Article:

Lorimer, B., Delgadillo, J. orcid.org/0000-0001-5349-230X, Kellett, S. et al. (1 more author) (2021) Dynamic prediction and identification of cases at risk of relapse following completion of low-intensity cognitive behavioural therapy. Psychotherapy Research, 31 (1). pp. 19-32. ISSN 1050-3307

https://doi.org/10.1080/10503307.2020.1733127

Dynamic Prediction and Identification of Cases at Risk of Relapse Following Completion of Low-Intensity Cognitive Behavioural Therapy

Ben Lorimer[1*], Jaime Delgadillo[2], Stephen Kellett[2] and James Lawrence[3]

[1] Department of Psychology, University of Sheffield, United Kingdom

[2] Clinical Psychology Unit, Department of Psychology, University of Sheffield, United Kingdom

[3] The Behavioural Insights Team, London, United Kingdom

* **Correspondence:** bdlorimer1@sheffield.ac.uk; Floor E, Cathedral Court, 1 Vicar Lane, University of Sheffield, S1 2LT

Abstract

Objective: Low-intensity cognitive behavioural therapy (LiCBT) can help to alleviate acute symptoms of depression and anxiety, but some patients relapse after completing treatment. Little is known regarding relapse risk factors, limiting our ability to predict its occurrence. Therefore, this study aimed to develop a dynamic prediction tool to identify cases at high risk of relapse.

Method: Data from a longitudinal cohort study of LiCBT patients was analysed using a machine learning approach (XGBoost). The sample included *n*=317 treatment completers who were followed-up monthly for 12 months (*n*=223 relapsed; 70%). An ensemble of XGBoost algorithms was developed in order to predict and adjust the estimated risk of relapse (vs maintained remission) in a dynamic way, at four separate time-points over the course of a patient's journey.

Results: Indices of predictive accuracy in a cross-validation design indicated adequate generalizability (AUC range = 0.72-0.84; PPV range = 71.2%-75.3%; NPV range = 56.0%-74.8%). Younger age, unemployment, (non-)linear treatment responses, and residual symptoms were identified as important predictors.

Discussion: It is possible to identify cases at risk of relapse and predictive accuracy improves over time as new information is collected. Early identification coupled with targeted relapse prevention could considerably improve the longer-term effectiveness of LiCBT.

Keywords: Depression; Anxiety; Relapse; Machine Learning; Cognitive Behavior Therapy

Introduction

Depression and anxiety disorders are associated with high rates of relapse (<12 months) and recurrence (≥12 months) after treatment (Bockting, Hollon, Jarrett, Kuyken & Dobson, 2015; Bruce et al., 2005; Burcusa & Iacono, 2007; Hardeveld, Spijker, De Graaf, Nolen & Beekman, 2010; Vervliet, Craske & Hermans, 2013). Patients who complete cognitive behavioural therapy (CBT) tend to have lower relapse rates compared to those who discontinue pharmacological treatment (Cuijpers et al., 2013; Hollon, Stewart & Strunk, 2006; Otto, Smits & Reese, 2005; Vittengl, Clark, Dunn & Jarrett, 2007). Despite this, relapse remains common after CBT for both depression and anxiety. For example, a meta-analysis estimated that 29% of patients with depression who completed CBT with remission of symptoms relapse within one year, with this increasing to 54% within two years when recurrence events are considered (Vittengl et al., 2007). More recently, a longitudinal cohort study also found that many patients relapse after completing the low-intensity version of CBT (i.e. brief, structured, guided psychoeducational self-help interventions based on CBT principles; LiCBT), as applied in the Improving Access to Psychological Therapies (IAPT) programme in England (Ali et al., 2017; Clarke, 2011). Specifically, they estimated that 53% of patients with depression and/or anxiety that complete LiCBT relapse within one year, with a further 13% experiencing a recurrence in the following year (Ali et al., 2017; Delgadillo et al., 2018). Therefore, it is important to understand what risk factors are associated with relapse after CBT, as many patients who complete therapy with remission of symptoms do not maintain their treatment gains long-term.

However, relatively little is known about factors associated with relapse after CBT for depression and anxiety disorders. A recent systematic review identified only two well-supported and replicated predictors of depressive relapse: residual symptoms, and previous number of depression episodes (Wojnarowksi, Firth, Finegan & Delgadillo, 2019). Despite the investigation of over twenty variables across studies included in this review, most studies were grossly underpowered to identify reliable predictors of relapse, especially in samples with a low event base-rate. Furthermore, these studies applied suboptimal analyses to perform variable selection in samples with multiple features, and to deal with multicollinearity when correlated features were examined as potential predictors of relapse. Contemporary machine learning approaches could potentially help to advance our understanding of relapse risk factors, since they are explicitly designed to optimize variable selection and to enhance predictive accuracy by leveraging the prognostic signal across multiple "weak" predictors (Hastie, Tibshirani & Friedman, 2009).

In *supervised* machine learning, patterns are identified in data with the goal of using input variables to predict the values of a target outcome measure (Hastie et al., 2009). Such models are often trained to maximize generalizability to new samples. In contrast to traditional statistical models, such as general linear model (GLM) approaches, supervised machine learning approaches focus on prediction rather than explanation (Yarkoni & Westfall, 2017). This focus provides certain advantages over GLM. For instance, the predictive accuracy of GLM equations is known to be limited when they are developed using relatively small samples and a large number of predictor variables (Iniesta, Stahl & McGuffin, 2016). Furthermore, GLM approaches are prone to *overfitting* (Babyak, 2004; Yarkoni & Westfall, 2017). This refers to situations where a prediction model is overly influenced by the idiosyncrasies of the cases used to develop it (i.e., it incorporates noise that is unique to this sample), and is consequently unreliable at predicting outcomes in new samples (i.e., poor generalizability). Machine learning approaches address overfitting by implementing cross-validation (i.e., training models and then examining their predictive ability in test samples) and regularization methods (i.e., penalizing overly complex models; Hastie et al., 2009).

