexity in psychological care, in view of the limited conceptual clarity and evidence base surrounding this topic. Our findings demonstrate that (1) several patient characteristics have a cumulative effect on treatment outcomes; (2) it is possible to make reasonably accurate prognoses using this information; (3) prognostic models can help us to operationalize case complexity in a way that is clinically useful. Cases classed as *Cx* (28.6%) on the basis of prognostic data tended to have significantly poorer outcomes after psychological treatment. Furthermore, *Cx* cases were two times (OR = 2.23) more likely to attain RCSI in depression symptoms if they were initially assigned to a high intensity intervention instead of usual stepped care. A similar trend was observed for anxiety symptoms, although this did not reach statistical significance.

A conceptual bridge between prognosis and case complexity

These results lend support to the clinical notion that some cases are more difficult to treat due to various complicating factors (Ruscio & Holohan, 2006), although clinicians' intuitions and treatment planning are often inconsistent with the evidence base (Garb, 2005). We found that treatment assignment decisions guided by clinical judgment were consistent with prognostic models in 65.6% of cases. This rate of

agreement could be achieved by chance, or simply by mechanically following stepped care guidelines and assigning all cases initially to LIT, since the base rate of Cx cases is relatively low (under 30%). This was evidenced more clearly by examining the aggregated TMaP score (0.09) which was close to zero. Overall, the findings indicate that depression improvement (RCSI) rates for *Cx* cases could be significantly increased (by approximately 16.3%) if clinical judgment was supported by prognostic treatment selection models.

This gap between practice and science is perhaps accentuated by an unwieldy literature on the topic of prognosis in psychological care. Previous authors have attempted to synthesize findings across multiple studies to elucidate predictors of depression and anxiety outcomes (e.g., Driessen & Hollon, 2010; Haby, Donnelly, Corry, & Vos, 2006; Hamilton & Dobson, 2002; Keeley et al., 2008; Kessler et al., 2017; Licht-Strunk et al., 2007; Nilsen, Eisemann, & Kvernmo, 2013). Although some convergent findings are evident, meta-analytic reviews that privilege data from clinical trials are limited by typically small samples with sparse and heterogeneous prognostic variables, often gathered in highly selected participants (i.e., those with specific disorders) that may not be representative of complex cases seen in routine care (Chambless & Ollendick, 2001). Naturalistic cohort studies can offer informative evidence to complement findings from controlled trials, especially where multiple variables are measured systematically across large healthcare populations, as exemplified in this study. Several such studies are yielding replicated findings (e.g., Beard et al., 2016; Delgadillo, Moreea,

& Lutz, 2016; Delgadillo, Dawson, Gilbody, & Böhnke, *in press*; Firth, Barkham, Kellett, & Saxon, 2015; Goddard, Wingrove, & Moran, 2015; Licht-Strunk et al., 2009).

Overall, the emerging literature on outcome prediction points to factors clustered around clinical (i.e., baseline symptom severity, diagnosis, comorbidity, functioning and disability, physical illnesses), demographic (i.e., age, ethnicity, employment, socioeconomic deprivation, marital status) characterological (i.e., personality disorder diagnoses or traits, interpersonal problems and style, trait anxiety and neuroticism) and dispositional domains (i.e., readiness to change, expectancy). Informed by advances in the biomedical literature (Safford et al., 2007; Shippee et al., 2012), we propose that complex cases in psychological care are characterised by the presence of measurable factors that map onto multiple domains (clinical, demographic, characterological and dispositional), which are statistically associated with clinical prognosis and have a cumulative –detrimental– effect on treatment outcomes. The concept of case complexity is, therefore, dimensional (i.e., degrees of complexity on a continuum), and complex cases can be distinguished from others using empirically derived population norms and classification rules.

