scientific reports

Check for updates

OPEN

Predicting anxiety treatment outcome in community mental health services using linked health administrative data

Kevin E. K. Chai[1]✉, Kyran Graham-Schmidt[2], Crystal M. Y. Lee[1], Daniel Rock[3,4,5], Mathew Coleman[6], Kim S. Betts[1], Suzanne Robinson[1,7] & Peter M. McEvoy[1,8]

Anxiety disorders is ranked as the most common class of mental illness disorders globally, affecting hundreds of millions of people and significantly impacting daily life. Developing reliable predictive models for anxiety treatment outcomes holds immense potential to help guide the development of personalised care, optimise resource allocation and improve patient outcomes. This research investigates whether community mental health treatment for anxiety disorder is associated with reliable changes in Kessler psychological distress scale (K10) scores and whether pre-treatment K10 scores and past health service interactions can accurately predict reliable change (improvement). The K10 assessment was administered to 46,938 public patients in a community setting within the Western Australia dataset in 2005–2022; of whom 3794 in 4067 episodes of care were reassessed at least twice for anxiety disorders, obsessive–compulsive disorder, or reaction to severe stress and adjustment disorders (ICD-10 codes F40–F43). Reliable change on the K10 was calculated and used with the post-treatment score as the outcome variables. Machine learning models were developed using features from a large health service administrative linked dataset that includes the pre-treatment K10 assessment as well as community mental health episodes of care, emergency department presentations, and inpatient admissions for prediction. The classification model achieved an area under the receiver operating characteristic curve of 0.76 as well as an F1 score, precision and recall of 0.69, and the regression model achieved an R^2 of 0.37 with mean absolute error of 5.58 on the test dataset. While the prediction models achieved moderate performance, they also underscore the necessity for regular patient monitoring and the collection of more clinically relevant and contextual patient data to further improve prediction of treatment outcomes.

Anxiety disorders are the most common class of mental illness in Australia, affecting 3.4 million adults aged 16 years and older or 17.2% of the population in 2020–2022[1]. Similarly in the United States, anxiety disorders are also the most common estimated to affect 30.6% of the population aged 18 years and older in 2020–2022[2]. These disorders are characterized by excessive worry, fear, and nervousness that can interfere with daily life. There are several different types of anxiety disorders, including generalized anxiety disorder, panic disorder, social anxiety disorder, and specific phobias. Historically, obsessive compulsive disorder and fear and stressor-related disorders (e.g., posttraumatic stress disorder) were considered anxiety disorders in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV, American Psychiatric Association, APA, 1994) although more recent nosologies consider them separate but related classes of disorders (DSM-5, APA, 2013). Within the International Classification of Diseases (ICD version 10, 2019; ICD version 11, 2023), these disorders are three categories within the mental, behavioural or neurodevelopmental disorders.

Primary care is the main source of treatment for anxiety disorders and, where required, providers more commonly refer patients to private specialist services than to public services[3]. Nonetheless, community mental

[1]School of Population Health, Curtin University, Perth, WA, Australia. [2]Department of Health, Perth, WA, Australia. [3]Western Australia Primary Health Alliance, Perth, WA, Australia. [4]Discipline of Psychiatry, Medical School, University of Western Australia, Perth, WA, Australia. [5]Faculty of Health, Health Research Institute, University of Canberra, Canberra, ACT, Australia. [6]Western Australia Country Health Service, Albany, WA, Australia. [7]Deakin Health Economics, Deakin University, Melbourne, VIC, Australia. [8]Centre for Clinical Interventions, North Metropolitan Health Service, Perth, WA, Australia. ✉email: k.chai@curtin.edu.au

health services remain important for patients who cannot afford or access private providers[4]. Public services refer to government funded and operated specialised mental health care provided by community and hospital based ambulatory care services, such as outpatient and day clinics[5] and offer a variety of ongoing treatment options including psychotherapy, medication, and support groups. A continuing challenge for clinicians and services in all settings is to predict how well an individual will respond to treatment. There are many factors that can influence outcomes, such as the severity of the disorder, the patient's readiness for change, the quality of the treatment they receive, and external factors that reflect the overall complexity of human lives (e.g., relationship breakdown, financial hardship, workplace redundancy, bereavement)[5–7].