Machine learning approaches have recently been utilized to address a series of mental health related prediction problems. For example: the prediction of persistent depressive symptoms in older adults (Hatton et al., 2019); the diagnosis of post-traumatic stress disorder (PTSD) three months after a severe injury (Papini et al., 2018); response to pharmacological treatment of depression (Chekroud et al., 2016); and targeted prescription of alternative psychological treatments for depression (Cohen, Kim, Van, Dekker & Driessen, 2019; Delgadillo & Gonzalez Salas Duhne, 2020). Set against this backdrop of emerging applications of machine learning in mental health, the present study aimed to (a) identify prognostic indicators of relapse after LiCBT and (b) to use this information to develop a relapse prediction tool that could be used to guided relapse prevention interventions in psychological services.

Method

Design, Setting and Interventions

This study analyzed data from the WYLOW study (Delgadillo et al., 2018), which was a naturalistic, prospective, longitudinal cohort study. *N*=439 LiCBT patients with remission of depression and anxiety symptoms after treatment were recruited into the study from a psychological therapy service in West Yorkshire, England, which was part of the IAPT programme (Clark, 2011). Participants were followed-up on a monthly basis after treatment completion, for up to 24 months. The overall objective of the study was to quantify relapse and recurrence rates after routinely-delivered LiCBT. Approval for the study was obtained from the NHS Health Research Authority and an independent ethics committee (Ref: 12/YH/0095).

Participants in the WYLOW study completed low intensity guided self-help interventions based on principles of CBT, which was consistent with national clinical guidelines and competency frameworks (National IAPT Team, 2015; National Institute for Health and Clinical Excellence, 2011). These interventions were delivered in one-to-one sessions, group settings, or via computerized CBT programmes with adjunct telephone support. LiCBT in this service was highly standardized, protocol-driven, and delivered by qualified psychological wellbeing practitioners (PWPs) who practiced under regular clinical supervision (weekly, or every other week).

LiCBT patients who completed treatment with sub-clinical depression and anxiety symptoms (see measures section) were eligible for inclusion and were recruited within one month of their last planned treatment session. Participants were contacted by independent researchers on a monthly basis and prompted to complete depression and anxiety questionnaires via email, telephone or postal survey (based on participants' preferences). Participants remained in the study until they met one of three criteria: (1) they were classified as having relapsed (as defined below); (2) they had failed to respond to two consecutive monthly assessments (lost to follow-up); or (3) their responses indicated that they had remained in remission throughout the follow-up period. Participants that were classified as having relapsed were provided with self-help information and supported to re-engage with mental healthcare services. Further details about the study setting, design, recruitment process, data collection and primary findings are reported elsewhere (Ali et al., 2017; Delgadillo et al., 2018).

Although the WYLOW study collected data for up to 24 months post-treatment, this study only analyzed data from the initial 12 month period. There were two reasons for this: 1) this study's purpose was to predict relapse (<12 months), not recurrence (≥12 months); and 2) the inclusion of recurrence events in the analysis would have limited the number of cases that remained in remission throughout the follow-up period to a point where the development of predictive algorithms would have been unfeasible.

Participants

The present study sample analyzed data for a subsample of *n*=317 cases from the WYLOW study, excluding *n*=122 cases that were lost to follow-up, and for whom the actual outcome of interest (relapsed or maintained remission) was unknown. Within the first half of the follow-up period, *n*=84 participants had relapsed after one month, *n*=125 after two months, *n*=155 after three months, *n*=171 after four months, *n*=178 after five months, and *n*=189 after six months. Within the second half, *n*=196 cases had relapsed after seven months, *n*=202 after eight months, *n*=208 after nine months, *n*=214 after 10 months, *n*=218 after 11 months, and *n*=223 after one year. Therefore, a total of *n*=223 (70%) cases in the present study sample relapsed during the 12-month follow-up period, while *n*=94 remained in remission.

The three most common primary presenting problems recorded in clinical records for the above cases were mixed anxiety and depression (32%); depressive episode (22%); and generalized anxiety disorder (19%). Other mental health problems (e.g., panic disorder, obsessive compulsive disorder) were recorded as the primary problem for less than 3% of cases each (*n*<10). The sample was characterized by a majority of female patients (59%) from a white British background (90%), with a mean age of 43 (*SD* = 14.8). Approximately 10% were unemployed at the start of treatment, 12% were unemployed at the end, 11% had a self-reported disability, and 26% had a self-reported long-term medical condition (e.g., asthma, diabetes, chronic pain, etc.). Patients accessed a mean of seven LiCBT sessions (*SD* = 2; range = 2-16).

Measures

Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, & William, 2001): A screening tool for major depression containing nine items. Each item assesses how often a specific symptom is experienced over a two-week period and is measured on a 0-3 scale (i.e. 0 = “not at all”, 3 = “nearly every day”). Responses are summed to calculate an overall severity score (range = 0-27). A cut-off of ≥10 is recommended to detect clinically significant depression symptoms (Kroenke et al., 2001).

Generalized Anxiety Disorder scale (GAD-7; Spitzer, Kroenke, Williams, & Love, 2006): A screening tool for anxiety disorders containing seven items. Similar to the PHQ-9, each item assesses how often a symptom is experienced by over a two-week period and the same scale (i.e. 0-3) is used. Responses are summed to calculate an overall severity score (range = 0-21), and a cut-off of ≥8 is used for the detection of an anxiety disorder (Kroenke, Spitzer, Williams, Monahan, & Löwe, 2007).

In order to monitor deterioration over time, a reliable change index of ≥5 for both the PHQ-9 (McMillan, Gilbody & Richards, 2010) and the GAD-7 (Richards & Borglin, 2011) was applied in this study.