Case complexity may challenge psychological improvement through several mechanisms. One possibility is that an accumulation of disadvantages (e.g., poverty, interpersonal difficulties, functional impairment, outgroup derogation due to minority ethnic status) could disrupt the balance between life stressors and coping resources

(Shippee et al., 2012). Complexity could also interfere with adequate engagement with therapy; for example by undermining expectancy, which is a well-established predictor of treatment outcomes (Constantino et al., 2011). Baseline severity is an important contributor to complexity, so another possibility is that high baseline severity does not completely block improvement but may dampen the effect of treatment (i.e., cases with high severity can attain reliable improvement even if their symptoms do not reach sub-clinical levels). Furthermore, our findings suggest that specific features (i.e., demographic, clinical, characterological) influence specific clinical outcomes (remission of depression, anxiety) differentially. For example, demographic factors (e.g., young age, unemployment) had a considerably larger influence over depression outcomes relative to clinical and characterological factors. Future research could focus on exploring the relative contribution of different prognostic domains to multiple outcome domains (symptoms, quality of life, functioning) and the mechanisms through which these cumulative disadvantages may complicate or undermine treatment.

Limitations

Some limitations should be considered when interpreting the results of this study. As is common in naturalistic datasets, we encountered several cases with missing data (>30%). To deal with this, we applied multiple imputation and sensitivity analyses which yielded similar prognostic models, albeit with some shrinkage observed in the

imputed dataset. On this basis, it was appropriate to perform further validation analyses using cases with complete data, to simulate how prognostic assessments would be applied in routine care, where data imputation of missing values is unfeasible.

Another limitation concerning the data used in this study was that we only had access to pre-post outcome measures for the entire treatment pathway, and it was not possible to disaggregate the outcomes for LIT and HIT for cases that accessed both steps. However, we were able to determine that *Cx* that only accessed HIT tended to have better outcomes compared to those who accessed LIT + HIT (a lengthier and costly treatment pathway). This suggests that there are no benefits of having LIT sessions preceding HIT, and hence the advantage of being initially assigned to HIT may not be solely due to having a lengthier treatment episode. Previous research using more granular outcomes data for each treatment step suggested that cases with poor prognostic features had a higher probability of dropout and lower probability of improvement at the LIT step by comparison to HIT (Delgadillo, Moreea, & Lutz, 2016). These emerging findings suggest that assigning complex cases directly to HIT seems justified, although future randomized controlled trials of this strategy are necessary to determine if it is indeed more cost-effective.

Other limitations include the lack of formal diagnostic assessments and the analysis of a limited number of prognostic variables. It is known, for example, that specific diagnoses such as post-traumatic stress disorder, eating disorders and obsessive-compulsive

disorder are associated with poorer outcomes in stepped care services (Delgadillo, Dawson, Gilbody, & Böhnke, *in press*), and it is plausible that such diagnoses could interact with other prognostic features. Notwithstanding these limitations, it is remarkable that this narrow range of variables yielded an accurate and clinically useful prognostic model. Other studies using routine practice data have shown that similar variables can be used to identify subgroups of cases with depression and anxiety problems that attain similar outcomes (Delgadillo, Moreea, & Lutz, 2016; Lutz, Lowry, Kopta, Einstein, & Howard, 2001; Lutz et al., 2005; Saunders, Cape, Fearon, & Pilling, 2016).

Clinical implications

In line with recent findings in stepped-care psychological treatment settings (Delgadillo, Moreea, & Lutz, 2016; Lorenzo-Luaces, DeRubeis, van Straten, & Tiemens, 2017), the present study provides further evidence that applying prognostic indices to guide personalized treatment recommendations is likely to improve treatment outcomes. Low intensity guided self-help interventions are recommended as first-line treatments for several common mental disorders (NICE, 2011) and are becoming widely available in routine stepped care services (Clark, 2011). The application of evidence-based treatment selection algorithms like the one demonstrated in this study could help to maximise the cost-effectiveness of LIT by selectively offering it to those

who are most likely to derive benefits. Equally, prognostic models could be used to fast-track complex cases to HIT in a timely way.

Acknowledgements

We thank Jan R. Böhnke for helpful comments on an earlier version of this manuscript.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & 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*, 341-382. doi: 10.1177/0011000005285875

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 3: receiver operating characteristic plots. *BMJ*, *309*(6948), 188.