Being able to accurately predict patient outcomes would be beneficial[7–10]. First, it would allow clinicians to tailor treatment plans to the individual needs of each patient, for example, by targeting known risk factors for disengagement or poor clinical outcomes. This could improve patient outcomes and reduce the need for patients to try multiple standardised treatments before finding one that works. Second, it would allow clinical planners in mental health services to allocate resources more effectively. For example, services could focus on providing more intensive treatment to patients who are at high risk of deterioration. Third, it could help identify patients who are unlikely to respond to treatment and may need additional support.

Promising methods for predicting patient outcomes for anxiety disorders and other mental illnesses include clinical prediction tools, patient-reported outcome measures, and machine learning[9–11]. These methods are commonly based on predictors such as patient demographics, clinical symptoms, treatment history, from different modes of data such as electronic health records, biometrics, and radiology and machine learning techniques such as logistic regression, random forests, support vector machines, gradient boosting and neural networks on datasets comprising of 4184 undergraduate students[9] and 1249 participants from a mental healthcare provider[11].

Research on the prediction of treatment outcomes in mental health show that it is difficult, either because treatment outcomes genuinely do not vary based on individual differences or due to a range of methodological limitations, such as investigations of variables based on convenience rather than strong theory; the lack of consideration of the complex interplay between relational and content components of psychotherapy; low statistical power due to studies being designed to evaluate main effects of treatments rather than moderators of symptom change; overly homogenous samples due to exclusion criteria in randomised trials; over-reliance on significance testing without due consideration to effect sizes; failure to probe interactions to understand patterns of effects; and neglecting non-linear relationships within the context of complex relationships for humans in the real world[8,12].

The alternative of relying on clinician intuition is also fraught. The biases clinicians bring to predicting psychotherapy outcomes have been long known[13–15]. Researchers have recently suggested that machine learning approaches that use large databases, theory-informed parameters and include complex relationships with multiple predictors of responder status, could address many of these issues[8,16,17]. Models that explain patterns in historical data and predict future outcomes, would hold promise for informing and improving the quality of care for people with anxiety disorders.

The aims of this study were to (a) investigate associations between demographic, treatment, and clinical variables and changes in psychological distress while patients were engaged with community mental health services and (b) develop machine learning models to predict reliable change in Kessler (K10) psychological distress scores using a patient's pre-treatment (K10) scores within a community mental health setting and their past health service interactions for anxiety disorders. No previous research has used a large sample of demographic, clinical, and treatment service data administratively collected within community mental health services over a 17-year period to predict changes in psychological distress using machine learning models.

Method

Study population

This study was approved by the Department of Health Western Australia Human Research Ethics Committee (approval number: RGS0000004782) and the Curtin University Human Research Ethics Committee (approval number: HRE2022-0001) with a waiver of informed consent obtained from the Department of Health Western Australia Human Research Ethics Committee. All methods in this study were performed in accordance with the relevant guidelines and regulations.

The study cohort was collated from a linked mental health dataset provided by the Department of Health Western Australia which is described elsewhere[18]. The linked dataset is comprised of records related to mental health assessments, community mental health service usage, emergency department presentations and inpatient admissions from 2005–2022.

For this study, we restricted the dataset to records from community mental health services where an anxiety disorder (International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification (ICD): F40–F43)[19] was recorded at any time in the episode of care and to episodes of care with at least two assessments (pre and post treatment ≥ 2 weeks and ≤ 4 months apart) for determining the outcome of the treatment. Based on community mental health dataset collection rules, assessments are not to be reported for brief community interventions (< 2 weeks) and that assessments should be completed at least every three months (we adjusted to 4 months to allow delays and scheduling issues). Data is included from eligible patient episodes of care, with the first pre/post assessment used for each individual episode. Allowing multiple care episodes per patient better represents real-world conditions, providing a more accurate evaluation of the predictive model's performance on each patient encounter. We conducted a sensitivity analysis comparing the use of single and multiple episodes of care in Supplementary Discussion 1. ICD-10 was used as 99% of records in the community mental health data collection period within the study population used this classification.

The dataset preparation steps for defining the study population (Table 1) and the number of records from each anxiety disorder ICD-10 code (Table 2) are presented below.

Primary outcome measure

The K10 assessment is a self-reported measure of anxiety and depression symptoms characteristic of the broad construct of psychological distress[20]. It comprises of 10 questions about emotional states assessed on a five-level response scale (1 = none of the time, 2 = a little of the time, 3 = some of the time, 4 = most of the time, 5 = all of the time). The responses from the 10 questions can be summed to a total ranging from 10 to 50, where lower scores represent lower levels of distress. The K10 has high internal reliability (Cronbach's alpha = 0.93)[21], distinguishes people with and without anxiety disorders[22], and has been shown to be highly sensitive to change during psychotherapy[23]. We calculated Cronbach's alpha for each ICD-10 code in our dataset using the Pingouin Python statistical package[24].