Work and Social Adjustment Scale (WSAS; Mundt, Marks, Shear & Greist, 2002): A questionnaire that assesses the impact of a mental health problem on five life domains (work, home management, social life, leisure activities, family and relationships). The impact on each domain is measured on 0-8 scales (i.e. 0 = "no impairment", 8 = "severe impairment"), and the five scores are summed to derive an overall score of functional impairment (range = 0-40).

A series of other clinical, demographic and treatment variables were also available. These are described in detail by Ali et al. (2017) and summarized in Table 1.

Primary Outcome

For a patient to be classed as having relapsed, the following two criteria must have been met: (a) at least one of the symptom measures (PHQ-9 or GAD-7) was above the clinical cut-off at a monthly follow-up review; and (b) the measure was indicative of statistically reliable deterioration relative to the score observed at the final treatment session (based on the reliable change index). Patients who scored above clinical cut-offs during the follow-up period (criterion a) but did not display statistically reliable deterioration (criterion b), were not classed as having relapsed and continued to be monitored on a monthly basis. Fourteen of the 94 participants who remained in remission after 12 months had met criterion a but not criterion b at least once during the follow-up period.

Analysis

A supervised machine learning approach was applied to develop an ensemble of four prognostic models that aimed to predict relapse in a dynamic (longitudinal) way during the treatment and follow-up phases. Each successive model was trained using additional information that became available over time, thus enabling the overall ensemble of algorithms to "learn" to "adjust" the prognosis in four steps. Model 1 used information only known at the time of pre-treatment assessments; Model 2 learned from information known at the final treatment session: Model 3 included information available one month after the end of treatment; and Model 4 included information available three months after the end of treatment. These models were trained to output a predicted classification for each patient (1 = *will relapse within 12-months of completing treatment*; 0 = *will have sustained remission for up to 12 months*), at each relevant time-point described above. Importantly, the dynamic nature of this overall modelling approach means that a case predicted at one time-point (e.g., last treatment session) to maintain remission could later be classed as at risk of relapse based on newly available information (e.g., if depression symptoms increase at 3-months follow-up).

No models were developed that used information collected after three months of follow-up as predictors of relapse, as these models would have been too underpowered due to the gradual decrease of the overall sample during the follow-up period. Indeed, due to this gradual decrease, Models 3 and 4 (i.e., predictions occurring during follow-up) were developed using smaller overall samples than Models 1 and 2 (i.e., predictions occurring before follow-u). The variables tested as potential predictors in each model are summarized in Table 1, and further described in Supplementary Materials A (glossary of terminology). The potential predictors ranged from baseline characteristics to factors related to treatment response.

Machine learning approach. We applied extreme gradient boosting (XGBoost; Chen & Guestrin, 2016), as implemented in the R package *xgboost* (Chen et al., 2019), to develop the four prognostic models. The XGBoost package implements gradient-boosted decision trees (Friedman, 2001). Gradient boosting is an ensemble machine learning approach, meaning that it combines several simple decision trees, that are each individually poor at predicting an outcome, to develop one collective, stronger prognostic model. Specifically, when each successive decision tree is trained, subsequent trees learn from the 'residuals' or 'errors' from the previous modelling attempt. Therefore, each gradient boosting model is developed iteratively in multiple steps, with each step placing more priority in making accurate predictions for those individuals for whom prior iterations made poor predictions.

Each single decision tree consists of 'branches', which represent logical structures. Each study participant has a 'path' through a decision tree that is determined by their data; for dichotomous features (e.g., employed/unemployed), paths diverge for yes/present or no/absent responses, while for continuous features (e.g., baseline depression severity), paths diverge based on empirically derived cut-offs. Different cut-offs are examined by the algorithm until a final cut-off is selected based on the optimization of predictive accuracy. Each path terminates at a 'leaf', which represents a predicted probability weight for the outcome of interest. The final predicted classification (relapse/sustained remission) for each case is derived by summing their individual probability weights across all decision trees in the ensemble. An example decision tree, extracted from a model developed in this study, is illustrated in Figure 1.

XGBoost has several advantages over conventional regression modelling (Chen & Guestrin, 2016). For instance, missing data does not need to be imputed or excluded, as the algorithm can use missing data as an informative splitting criterion. Moreover, continuous variables do not need to be transformed, as path divergences are established based on cut-offs determined by the algorithm. Gradient boosting allows for complex, non-linear interactions among categorical and continuous variables to be incorporated into predictive models. In addition, XGBoost can also minimize overfitting through the implementation of L1 (LASSO) and L2 (Ridge) regularization procedures, and cross-validation loops.

Parameters. For the development of each of the four models, hyperparameters were set for the implementation of XGBoost. One parameter, *eta*, controls the learning rate of the algorithm, and can minimize overfitting by shrinking the weights of variables after each boosting step, thus making the boosting process more conservative. *Eta* can range from 0-1, and was manually set in this study at 0.1. Another parameter manually set was the *maximum depth* of a tree, which was set at 3. This meant that during the development of each individual decision tree, there can only be a maximum of two further 'branch splits' along a path after the initial split. Trees with larger depths are more complex and allow models to learn relations that are highly specific to a particular sample. Therefore, having a smaller maximum depth for trees minimizes overfitting. We used the XGBoost default values for regularization, where *alpha* (L1) is set to 0 and *lambda* (L2) is set to 1. Therefore, no LASSO regularization was applied in the development of predictive models, while a Ridge penalty was applied.

The evaluation metric used to assess the predictive ability of models was the area under the receiver operating characteristic curve (AUC). This measure assesses how capable models are at distinguishing between binary outcome classes (Bradley, 1997). The metric ranges from 0-1, with 1 representing a perfect predictive model and 0.5 indicating that predictions are no better than chance. The AUC statistic has been recommended as an appropriate evaluation metric for healthcare research, due to the clinical utility of model predictions being more important than statistical power in these contexts (Moons et al., 2015).