Beard, C., Stein, A. T., Hearon, B. A., Lee, J., Hsu, K. J., & Björgvinsson, T. (2016). Predictors of Depression Treatment Response in an Intensive CBT Partial Hospital. *Journal of Clinical Psychology*, 72, 297-310. doi: 10.1002/jclp.22269

Beck, J. S. (1998). Complex cognitive therapy treatment for personality disorder patients. *Bulletin of the Menninger Clinic*, 62, 170–194.

Bennett-Levy, J., Richards, D.A., Farrand, P. (2010). Low intensity CBT interventions: a revolution in mental health care. In: *Oxford guide to low intensity CBT interventions*. Oxford (e.d. Bennett-Levy J, Richards DA, Farrand P, Christensen H, Griffiths KM, Kavanagh DJ, Klein B, Lau MA, Proudfoot J, Ritterband L, White J and Williams C), pp. 3-18. Oxford University Press: Oxford.

Boyd, C.M., Darer, J., Boult, C., Fried, L.P., Boult, L., Wu, A.W. (2005). Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. *JAMA*, *294*, 716–24. doi: 10.1001/jama.294.6.716

Clark, D. M., Layard, R., Smithies, R., Richards, D. A., Suckling, R., & Wright, B. (2009). Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy*, *47*, 910-920. doi: 10.1016/j.brat.2009.07.010

Clark D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry*, 23, 318–327. doi: 10.3109/09540261.2011.606803

Chambless, D. L. and Ollendick, T. H. (2001). Empirically supported psychological interventions: controversies and evidence. *Annual Review of Psychology*, *52*, 685–716. doi: 10.1146/annurev.psych.52.1.685

Constantino, M. J., Arnkoff, D. B., Glass, C. R., Ametrano, R. M., & Smith, J.Z. (2011). Expectations. Journal of Clinical Psychology: In Session, 67, 184–192. doi: 10.1002/jclp.20754

Delgadillo, J., Dawson, A., Gilbody, S., Böhnke, J.R. (*in press*). The impact of long term medical conditions on the outcomes of psychological therapy for depression and anxiety. *British Journal of Psychiatry*. doi: 10.1192/bjp.bp.116.189027

Delgadillo, J., Moreea, O., Lutz, W. (2016). Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour Research and Therapy*, *79*, 15-22. doi: 10.1016/j.brat.2016.02.003

Driessen, E., & Hollon, S. D. (2010). Cognitive behavioral therapy for mood disorders: efficacy, moderators and mediators. *Psychiatric Clinics of North America*, *33*, 537-555. doi: 10.1016/j.psc.2010.04.005

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438),548-560. doi: 10.1080/01621459.1997.10474007

Firth, N., Barkham, M., Kellett, S., & Saxon, D. (2015). Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis. *Behaviour Research and Therapy*, *69*, 54-62. doi: 10.1016/j.brat.2015.04.001

Garb, H.N. (2005). Clinical judgement and decision making. *Annual Review of Clinical Psychology*, *1*, 67–89.

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin, & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed.). New York: Wiley & Sons.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, England: Wiley.

Goddard, E., Wingrove, J., & Moran, P. (2015). The impact of comorbid personality difficulties on response to IAPT treatment for depression and anxiety. *Behaviour Research and Therapy*, *73*, 1–7. doi: 10.1016/j.brat.2015.07.006

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy and Law*, *2*, 293–323. doi: 10.1037/1076-8971.2.2.293

Haby, M. M., Donnelly, M., Corry, J., & Vos, T. (2006). Cognitive behavioural therapy for depression, panic disorder and generalized anxiety disorder: a meta-regression of factors that may predict outcome. *Australian and New Zealand Journal of Psychiatry*, *40*, 9-19.

Hamilton, K.E., Dobson, K.S. (2002). Cognitive therapy of depression: pretreatment patient predictors of outcome. *Clinical Psychology Review*, *22*, 875–894. doi: 10.1016/S0272-7358(02)00106-X

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/In Session*, *61*, 1–9. doi: 10.1002/jclp.20108

Hawton, K., Salkovskis, P., Kirk, J., Clark, D.M., (Eds). (1989). *Cognitive behaviour therapy for psychiatric problems: a practical guide*. Oxford: Oxford University Press.