Data analysis plan

Treatment outcome

The treatment outcome and its effectiveness were determined by subtracting the post-treatment score from the pre-treatment score. Given that changes in scores reflect true change plus measurement error, Jacobson and Traux proposed the Reliable Change Index (RCI) to evaluate the effectiveness of therapies and interventions based on pre/post treatment scores[25]. The RCI estimates the magnitude of change in a measure's observed score required before assuming that true change has occurred (i.e., not attributable to measurement error). The RCI is calculated by dividing the difference between the two scores by the standard error of the difference. RCI values ≥ 1.96 represent reliable improvement, RCI values ≤ 1.96 represent reliable deterioration and RCI values between − 1.96 and 1.96 represent no reliable change. The K10 was used as both a continuous outcome variable (post-treatment score) and to classify individuals with respect to whether they reliably improved, deteriorated, or remained unchanged between pre-treatment and post-treatment. The calculation of the RCI and subsequent analysis were conducted using Python 3.9.

Dataset

The dataset of the study population was prepared with the prediction model features restricted to data from the K10 pre-assessment and previous community mental health episodes of care, in addition to emergency department and inpatient mental health service events (Fig. 1).

The features extracted and created from these data sources are presented in Table 3 with definitions provided in Supplementary Table 1. The dataset is split into a 70%/30% training and test set using fivefold random subsampling stratified cross validation in machine learning experiments.

Machine learning

Classification and regression models are used to predict the reliable change category (deterioration/no reliable change vs. reliable improvement) and post treatment score as a continuous variable, respectively. Models were trained using the Python scikit-learn library[26]. Training (70%) and testing (30%) datasets were created using a stratified fivefold repeated random sub-sampling cross-validation method.

Model selection

PyCaret[27], an automated machine learning (AutoML) software library, was used to initially experiment with several machine learning algorithms by splitting only the training dataset into 70/30% using fivefold random

	Records	Episodes	Patients
Mental health assessments	1,171,287	299,055	103,691
K10 assessments	225,088	123,647	61,766
Community setting	147,928	70,337	46,938
F40–F43 diagnosis	4067	4067	3794

Table 1. Study population. *Episodes* episodes of care, *K10* Kessler psychological distress scale, *F40-F43* anxiety disorder ICD-10 codes for primary and additional diagnoses.

ICD-10 code	Records
F40: Phobic anxiety disorders	702 (17%)
F41: Other anxiety disorders	1937 (48%)
F42: Obsessive–compulsive disorder	129 (3%)
F43: Reaction to severe stress, and adjustment disorders	1299 (32%)

Table 2. ICD-10 codes.

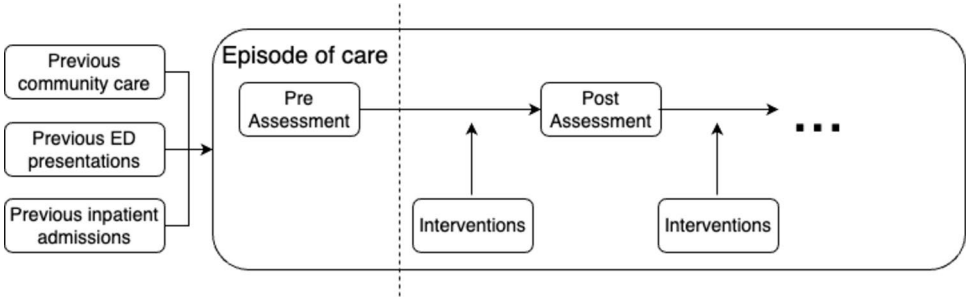

Fig. 1. The data sources and features that are available for the prediction model at pre-assessment are depicted to the left of the dashed line. The first pre/post assessment is used for each episode of care and patients may have multiple eligible episodes of care in the dataset. *ED* emergency department.

Pre-assessment	Previous community mental health contacts	Previous emergency department presentations	Previous inpatient admissions
Age	Number of episodes	Number of presentations	Number of admissions
Sex	Years since	Years since	Years since
Focus of care	Client sessions		Had psychiatric care
Phase of care	Client duration mean		
Collection stage	Associate sessions		
Stream type	Associate duration mean		
Legal status	Deactivation outcome		
Diagnosis count	Presenting complaint		
Remoteness area	Triage outcome		
Relative disadvantage index	Legal status		
Score			

Table 3. Prediction model features.

sampling cross validation. These initial results will be used to select the most suitable classification and regression methods for subsequent experiments.