Internal cross-validation. Internal cross-validation loops were used to develop an XGBoost ensemble model. To achieve this, the overall sample was partitioned into five subsamples using stratified randomization so that each subsample had approximately the same rate of relapse cases. Each loop would use 4/5 of the data as a training sample and 1/5 as a test sample. Each subsequent decision tree was trained using this five-fold cross-validation process, which continued until a 'stopping rule' was triggered if the addition of new trees resulted in overfitting with reference to the AUC observed in the test sample. The cross-validation process therefore identified the highest number of decision trees an XGBoost model can incorporate before overfitting to the dataset.

Variable selection. Implementation of XGBoost produces a model with: (1) a final AUC statistic that assesses its 'out-of-sample' performance; and (2) information regarding the variables that the model has selected for use in making its predictions. It is desirable to train a model with the highest possible AUC index, and with the lowest number of variables selected (i.e., a more parsimonious model). If superfluous variables are made available to the algorithm, then given the large number of decisions the model must take, it will occasionally happen across a subset of data for which these extra variables appear to be predictive but do not extend to the test set. Thus, we can improve the model by iteratively testing whether removing a variable from the dataset improves the AUC index.

We therefore performed variable selection through the additional incorporation of 'leave-one-variable-out loops'. One loop involved additional XGBoost models to be developed, with the same number of models being developed as there were variables input into the base model (*k*). For each of the models produced in this step of model development, one different variable was excluded from analysis and therefore not incorporated into the model (i.e., *k*-1). The model with the highest associated AUC statistic was considered the best model in this step of the process, and kept for further development. Following this, another leave-one-out loop was implemented using the variables included in the best model from the previous loop, and this process continued until there was only one variable remaining. The model with the highest AUC across all of the loops was considered the best overall model developed. This incorporation of leave-one-out loops ensured that the final selected model had the highest AUC possible with the lowest number of variables involved.

Model evaluation. The analytical process described thus far was followed to develop each of the four prognostic models (Models 1-4). Using each of these models, predicted probabilities of relapse for each patient were calculated at four time-points during their pathway through treatment and follow-up. Probabilities ranged from 0-1 with higher scores indicating greater probability of relapse, and a patient was predicted to relapse when their calculated probability was greater than 0.5. Conventional performance metrics were calculated to assess the predictive value of each model when tested out-of-sample (in the 1/5 test set): accuracy; sensitivity; specificity; positive predictive value (PPV); and negative predictive value (NPV).

Exploring individual predictors. Before individually exploring the predictors identified by each model, a sensitivity analysis was conducted in which the four final models were reanalyzed with the application of L1 regularization (*alpha*=0.5). In contrast to L2 regularization, L1 regularization is capable of shrinking the coefficients of variables to be exactly zero, thus eliminating likely irrelevant variables from the model (Hastie, Tibshirani & Friedman, 2009). Therefore, this sensitivity analysis was conducted to explore if the addition of a LASSO penalty identified any variables in the models as being spurious or potential false positives, and thus ensure that only the most important individual predictors were explored.

Following this, the variables that were (1) determined to be important by at least one of the four final XGBoost models, and (2) not deemed spurious by the addition of L1 regularization, were explored in terms of their associations with relapse using relative importance metrics, partial dependence plots, and descriptive statistics. Relative importance is a metric that indicates how useful a predictor was in the construction of boosted decision trees within a model, relative to the other predictors used by the model. It is calculated for each predictor by assessing the degree to which a model's predictive ability improves when key decisions within decision trees use that specific variable (Hastie, Tibshirani & Friedman, 2009). Meanwhile, partial dependence plots demonstrate the relationship between a predictor and the outcome by aggregating all of the decisions in all of the decision trees in the model which select on that predictor (Friedman, 2001). Therefore, they illustrate the relationship that the XGBoost model judges is directly attributable to that predictor, after partialling out the effects of other predictor variables, even if they might be correlated with the predictor of interest.

Results

Performance of XGBoost Models

The predictive value of the four models when evaluated on the test set is summarized in Table 2. The AUC became larger as more information was available to the models over time. Model 1, which only used pre-treatment assessment information, had the lowest AUC (0.72), while Model 4, which used information available up to the third month of post-treatment follow-up, had the highest AUC (0.84). Other metrics of predictive value displayed in Table 2 indicated that Models 1 and 2 had highly similar performances. Both models had high sensitivity (91% and 90.1% respectively), PPV (74.9% and 75.3% respectively), and accuracy (both 72.2%). However, both models also had low specificity (27.7% and 29.8% respectively) and NPV (56.5% and 56% respectively). Model 3 was found to have greater specificity (52.1%) and NPV (63.6%), however it also had the lowest accuracy (68.7%) and PPV (71.2%) of all four models, and the second lowest sensitivity (79.9%). Overall, Model 4 was found to have the best relative balance in performance indices across all four models. The distributions of the probabilities of relapse predicted by each model for each patient can be seen in Figure 2.

Important Predictors

Of the 141 variables input as potential predictors of relapse across all four models, a total of 42 predictors were deemed important. Model 2 was the most complex, selecting 13/32 input variables as important predictors. Models 3 and 4 each identified 11 predictors as important, with 39 and 53 variables initially being entered into each model respectively. Finally, Model 1 was the simplest, selecting only seven of the 17 variables initially entered into the model. The predictors identified as being important by each model, and the relative importance of each variable for the prediction of relapse, are displayed in Table 3.

The sensitivity analyses (i.e., re-analysis of the final models with L1 regularization applied) yielded models with similar but slightly lower cross-validated AUC indices. They also retained the same number of important predictors for Models 2, 3, and 4. However, the sensitivity analysis for Model 1 did not select the variable denoting missing vs. complete response to the *treatment expectancy* variable as an important predictor, indicating that this variable is likely to be spurious. Therefore, this variable was not explored further. For the sake of parsimony, the important predictors can be grouped according to four categories: demographics; baseline clinical features; treatment process features; and residual symptoms. Partial dependence plots can be found in Supplementary Materials B.