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

Keeley, M. L., Storch, E. A., Merlo, L. J., & Geffken, G. R. (2008). Clinical predictors of response to cognitive-behavioral therapy for obsessive–compulsive disorder. *Clinical Psychology Review*, *28*, 118-130. doi: 10.1016/j.cpr.2007.04.003

Kessler, R., van Loo, H., Wardenaar, K., Bossarte, R., Brenner, L., Ebert, D., . . . Sampson, N. (2017). Using patient self-reports to study

heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, *26*, 22-36. doi: 10.1017/S2045796016000020

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

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine*, *146*, 317–325. doi: 10.7326/0003-4819-146-5-200703060-00004

Licht-Strunk, E., Windt, D.A.W.M., van Marwijk, H.W.J., de Haan, M., Beekman, A.T.F. (2007). The prognosis of depression in older patients in general practice and the community. A systematic review. *Family Practice*, *24*, 168-80. doi: 10.1093/fampra/cml071

Licht-Strunk, E., Van Marwijk, H. W. J., Hoekstra, T., Twisk, J. W. R., De Haan, M., & Beekman, A. T. F. (2009). Outcome of depression in later life in primary care: longitudinal cohort study with three years' follow-up. *British Medical Journal*, *338*, a3079. doi: 10.1136/bmj.a3079

Lorenzo-Luaces, L., DeRubeis, R. J., van Straten, A., & Tiemens, B. (2017). A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. *Journal of Affective Disorders*, 213, 78-85. doi: 10.1016/j.jad.2017.02.010

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

Lutz, W., Lowry, J., Kopta, S.M., Einstein, D.A., & Howard, K.I. (2001). Prediction of Dose–Response Relations Based on Patient Characteristics. *Journal of Clinical Psychology*, *57*, 889–900. doi: 10.1002/jclp.1057

Moran, P., Leese, M., Lee, T., Walters, P., Thornicroft, G., & Mann, A. (2003). Standardised assessment of personality e abbreviated scale (SAPAS): preliminary validation of a brief screen for personality disorder. *British Journal of Psychiatry*, *183*, 228-232. doi: 10.1192/bjp.183.3.228

Mundt, J. C., Mark, I. M., Shear, M. K., Griest, J. M. (2002). The Work and Social Adjustment Scale: a simple measure of impairment in

functioning. *British Journal of Psychiatry*, *180*, 461-464. doi: 10.1192/bjp.180.5.461

Myhr, G., J. Talbot, L. Annable and G. Pinard. (2007). Suitability for Short-Term Cognitive Behavioral Therapy. *Journal of Cognitive Psychotherapy*, *21*, 334–45. doi: 10.1891/088983907782638743

National Institute for Health and Care Excellence. (2010). *Depression: the treatment and management of depression in adults (updated Ed.)*. London: The British Psychological Society and The Royal College of Psychiatrists.

National Institute for Health and Care Excellence (2011). *Common mental health disorders: Identification and pathways to care*. London: NICE. Retrieved from http://www.nice.org.uk/guidance/CG123

Nilsen, T. S., Eisemann, M., & Kvernmo, S. (2013). Predictors and moderators of outcome in child and adolescent anxiety and depression: a systematic review of psychological treatment studies. *European Child & Adolescent Psychiatry*, *22*, 69-87. doi: 10.1007/s00787-012-0316-3

Richards, D. A., & Borglin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: Two year prospective cohort study. *Journal of Affective Disorders*, *133*, 51-60. doi: 10.1016/j.jad.2011.03.024

Roth, A. D., & Pilling, S. (2008). Using an evidence-based methodology to identify the competencies required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behavioural and Cognitive Psychotherapy*, *36*, 129–147. doi: 10.1017/S1352465808004141

Safford, M.M., Allison, J.J., and Kiefe, C.I. (2007). Patient Complexity: More Than Comorbidity. The Vector Model of Complexity. *Journal of General Internal Medicine*, *22* Suppl. 3, 382–90. doi: 10.1007/s11606-007-0307-0

Saunders, R., Cape, J., Fearon, P., & Pilling, S. (2016). Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. *Journal of Affective Disorders*, *197*, 107-115. doi: 10.1016/j.jad.2016.03.011

Schaink, A. K., Kuluski, K., Lyons, R. F., Fortin, M., Jadad, A. R., Upshur, R., & Wodchis, W. P. (2012). A scoping review and thematic classification of patient complexity: offering a unifying framework. *Journal of Comorbidity*, *2*, 1-9. doi: 10.15256/joc.2012.2.15

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective.