Model evaluation

The classification models are evaluated using the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), precision, recall, F1 score (harmonic mean of precision and recall) and a confusion matrix to identify how often a model gets predictions right (true positives/negatives) and wrong (false positives/negatives) for each reliable change category. An AUC of 1 is considered to have perfect predictive power while an AUC 0.5 suggests no predictive power beyond random chance[28]. The regression models are evaluated using predicted R squared (R^2) and the mean absolute error[29]. The predicted post-treatment scores from the regression model were also used to classify episodes of care into the reliable change categories for evaluation.

Feature importance and selection

Shapley Additive Explanations (SHAP) is a game theory inspired technique commonly used to explain the importance and contribution of features in prediction modelling[30,31]. It is a model agnostic approach applied to both classification and regression models in our experiments using the SHAP Python library[31]. Furthermore, a greedy forward feature selection method[32] was applied, which involved sequentially adding the feature that provides the largest contribution to the model until a pre-defined stopping criterion was met. The stopping criteria used in experiments for classification were F1 improvement > 0.01 and mean absolute error (MAE) improvement < 0.001 for regression.

Results

Treatment outcome

The distribution of score changes between pre/post-treatment is shown in Fig. 2. 2882 (71%) episodes of care showed a reduction in K10 score after treatment, 872 (21%) exhibited an increase in K10 after treatment and 313 (8%) remained unchanged.

The RCI method was applied on the dataset, where K10 reliability coefficients (Cronbach's alphas) of 0.92–0.94 were calculated for each of the ICD codes. The pattern of reliable change for F43 (Reaction to severe stress, and adjustment disorders) is illustrated in Fig. 3. These boundaries vary for other ICD codes (F40, F41, F42) as the reliable change index was calculated and applied separately for each diagnosis (Supplementary Fig. 1).

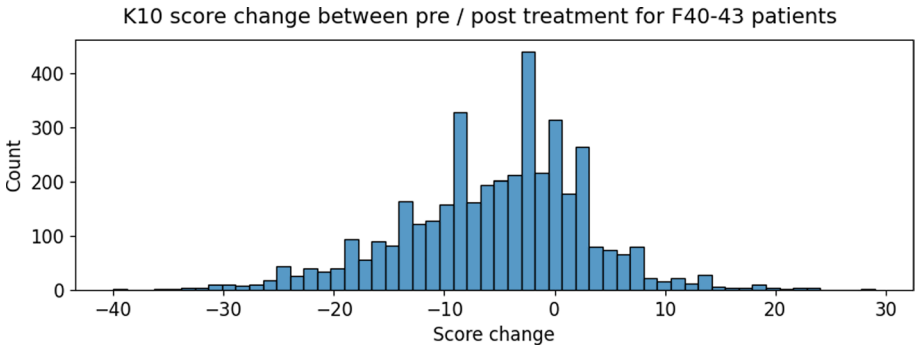

Fig. 2. The difference (score change) between pre/post treatment Kessler psychological distress scale (K10) total scores.

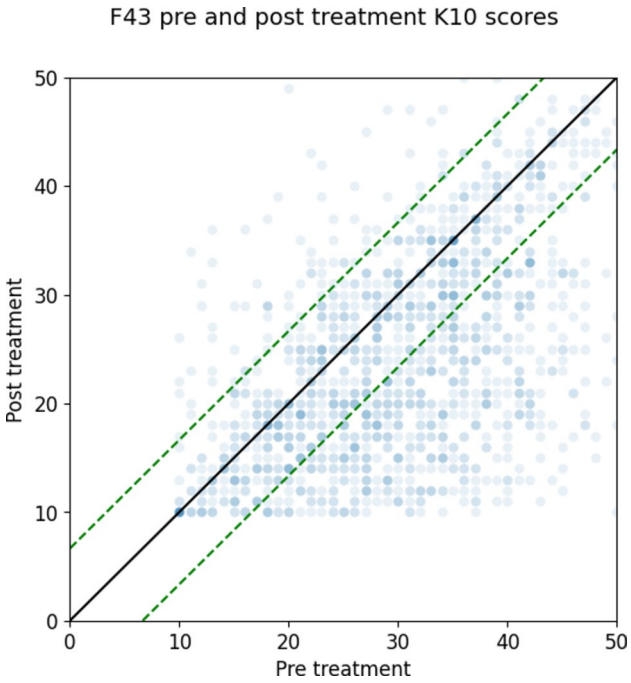

Fig. 3. Pre/post treatment scores for F43: Reaction to severe stress, and adjustment disorders. The dashed green lines represent the boundaries of the reliable change index, with the area to the left representing reliable deterioration and the area to the right representing reliable improvement. The area between the green lines represents no reliable change.