Demographics. Demographic predictors leveraged a combined importance (i.e. the sum of relative importance metrics of each predictor of the same type) of 63% in Model 1, however this decreased when subsequent models had access to further information. Demographic predictors only had 30% combined importance for Model 2, 37% for Model 3, and 11% for Model 4. Younger age, unemployment (pre-treatment and post-treatment), disability, and female gender featured as important predictors across at least two models (the latter two features had relatively little importance). Unemployment was a particularly important risk factor, with all 31 participants who were unemployed before starting treatment relapsing within 12 months, and 37/39 participants who were unemployed at the end of treatment relapsing.

Baseline clinical features. Pre-treatment clinical features had a combined importance of 37% for Model 1, 15% for Model 2, and 14% for both Models 3 and 4. The following risk factors were selected as important by at least two models: use of psychotropic medication at the start of treatment, and family history of mental health problems. Both features had comparatively small importance indices. Six other baseline clinical features were deemed important by only one model each: higher baseline WSAS severity, higher baseline GAD-7 severity, previous treatment episodes, low expectancy of treatment, diagnosis of an affective or an anxiety disorder (*vs* a 'mixed' or 'other' diagnosis), and a chronic mental health problem.

Treatment process features. Treatment process features were only relevant to Models 2-4, since Model 1 only included pre-treatment information. The combined importance of these features was 40% in Model 2, 0% in Model 3 and 11% in Model 4. The following risk factors were selected as important in two models: early treatment response (higher improvement by session 3) in the PHQ-9 measure, a more linear (versus nonlinear) trajectory of improvement in the WSAS measure during treatment, and a more nonlinear (versus linear) trajectory of improvement in GAD-7. However, early treatment response had less importance compared to the latter two features, while the findings for the GAD-7 measure were relatively inconsistent and less clear from the partial dependence plots than the findings for the WSAS measure.

Residual symptoms. The combined importance of residual symptoms (sub-threshold symptom scores close to the diagnostic cut-offs) post-treatment was only 15% for Model 2, however this increased to 48% and 47% for Models 3 and 4 respectively. Higher post-treatment scores and increases in PHQ-9, GAD-7 and WSAS scores between the last treatment session and follow-up assessments were selected as important predictors across at least two models.

Discussion

This study demonstrates the utility of a data-driven prognostic method capable of identifying cases at high risk of relapse after the completion of LiCBT. Previous studies have used traditional statistical models to predict relapse after routinely delivered CBT interventions (see review by Wojnarowski et al., 2019). The present study represents a novel and considerable advance in this line of research through the application of a machine learning ensemble of models that dynamically predict and adjust expected prognoses at multiple time-points during the course of a patient's treatment journey. The AUC statistics for these models ranged from 0.72-0.84, with predictive accuracy tending to improve over time, when models were trained using more information. The best performance (AUC = 0.84) was observed for Model 4, which used information collected before treatment, at the end of treatment, and up to the third month of post-treatment follow-up. This is likely to be explained by the use of the maximum number of available predictors, but also due to the temporal proximity of the prediction time-point to the target outcome. A close examination of performance metrics across these models indicated that positive predictive values were superior to negative predictive values. Put simply, these models performed well at identifying cases at high risk of relapse, but they were less capable of accurately identifying patients that would maintain remission of symptoms during the 12-month follow-up period. However, this was not the case for Model 4, which had both high PPV (74.6%) and high NPV (74.8%).

In terms of predictor categories, demographic and baseline clinical features exhibited high importance in the early models. However, this decreased in later models when additional variables related to the therapeutic process and residual symptoms were also included. Residual symptoms had by far the highest combined importance of the four categories in Models 3 and 4, having the greatest temporal proximity to the target outcome. In contrast, although therapeutic process features had high combined importance in Model 2, they had no importance in Model 3, and little importance in Model 4. This inconsistent pattern may indicate limited robustness of the findings related to these features, and thus further research is required.

Specific patient features that were most informative for the prediction of relapse included younger age, unemployment, treatment response trends (linear versus nonlinear), and the presence of residual symptoms. Young age was identified as a risk factor, contrary to prior studies in this area (Evans et al., 1992; Lincoln et al., 2005; Heldt et al., 2011). However, previous studies had less than *n*=20 relapse cases in their samples, indicating that they were grossly underpowered to detect even a large effect size. Indeed, a relatively larger study (*n*=31 recurrence cases) conducted by Fava et al. (2001), found a significant effect of young age on recurrence of panic disorder with agoraphobia after exposure-based treatment.

Unemployment was a highly important risk indicator in the current sample, with all 31 participants who were unemployed at the start of treatment being classed as having relapsed within 12 months. In addition, the majority of these relapse cases occurred in the first month of follow-up, indicating that the time-to-relapse was accelerated for unemployed patients. Interestingly, cases that were unemployed before treatment did not entirely overlap with cases unemployed at the end of treatment, yet unemployment at both time-points was identified as a risk factor. This indicates that any experience of unemployment during the therapy process increases a patient's risk of relapse. This adds further credibility to the wealth of evidence concerning the harmful effects of unemployment on mental health (Waddell & Burton, 2006).

Patterns of change during treatment were also informative for relapse prediction. Patients who displayed a rapid reduction of depression symptoms during the first 3 sessions had higher risk of relapse. Although early response is a well-established predictor of end-of-treatment outcomes in psychotherapy, a recent systematic review suggested that this could be at least partly influenced by a quasi-placebo effect, since rapid response is also evident in placebo control groups (Beard & Delgadillo, 2019). It is, therefore, possible that patients with early response may mostly be improving due to a *remoralisation effect* (Howard, Lueger, Maling, & Martinovich, 1993), but remain at risk of relapse because their learning of coping strategies has not been sufficiently developed or consolidated. A linear improvement in functional impairment over the course of treatment also appeared to predict relapse. Previous studies on the relationship between treatment duration and clinical outcomes suggest that domains such as interpersonal and social functioning tend to improve at later stages of treatment (see review by Robinson, Delgadillo, & Kellett, 2019). Therefore, a gradual and linear improvement in functioning in relatively brief interventions (less than 8 sessions) possibly indicates that such changes may possibly be due to the temporary resolution of life problems and current stressors rather than therapy, leaving patients at risk of relapse when future stressors arise if they have not had an opportunity to consolidate coping skills. Findings concerning trajectories of anxiety symptoms were mixed and therefore an interpretation of these patterns would be premature at this stage and warrants further investigation.