Multivariate Behavioral Research, *33*, 545-571. doi: 10.1207/s15327906mbr3304_5

Shippee, N. D., Shah, N. D., May, C. R., Mair, F. S., & Montori, V. M. (2012). Cumulative complexity: a functional, patient-centered model of patient complexity can improve research and practice. *Journal of Clinical Epidemiology*, *65*, 1041-1051. doi: 10.1016/j.jclinepi.2012.05.005

Smith, S.M., O'Dowd, T. (2007). Chronic diseases: what happens when they
come in multiples? *British Journal of General Practice*, *57*, 268–70.

Spitzer, R., Kroenke, K., Williams, J. B. W., & Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*, 1092–1097. doi: 10.1001/archinte.166.10.1092

Tarrier, N., Wells, A., & Haddock, G., (Eds.). (1998). *Treating complex cases: The cognitive behavioural therapy approach*. Chichester: John Wiley & Sons Ltd.

Tarrier, N. (Ed.). (2006). *Case formulation in cognitive behaviour therapy: The treatment of challenging and complex clinical cases*. London: Routledge.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological*), 58, 267-288.

Tinetti, M.E., Bogardus, S.T., Agostini, J.V. (2004). Potential pitfalls of disease specific guidelines for patients with multiple conditions. *New England Journal of Medicine*, *351*, 2870–4. doi: 10.1056/NEJMsb042458

Van Straten, A., Tiemens, B., Hakkaart, L., Nolen, W., & Donker, M. (2006). Stepped care vs. matched care for mood and anxiety disorders: a randomized trial in routine practice. *Acta Psychiatrica Scandinavica*, *113*, 468-476. doi: 10.1111/j.1600-0447.2005.00731.x

Table 1. Estimation of prognostic indices using penalized categorical regression with optimal scaling

	Depression (PHQ-9) model parameters				Anxiety (GAD-7) model parameters			
	F(591) = 4.13, $p < 0.001$				F(643) = 6.85, $p < 0.001$			
	Adjusted $R^2 = 0.09$				Adjusted $R^2 = 0.15$			
	AUC = 0.67, SE = 0.02 (0.64, 0.71), p < 0.001				AUC = 0.74, SE = 0.02 (0.70, 0.78) , p < 0.001			
	Lasso-based coefficients				Lasso-based coefficients			
			importance				importance	
Predictors	**B**	**SE**	**R^2 share**	**removal impact**	**B**	**SE**	**R^2 share**	**removal impact**
Gender	0.000	0.002			0.000	0.007		
Ethnicity	0.000	0.009			0.007*	0.023	3.9%	$p = 0.77$
Age group	0.019*	0.028	8.8%	$p = 0.72$	0.022*	0.030	4.4%	$p = 0.65$
Employment	0.065*	0.037	13.7%	$p = 0.08$	0.067*	0.038	6.9%	$p = 0.07$
SAPAS item 1	0.000	0.026			0.000	0.015		
SAPAS item 2	0.050*	0.037	9.4%	$p = 0.18$	0.072*	0.040	7.3%	$p = 0.08$
SAPAS item 3	0.016*	0.025	1.7%	$p = 0.52$	0.052*	0.036	4.1%	$p = 0.15$
SAPAS item 5	0.000	0.019			0.056*	0.034	9.8%	$p = 0.11$
SAPAS item 7	0.007*	0.024	3.6%	$p = 0.77$	0.006*	0.023	2.7%	$p = 0.79$
Baseline PHQ-9	0.085*	0.040	9.0%	$p < 0.01$	0.138*	0.038	19.0%	$p < 0.001$
Baseline GAD-7	0.034*	0.035	6.0%	$p = 0.39$	-0.094*	0.041	30.8%	$p = 0.02$
Baseline WSAS	0.000	0.023			0.057*	0.037	6.1%	$p = 0.07$