Dataset

Descriptive statistics for the dataset are reported in Table 4. Altogether, 4067 episodes of care were available for analysis that comprised predominately of females (67%) and a mean (SD) age of 40.2 (17.9) years. The deteriorated reliable change category had low representation (212 records or 5%) and was merged with the no reliable change category (total of 2446 records or 60%) for machine learning experiments.

Machine learning

The machine learning results are presented in two sections (a) classification for predicting the reliable change category and (b) regression for predicting post-assessment scores.

Classification

Model selection

PyCaret (AutoML) was used to initially experiment with several classification models on the training dataset using cross-validation as presented in Table 5. Gradient boosting achieved the highest AUC (0.72) and F1 score (0.57). All the models outperform the baseline classifier (AUC = 0.5) that predicts all records as the majority class (deteriorated/no reliable change). Based on these results, gradient boosting was selected for subsequent experiments.

	N (%)
n (episodes of care)	4067
Sex	
Male	1335 (33%)
Female	2729 (67%)
Other	≤ 5 (0%)
Age	
Mean (SD)	40.2 (17.9)
Adult stream	3601 (88%)
Elderly stream	466 (12%)
Assessments	
Pre score mean (SD)	29.1 (8.8)
Post score mean (SD)	23.8 (8.9)
Reliable change	
Improved	1621 (40%)
Deteriorated	212 (5%)
No change	2234 (55%)
Focus of care	
Acute	313 (8%)
Functional gain	1845 (45%)
Intensive extended	171 (4%)
Maintenance	801 (20%)
Not specified	937 (23%)
Remoteness	
Metro	2423 (59%)
Rural	633 (16%)
Remote	374 (9%)
Not specified	637 (16%)
IRSD	
Quintile 1	90 (2%)
Quintile 2	1703 (42%)
Quintile 3	125 (3%)
Quintile 4	1069 (26%)
Quintile 5	861 (21%)
Not specified	219 (6%)

Table 4. Dataset characteristics. *SD* standard deviation, *IRSD* Index of Relative Socio-economic Disadvantage.

Model	AUC	Precision	Recall	F1
Gradient Boosting	0.72	0.57	0.57	0.57
Logistic Regression	0.72	0.58	0.49	0.53
Linear Discriminant Analysis	0.71	0.58	0.51	0.54
Random Forest	0.71	0.59	0.45	0.51
Naive Bayes	0.62	0.40	0.92	0.56
Decision Tree	0.58	0.49	0.51	0.50
Baseline	0.50	0.00	0.00	0.00

Table 5. AutoML classification results.

Model evaluation

The gradient boosting model was run on both the train and test datasets achieving an average F1 score of 0.66 (0.66–0.69) over fivefold cross validation, with the best model achieving an AUC of 0.77 and F1 of 0.69 (Table 6).

The confusion matrix and ROC of the best model is presented in Fig. 4. The confusion matrix highlighted that the model performed better in classifying episodes of care with deterioration/no reliable change (551 out of 734 (75%) correctly classified) than those that demonstrated reliable improvement (306 out of 487 (63%) correct).

Dataset	AUC	Precision	Recall	F1	Records
Train	0.76	0.71	0.71	0.71	2846
Test	0.76	0.69	0.69	0.69	1221

Table 6. Classification best model results. *AUC* area under the receiver operating characteristic curve.