The identification of residual symptoms after treatment as an important predictor of relapse is consistent with previous research (Bockting et al., 2015; Wojnarowski et al., 2019). One proposed explanation is that residual symptoms represent the persistence of the underlying disorder, albeit in a milder form (Paykel, 2008). Another possibility is that the presence of residual symptoms is associated with a greater vulnerability to experiencing future life events as intolerably stressful and de-stabilizing, thus increasing the likelihood of relapse (Harkness, Theriault, Stewart & Bagby, 2014).

Limitations and Future Research

A limitation concerning the study sample concerns *class-imbalance* (i.e., the number of observations belonging to each outcome class was not the same), since relapse cases in this sample outnumbered remission cases (223 versus 94). Class-imbalance can undermine the accuracy of machine learning algorithms, with poorer predictive accuracy for patients in the minority class (Lopez, Fernandez, Garcia, Palade & Herrera, 2013). Indeed, Models 1 and 2, which were both developed using the total class-imbalanced sample, were highly effective at classifying relapse cases correctly (high PPV), but less capable of classifying remission cases accurately (low NPV). In contrast, Models 3 and 4, which were developed using smaller, more balanced samples, had an improved ability to accurately predict remission cases (higher NPVs). This pattern (Figure 2) may be influenced by class-imbalance bias. Therefore, Model 4 being superior at prediction compared to the other models may not be because prediction of relapse is most accurate during follow-up, but rather an artifact of class-imbalance.

One potential solution for attenuating class-imbalance bias is to adopt resampling techniques. One such technique is over-sampling, in which the skewed distribution is addressed by generating new minority class observations (e.g. synthetic minority over-sampling technique [SMOTE]; Chawla, Bowyer, Hall & Kegelmeyer, 2002). A different technique is under-sampling, in which observations from the majority class are discarded instead (e.g. Random Undersampling; Tahir, Kittler, Mikolajczyk & Yan, 2009). The technique that is most appropriate may depend on the form of class-imbalance exhibited by a sample (Haixiang et al., 2017). Future replications of this study that also use class-imbalanced samples should consider using resampling techniques to limit this bias and improve model development. In the current study, we chose not to apply such methods to minimize modelling complexity since the primary goal was to test the accuracy of dynamic prediction models, but also because it is instructive to show the effects of class-imbalance so as to inform future research.

Another important limitation is that this study only applied an internal cross-validation loop, with the models being evaluated on subsamples that were also involved in the training of the models. Although internal validation can provide an indication of how models may generalize to new samples, a more robust method to determine out-of-sample generalizability is to apply an additional external validation in a statistically independent sample that was not involved in model development (Steyerberg, Bleeker, Moll, Grobbee & Moons, 2003). This can be done using a 'holdout' subsample of the overall original sample, which was not possible in this study due to the limited sample size, and/or using a sample of participants from a different population altogether (but for whom the same variables are available). Applications of machine learning algorithms using this rigorous external cross-validation process in mental healthcare research have shown that out-of-sample predictions tend to be valid and clinically useful, but less accurate in new samples (Chekroud et al., 2016; Delgadillo & Gonzalez Salas Duhne, 2020; Delgadillo, Huey, Bennett, & McMillan, 2017; Leighton et al., 2019).

There were additional limitations related to the population investigated and data collection methods used. For instance, loss-to-follow-up in the primary study sample meant that data from 122 participants were not taken into account in the development of the present models. Furthermore, participants only received LiCBT interventions, so these models may not apply to patients who receive high-intensity CBT. In contrast to high-intensity CBT, LiCBT interventions are brief, highly manualized, non-specialist, and driven by the use of highly structured psychoeducational workbooks (Bennett-Levy, Richards & Farrand, 2010). Due to these differences, it is plausible that there are differences between the two therapeutic formats in terms of the processes that underlie relapse. Therefore, further research needs to explore any potential differences between LiCBT and high-intensity CBT in terms of predicting relapse, and in terms of the individual risk factors associated with relapse. Another limitation of this study was the lack of an assessment of treatment fidelity or competency, thus making it impossible to know if PWPs had adhered to LiCBT treatment protocols. A recent study has developed and validated two measures for the purposes of assessing LiCBT competencies (Kellett et al., 2020). These measures could be applied in future research.

Future research into the prediction of relapse should also investigate predictors that were not considered in this study. For example, previous research has indicated that the experience of stressful life events during follow-up increases the risk of relapse for both depression and anxiety (Harkness et al., 2014; Heldt et al., 2011). Use of experience sampling methods using mobile-phone and passive sensing technology could be fruitful to track life events and/or physiological and subjective responses to such events.

Clinical Implications

This study highlights the need for relapse prevention to be a core component of low-intensity CBT interventions. Although relapse prevention is considered an important aspect of LiCBT (Rodgers et al., 2012), the lack of treatment fidelity measures being applied means we do not know if this essential component of treatment is being delivered. Furthermore, this study also highlights the importance of providing relapse prevention interventions. Although offering these interventions incurs a financial cost, the effective prevention of relapse has been argued to save money in the long-term, due to a reduction in the ‘revolving door’ phenomenon where patients continually return for further treatment (Scott, Palmer, Paykel, Teasdale & Hayhurst, 2003; Wojnarowski et al., 2019). Furthermore, the development of prognostic algorithms using machine learning approaches potentially allows for patients at risk of relapse to be identified and targeted with maintenance interventions. If such algorithms were to be implemented in psychological services, it is important that models can effectively identify both patients who are vulnerable to relapse, and those who are not. This is to ensure that the prescription of these interventions is provided to patients that are in need of them, but also performed in a cost-effective manner (i.e. not offered to every patient who successfully responds to treatment). Therefore, Models 1 and 2 may be inappropriate for application in clinical contexts due to these models possessing low NPVs, although these models could potentially be improved with further research that utilizes larger, class-balanced samples. In contrast, Model 4 would be the most appropriate model for use in clinical contexts, due to it possessing a good balance between high PPV and NPVs. However, Model 4 is limited by not being able to predict relapses that occur in the first three months of follow-up. Model 4 possessing the greatest predictive power may indicate that follow-up appointments improve our ability to understand and predict relapse. Incorporating structured follow-up into clinical practice would allow for prognostic algorithms such as Model 4 to be applied by clinicians, while also enabling larger datasets to be constructed and thus allow more improved predictive models to be developed.