Notes: Dependent variables in both models are categorical markers for post-treatment remission of symptoms (0 = remission; 1 = no remission); AUC = area under the curve statistic; Beta coefficients are expressed in a categorical quantification scale; * predictors selected into optimal Lasso model; SE = standard errors aggregated over 1000 bootstrap samples; R^2 share = squared partial correlation between predictor and outcome / adjusted R^2; removal impact = *F* test probability of model deterioration if the predictor is removed; Gender: female = 0, male =1; Ethnicity: white British = 0, minority ethnic group = 1; Employment: employed = 0, unemployed = 1; SAPAS: item not endorsed = 0, item endorsed = 1, a “no” answer to item 3 is reverse scored = 1

Table 2. Validation of prognostic indices applied out-of-sample using logistic regression

	Depression (PHQ-9) model parameters				Anxiety (GAD-7) model parameters			
	Nagelkerke R^2 = 0.08				Nagelkerke R^2 = 0.08			
Predictors	**B**	**SE**	**OR**	**(95% CI)**	**B**	**SE**	**OR**	**(95% CI)**
Baseline severity †	-0.06**	0.02	0.94	0.90, 0.98	-0.001	0.02	0.99	0.96, 1.04
Case complexity	0.59**	0.20	1.81	1.21, 2.69	1.03***	0.18	2.81	1.98, 3.98
Constant	0.57	0.46	1.77		-0.70	0.40	0.50	

Notes: Dependent variables in both models are categorical markers for post-treatment remission of symptoms (0 = no remission; 1 = remission); ** $p < 0.01$; *** $p < 0.001$; SE = standard error; † the baseline severity measure (either PHQ-9 or GAD-7) entered in each model matched the relevant outcome variable; Case complexity: *Cx* = 0, *St* =1

Table 3. Logistic regression models assessing case complexity and treatment selection

	Depression (PHQ-9) model parameters				Anxiety (GAD-7) model parameters			
Predictors	**B**	**SE**	**OR**	**(95% CI)**	**B**	**SE**	**OR**	**(95% CI)**
Subsample of *Cx* cases	Nagelkerke R^2 = 0.11				Nagelkerke R^2 = 0.10			
Baseline severity †	-0.10**	0.03	0.90	0.84, 0.97	-0.05	0.04	0.95	0.88, 1.03
Propensity score	-3.28*	1.27	0.04	0.003, 0.46	-4.70***	1.29	0.01	0.001, 0.11
Treatment = LIT+HIT (ref)								
Treatment = LIT	-0.02	0.34	0.98	0.50, 1.92	0.13	0.34	1.14	0.58, 2.21
Treatment = HIT	0.80*	0.31	2.23	1.21, 4.13	0.55	0.32	1.74	0.93, 3.25
Constant	3.67**	1.23	39.29		3.46**	1.17	31.77	
Subsample of *St* cases	Nagelkerke R^2 = 0.01				Nagelkerke R^2 = 0.06			
Baseline severity †	-0.03	0.03	0.97	0.91, 1.03	0.05	0.03	1.05	0.98, 1.11
Propensity score	-0.51	0.98	0.60	0.09, 4.06	-3.69***	0.92	0.03	0.004, 0.15
Treatment = LIT+HIT (ref)								
Treatment = LIT	-0.14	0.25	0.87	0.53, 1.43	-0.18	0.24	0.83	0.52, 1.34
Treatment = HIT	0.32	0.34	1.38	0.70, 2.70	0.00	0.32	1.00	0.54, 1.86
Constant	1.03	0.71	2.79		1.81**	0.66	6.13	

Notes: Dependent variables in both models are categorical markers for post-treatment remission of symptoms (0 = no remission; 1 = remission); * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; SE = standard error; † the baseline severity measure (either PHQ-9 or GAD-7) entered in each model matched the relevant outcome variable; OR = odds ratios and 95% confidence intervals; *Cx* = complex cases; *St* = standard cases; ref = reference category