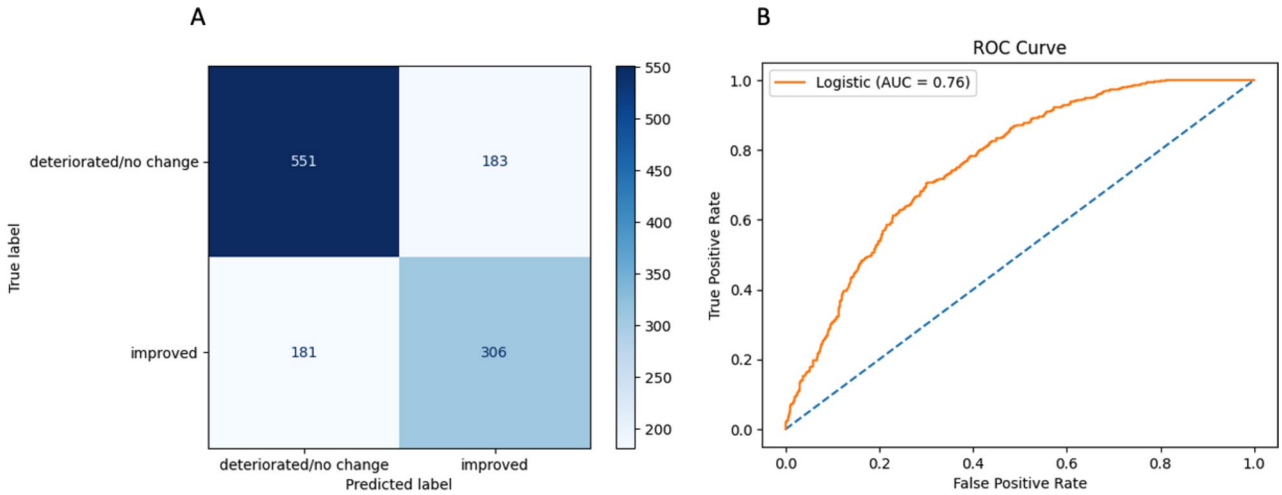

Fig. 4. (**A**) Classification confusion matrix shows how often the model correctly predicted each class (true positives/negatives) and how often it made mistakes (false positives/negatives). (**B**) The receiver operating characteristic curve on the test dataset shows the sensitivity and specificity at different thresholds for prediction.

Feature importance and selection

The top 20 features based on the SHAP values and feature selection results are shown in Supplementary Table 2 and Supplementary Fig. 2. The top 2 features from both methods were the pre-assessment score and the collection stage (review). Only using the pre-assessment score achieved a 0.62 F1 score with the admission collection stage increasing the prediction performance to 0.66 and years since the previous emergency contact to 0.69. The additional 4 selected features only improve the model performance to 0.70 (+ 0.1 F1 score).

Regression

Model selection

AutoML was applied to experiment with several regression models on the training dataset using cross-validation as presented in Table 7. Gradient boosting achieved the top performance with a 0.33 R^2 and 5.82 MAE. All models, except for decision tree, outperformed the baseline regressor that predicts the mean post-treatment score for all records. The gradient boosting model was selected for subsequent experiments.

Model evaluation

The gradient boosting model achieved an average MAE of 5.73 (5.58–5.83) over fivefold cross validation with the best model achieving an R^2 of 0.39, 0.37 and MAE values of 5.65, 5.58 on the train and test dataset, respectively (Table 8).

Feature importance and selection

The top 20 features based on the absolute SHAP values and feature selection results are shown in Supplementary Table 3 and Supplementary Fig. 3. Feature selection identified the pre-assessment score and the collection stage

Model	R^2	MAE
Gradient Boosting	0.33	5.82
Linear Regression	0.32	5.91
Random Forest	0.29	5.98
K Neighbours	0.13	6.67
Baseline	−0.01	7.48
Decision Tree	−0.35	8.04

Table 7. AutoML regression results.

Dataset	R^2	MAE	Records
Train	0.39	5.65	2846
Test	0.37	5.58	1221

Table 8. Regression best model results.

(admission) as the top features achieving a 5.75 and 5.59 MAE. The other 5 selected features only reduced the MAE to 5.52 (− 0.07).

Regression applied classification
The regression model predicted the post-assessment score and was used to classify episodes of care into reliable change. The regression applied classification results (Table 9) showed a decline when compared to the classification model with an F1 score of 0.69 vs. 0.67 on the test set. The AUC cannot be computed for comparison as the regression model does not generate classification probabilities.

The confusion matrix of the regression applied classification is shown in Fig. 5. These results when compared to the classification model showed that the regression model performed poorer in predicting improved reliable change (306 vs. 304), and deterioration/no reliable change (551 vs. 533).

Discussion

This study aimed to investigate whether community mental health treatment is related to improvements in psychological distress and develop machine learning models for predicting reliable change and post-treatment scores in anxiety disorder treatments. The discussion will now assess whether the results and findings adequately achieved these aims.