Conclusions

This study demonstrates that it is possible to identify cases at high risk of relapse using routinely available data, with considerable accuracy even before the start of treatment, and predictive accuracy improving as new information becomes available over time. This dynamic prognostic system could be used in routine care to identify 'high risk' cases that could be offered evidence-based relapse prevention interventions in a targeted way, thus making best use of limited resources in publicly funded mental health services.

Disclosure of Interest: The authors report no conflict of interest.

Data Availability Statement: In line with the requirements of the ethics review board for this study, requests for access to data are to be made in writing to the principal investigator and data custodian for the WYLOW study: jaime.delgadillo@nhs.net.

Funding: This work was supported by the Economic and Social Research Council [grant number ES/R500975/1], and by a travel grant awarded by MindLife UK Ltd.

References

Ali, S., Rhodes, L., Moreea, O., McMillan, D., Gilbody, S., Leach, C., . . . Delgadillo, J. (2017). How durable in the effect of low intensity CBT for depression and anxiety? Remission and relapse in a longitudinal cohort study. *Behaviour Research and Therapy, 94*, 1–8.

Babyak, M.A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411-421.

Beard, J.I.L., & Delgadillo, J. (2019). Early response to psychological therapy as a predictor of depression and anxiety treatment outcomes: A systematic review and meta-analysis. *Depression and Anxiety*. doi: 10.1002/da.22931

Bennett, J., Richards, D. A., & Farrand, P. (2010). Low intensity CBT interventions: A revolution in mental health care. In J. Bennett-Levy, D. A. Richards, P. Farrand, H. Christensen, K. M. Griffiths, D. J. Kavanaugh, . . . C. Williams (Eds.), *Oxford guides in cognitive behavioural therapy: Oxford guide to low intensity CBT interventions* (pp. 3-18). New York, NY, US: Oxford University Press

Bockting, C.L., Hollon, S.D., Jarrett, R.B., Kuyken, W. & Dobson, K. (2015). A lifetime approach to major depressive disorder: The contributions of psychological interventions in preventing relapse and recurrence. *Clinical Psychology Review, 41,* 16-26.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30(7),* 1145-1159.

Bruce, S.E., Yonkers, K.A., Otto, M.W., Eisen, J.L., Weisberg, R.B., Pagano, M., . . . & 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. *American Journal of Psychiatry, 162,* 1179-1187.

Burcusa, S.L. & Iacono, W.G. (2007). Risk for recurrence in depression. *Clinical Psychology Review, 27,* 959-985.

Chawla, N.V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16,* 321-357.

Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H. . . . Corlett, P.R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *Lancet Psychiatry, 3*, 243-250.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., . . . Li, Y. (2019). Xgboost: Extreme Gradient Boosting. R package version 0.82.1. https://CRAN.R-project.org/package=xgboost

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.

Cohen, Z.D., Kim, T.T., Van, H.L., Dekker, J.J. & Driessen, E. (2019). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive behavioural versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research, 11*, 1-14.

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A. & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Canadian Journal of Psychiatry, 58,* 376–385.

Delgadillo, J. & Gonzalez Salas Duhne, P. (2020). Targeted prescription of cognitive-behavioral therapy versus person-centred counselling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology, 88,* 14-24.

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology, 85*, 835-853.

Delgadillo, J., Rhodes, L., Moreea, O., McMillan, D., Gilbody, S., Leach, C., . . . Ali, S. (2018). Relapse and recurrence of common mental health problems after low intensity cognitive behavioural therapy: The WYLOW longitudinal cohort study. *Psychotherapy and Psychosomatics, 87*, 116-117.

Evans, M.D., Hollon, S.D., DeRubeis, R.J., Piasecki, J.M., Grove, W.M., Garvey, M.J. & Tuason, V.B. (1992). Differential relapse following cognitive therapy and pharmacotherapy for depression. *Archives of General Psychiatry, 49,* 802–808.

Fava, G.A., Rafanelli, C., Grandi, S., Conti, S., Ruini, C., Mangelli, L. & Belluardo, P. (2001). Long-term outcome of panic disorder with agoraphobia treated by exposure. *Psychological Medicine, 31,* 891-898.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29(5)*, 1189-1232.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73,* 220-239.

Hardeveld, F., Spijker, J., De Graaf, R., Nolen, W.A. & Beekman, A.T.F. (2010). Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatra Scandinavica, 122,* 184-191.

Harkness, K.L., Theriault, J.E., Stewart, J.G. & Bagby, R.M. (2014). Acute and chronic stress exposure predicts 1-year recurrence in adult outpatients with residual depression symptoms following response to treatment. *Depression and Anxiety, 31*, 1-8.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* New York, NY: Springer.

Hatton, C.M., Paton, L.W., McMillan, D., Cussens, J., Gilbody, S. & Tiffin, P. (2019). *Journal of Affective Disorders, 246*, 857-860.

Heldt, E., Kipper, L., Blaya, C., Salum, G.A., Hirakata, V.N., Otto, M.W. & Manfro, G.G. (2011). Predictors of relapse in the second follow-up year post cognitive-behavior therapy for panic disorder. *Revista Brasileira de Psiquiatria, 33(1),* 23-29.