Figure 1. Reliable and clinically significant improvement (RCSI) in cases classified as standard (*St*) and complex (*Cx*)

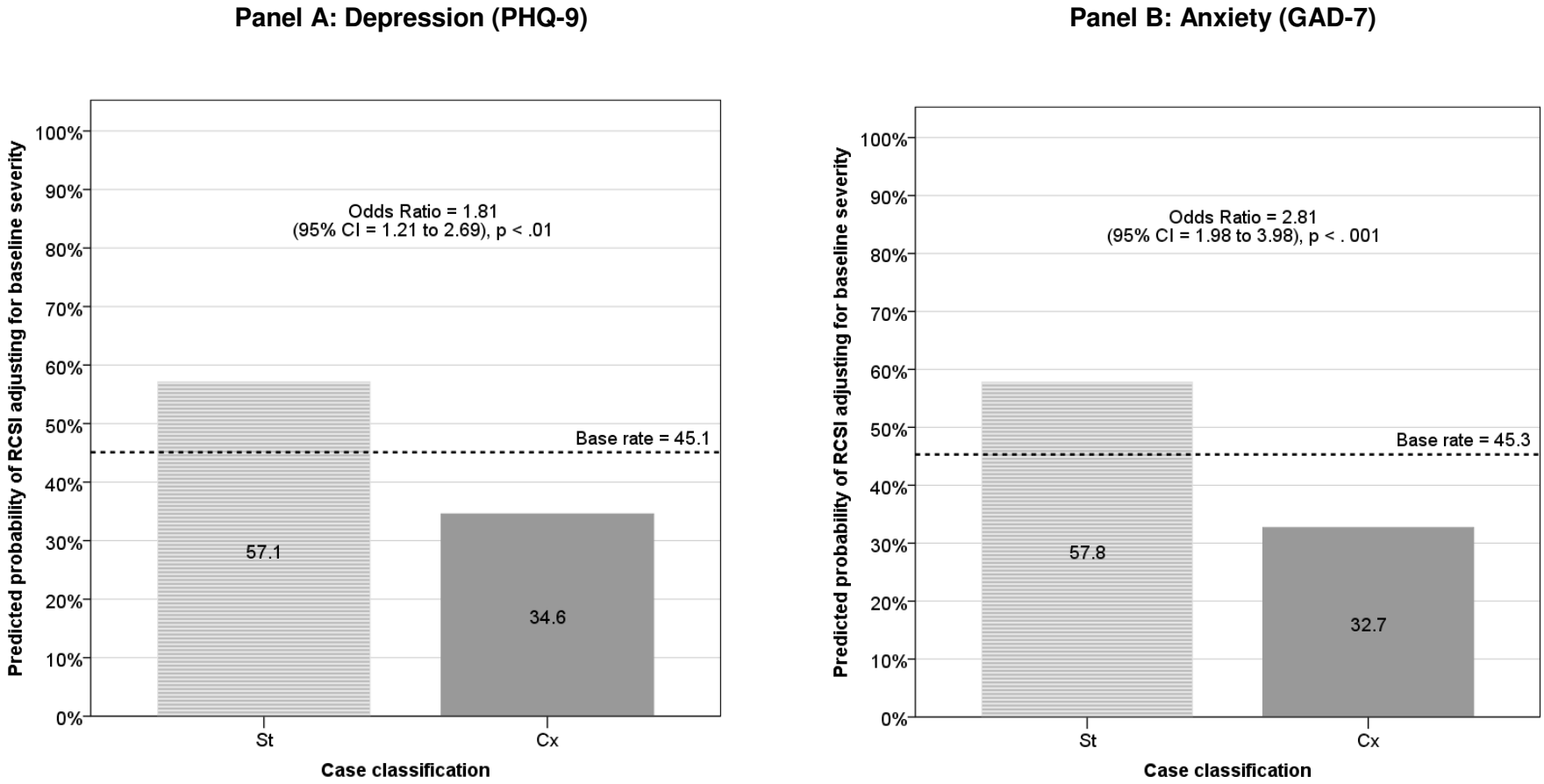

Figure 2. Reliable and clinically significant improvement (RCSI) in cases classified as standard (*St*) and complex (*Cx*) according to treatment pathway