Prediction performance

The classification model achieved an AUC of 0.76 on the test dataset of 1193 patients and an AUC between 0.75 and 0.90 indicates a moderate score in psychology and human behavioural research[33,34]. Our results are similar to a study that achieved an AUC of 0.73 on a test dataset of 1255 undergraduate students[9] and outperformed another study that achieved an AUC of 0.60 on 279 patients in their test dataset[11]. The regression model

Dataset	Precision	Recall	F1	Records
Train	0.67	0.67	0.67	2846
Test	0.67	0.67	0.67	1221

Table 9. Regression applied classification results.

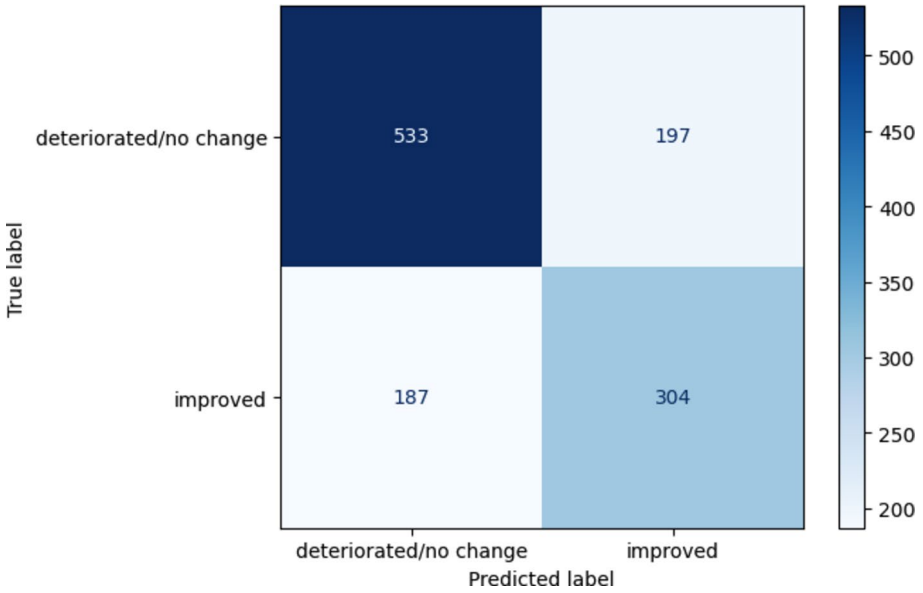

Fig. 5. Regression applied classification confusion matrix shows how often the model correctly predicted each class (true positives/negatives) and how often it made mistakes (false positives/negatives).

achieved a R^2 on the test dataset and a R^2 between 0.3 and 0.5 is generally considered a weak effect[35] but can be considered as moderate in the context of human behavioural and psychology research[36]. Furthermore, A MAE of 5.58 for the regression model could be interpreted as a relatively large error for downstream tasks such as using the predicted post-treatment scores to classify reliable change. The classification and regression applied classification model achieved similar performance and both outperformed the baseline models. The moderate performance indicates that the models could be further improved with more data and/or better discriminating features. However, there is likely to be an upper limit on prediction performance given the inherent complexity of human lives in predicting the outcome of patient treatments (i.e. Bayes error)[29].

Classification and regression

The classification model generated probabilities for each class, which helped identify appropriate classification thresholds using the ROC and AUC evaluation metrics. However, a strength of the regression model is that it predicted the post-treatment score, which allowed for the use of classification systems such as reliable change and could potentially be used for other metrics of recovery. Furthermore, the SHAP values of the regression model were easier to interpret as a higher SHAP value indicated a higher predicted post-assessment score (poorer outcomes) compared to classification where a higher SHAP value represents as a lower post-assessment score (improved reliable change). For example, a high pre-assessment score (poor outcome) for classification resulted in the model predicting towards reliable improvement, possibly due to higher pre-assessment scores having more potential to change by post-assessment (i.e. lower scores experiencing a floor effect). However, for regression, a high pre-assessment score (poor outcome) would predict towards high post-assessment scores (poor outcomes).

Model features

The SHAP analysis and feature selection experiments showed that the pre-assessment score was the most important feature, with the assessment collection stage (admission, review) improving prediction with the remaining features providing only a minor contribution to the overall performance. However, a strength of having fewer contributing features is that the model is simpler to implement and translate into clinical software. These top features were, however, not particularly helpful for future treatment-matching, although the challenge of discovering robust predictors of mental health treatment outcomes is well known[8,12]. A shift from capturing predominantly health service activity data to capturing more clinically relevant data (e.g., therapeutic process, treatments delivered) along with contextual factors (i.e., non-therapy factors such as life stressors), and implementing more regular patient outcome monitoring[37] to more readily identify when a clinical intervention is not working and could be adapted or stopped, may be required to improve prediction. A cardiologist would not contemplate diagnosing and evaluating interventions for heart disease from single datapoints three months apart, and yet mental health services are expected to do so.