Hollon, S.D., Stewart, M.O. & Strunk, D. (2006). Enduring effects for cognitive behaviour therapy in the treatment of depression and anxiety. *Annual Review of Psychology, 57,* 285-315.

Howard, K. I., Lueger, R. J., Maling, M. S., and Martinovich, Z. (1993). A phase model of psychotherapy outcome: causal mediation of change. *Journal of Consulting and Clinical Psychology, 61*, 678-685.

Iniesta, R., Stahl, D. & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine, 46*(12), 2455-2465.

Kellett, S., Simmonds-Buckley, M., Limon, E., Stride, C., Hughes, L., Hague, J. & Millings, A. (2020). Defining the assessment and treatment competencies to deliver low intensity cognitive behavioural therapy: A multi-centre validation study. *Behavior Therapy.*

Kroenke, K., Spitzer, R.L. & Williams, J.B. (2001). The PhQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16,* 606-613.

Kroenke, K., Spitzer, R. L., Williams, J. B., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine, 146*, 317-325.

Leighton, S.P., Upthegrove, R., Krishnadas, R., Benros, M.E., Broome, M.R., Gkoutos, G.V., . . . Mallikarjun, P.K. (2019). Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: A machine learning approach. *The Lancet Digital Health, 1(6),* 261-270.

Lincoln, T.M., Rief, W., Hahlweg, K., Frank, M., von Witzleben I., Schroeder, B. & Fiegenbaum, W. (2005). Who comes, who stays, who profits? Predicting refusal, dropout, success, and relapse in a short intervention for social phobia. *Psychotherapy Research, 15(3),* 210-225.

Lopez, V., Fernandez, A., Garcia, S., Palade, V. & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250,* 113-141.

McMillan, D., Gilbody, S., & Richards, D. A. (2010). Defining successful treatment outcome in depression using the PHQ-9: A comparison of methods. *Journal of Affective Disorders, 127,* 122-129.

Moons, K.G.M., Altman, D.G., Reitsma, J.B., Ioannidis, J.P.A., Macaskill, P., Steyerberg, E.W., . . . Collins, G.S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine, 162(2),* 1-73.

Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. M. (2002). The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *The British Journal of Psychiatry*, *180*(5), 461-464.

National IAPT Team. (2015). *National curriculum for the education of Psychological Wellbeing Practitioners, Third edition*. London: University College London.

National Institute for Health and Clinical Excellence. (2011). *Common mental health disorders: Identification and pathways to care*. National Collaborating Centre for Mental Health, London. Retrieved from https://www.nice.org.uk/guidance/cg123/resources/common-mental-health-problems-identification-and-pathways-to-care-pdf-35109448223173

Otto, M. W., Smits, J. A. and Reese, H. E. (2005). Combined psychotherapy and pharmacotherapy for mood and anxiety disorders in adults: review and analysis. *Clinical Psychology: Science and Practice, 12,* 72-86.

Papini, S., Pisner, D., Shumake, J., Powers, M.B., Beevers, C.G., Rainey, E.E. . . . Warren, A.M. (2018). Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization. *Journal of Anxiety Disorders, 60,* 35-42.

Paykel, E.S. (2008). Partial remission, residual symptoms, and relapse in depression. *Dialogues in Clinical Neuroscience, 10(4)*, 431-437.

Richards, D. A., & Borglin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: Two year prospective cohort evaluation. *Journal of Affective Disorders, 133,* 51–60.

Robinson, L., Delgadillo, J., Kellett, S. (2019). The dose-response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy Research.* doi: 10.1080/10503307.2019.1566676

Rodgers, M., Asaria, M., Walker, S., McMillan, D., Lucock, M., Harden, M., … Eastwood, A. (2012). The clinical effectiveness and cost-effectiveness of low-intensity psychological interventions for the secondary prevention of relapse after depression: a systematic review. *Health Technology Assessment, 16(28),* 1–130.

Scott, J., Palmer, S., Paykel, E., Teasdale, J., & Hayhurst, H. (2003). Use of cognitive therapy for relapse prevention in chronic depression: cost effectiveness study. *British Journal of Psychiatry, 182,* 221–227.

Spitzer, R.L., Kroenke, R., Williams, J.B. & Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal medicine, 166,* 1092-1097.

Steyerberg, E.W., Bleeker, S.E., Moll, H.A., Grobbee, D.E. & Moons, K.G.M. (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology, 56,* 441-447.

Tahir, M.A., Kittler, J., Mikolajczyk, K. & Zhou, Y. (2009). A multiple expert approach to the class imbalance problem using inverse random under sampling. In J.A. Benediktsson,

J. Kittler & F. Roli. (Eds.), *Multiple Classifier Systems: MCS 2009: Lecture Notes in Computer Science: Vol 5519* (pp. 82-91). Berlin, Heidelberg: Springer.

Vervliet, B., Craske, M.G. & Hermans, D. (2013). Fear extinction and relapse: State of the art. *Annual Review of Clinical Psychology, 9,* 215-248.

Vittengl, J. R., Clark, L. A., Dunn, T. W. and Jarrett, R. B. (2007). Reducing relapse and recurrence in unipolar depression: A comparative meta-analysis of cognitive behavioural therapy's effects. *Journal of Consulting and Clinical Psychology, 75,* 475-488.

Waddell, G., & Burton, A. K. (2006). *Is work good for your health and well-being?* London, UK: The Stationery Office.

Wojnarowski, C., Firth, N., Finegan, M. & Delgadillo, J. (2019). Predictors of depression relapse and recurrence after cognitive behavioural therapy: A systematic review and meta-analysis. *Behavioural and Cognitive Psychotherapy, 47*(5), 514-529.

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in Psychology: Lessons from Machine Learning. *Perspectives on Psychological Science, 12*(6), 1100-1122.