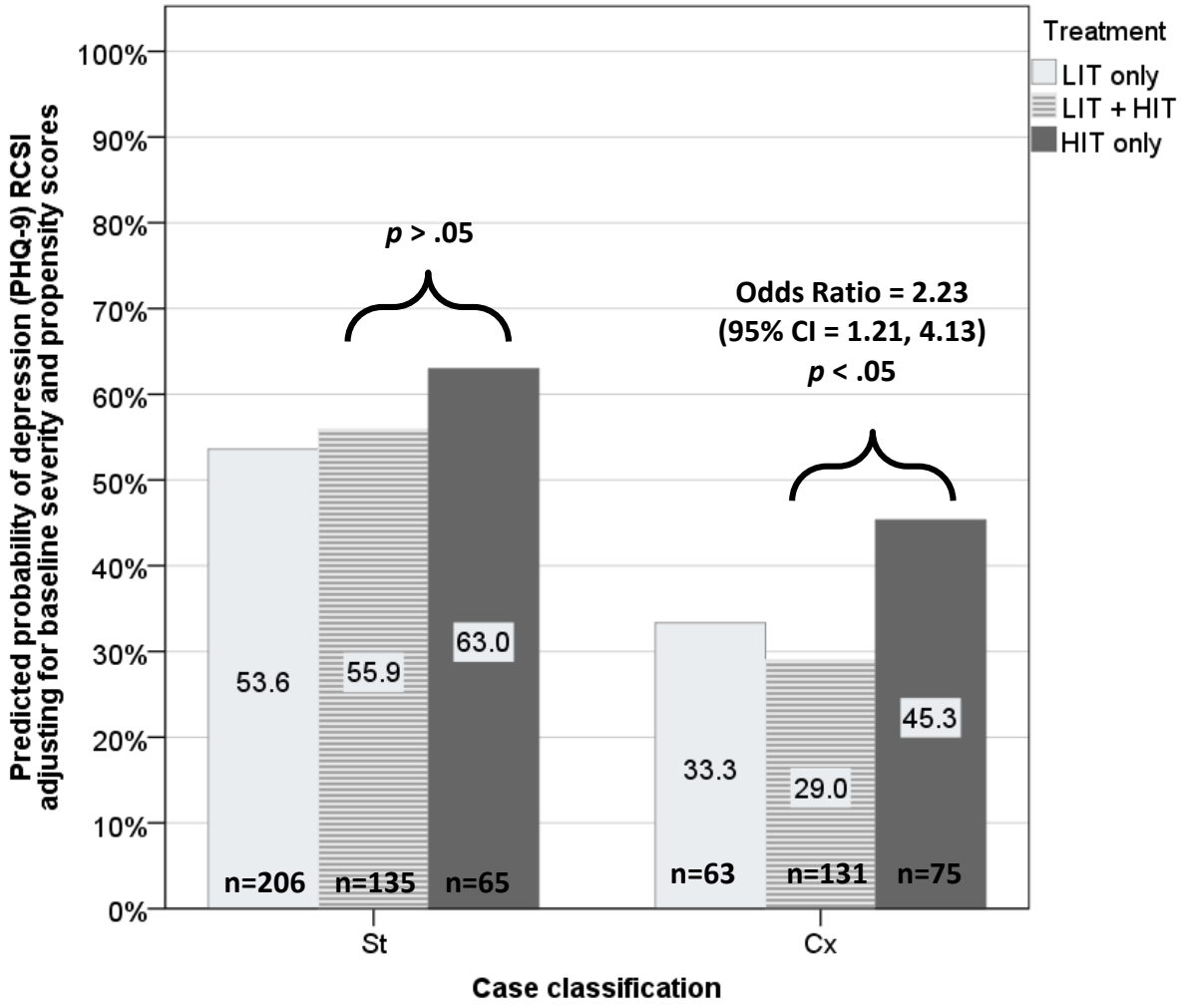

Figure 3. Distribution of hit rates and treatment matching precision (TMaP) scores across 26 therapists