Clinically relevant data

While the study dataset can be seen as a strength (i.e. linked population dataset collected over a 17-year period for training and evaluating prediction models) it is still limited and can be further enhanced. The collection of administrative patient data is often driven by compliance and reporting requirements rather than a clear understanding of its clinical utility. This can lead to the accumulation of vast amounts of data that are difficult to analyse and interpret, providing limited insights into patient care and outcomes. Moreover, the focus on compliance can divert resources away from efforts to collect and curate data that is directly relevant to clinical decision-making while burdening clinicians with onerous data entry administrative tasks. For instance, measures of key individual differences theorised to play a critical role in the aetiology and maintenance of anxiety disorders, such as anxiety sensitivity[38], intolerance of uncertainty[39], and experiential avoidance[40], may help with case formulation, treatment planning, and outcome monitoring. The degree to which interventions successfully modify these factors would be expected to determine downstream impacts on symptom change across the anxiety disorders. Patients' satisfaction and engagement with the service (e.g., attendance frequency and duration), relational factors between the clinician and patient (e.g., working alliance[41]), and social determinants (e.g., interpersonal supports and stressors, financial stressors, adverse childhood experiences[42,43]) may also help focus clinicians' and consumers' attention on factors likely to have the largest impact on mental health and wellbeing and thereby improve outcomes and their prediction. Outcomes beyond symptom change that capture broader intervention impacts (e.g., quality of life), or monitoring progress on idiographic presenting problems (those specific and of highest priority to the individual), may be particularly valued by consumers[44], although there is evidence that improvements in quality of life are largely mediated by symptom change[45]. Routine monitoring of known predictors of mental health and wellbeing would facilitate outcome evaluation and benchmarking, whereby novel interventions and service models can be compared over time to previous benchmarks. Without these data, services have no way of knowing if outcomes are worsening, maintaining, or improving over time, which would help with treatment planning. There is evidence that regular and routine outcome monitoring (e.g., session-by-session) that is used collaboratively by consumers and clinicians can improve outcomes, decrease negative outcomes for consumers at risk of not benefiting from treatment, and increase cost-effectiveness of interventions[46]. Future research incorporating and documenting these measures and processes would likely produce more robust and informative predictive models.

The inability of the prediction models to produce higher or more robust performance might suggest that the health administrative data being collected and made available for research lacks clinical relevance, which makes its collection and use difficult to justify. The resources invested in collecting and storing this data could be better utilised towards initiatives that directly improve patient care. Moreover, relying on data that fails to provide meaningful insights could lead to misguided policy decisions and interventions that may not produce the desired

outcomes. Administrative data collected solely for service utilisation and planning metrics are insufficient for evaluating quality of care, identifying impacts of service innovations, and ensuring consumer outcomes improve over time. If the priority is maximising patient recovery, then infrastructure (e.g., digital platforms) and measures that routinely, regularly and effectively capture consumer-driven priorities are required to ensure interventions are on track for positive outcomes, or, if not, can be, collaboratively and rapidly responded to by the consumer and healthcare worker to process back on track.

Clinical assessment

A limitation of the model and experiments are features provided by clinicians in their assessments of the patients such as unstructured clinical notes. While these features could aid in prediction, it is noteworthy to highlight that it is also difficult for clinicians to predict, based only from the initial pre-assessment, whether a patient will drop out, be treatment resistant or improve. If this cannot be predicted accurately and reliably by clinical experts[13–15], then it may be no different when developing and using predictive models. Future research including a combination of clinician, consumer, and administrative data may improve predictive models.

Conclusion

Predicting patient outcomes in mental health is a complex and difficult task but is essential for improving the quality of care for people with anxiety disorders. Research on the prediction of patient outcomes is ongoing and the preliminary findings to date are promising. This study developed classification and regression models that showed moderate prediction performance with features that would be relatively easy to collect and implement in health services organisations and clinics on a linked health administrative dataset collected over a 17-year period. Future research using regular patient outcome monitoring, clinical assessment, consumer and administrative data, may yield more accurate and reliable models for predicting patient outcomes. This will have a significant impact on the lives of people with anxiety disorders and will inform healthcare policy planning.

Data availability

The data that support the findings of this study are available from Government of Western Australia Department of Health (https://www.datalinkage-wa.org.au/) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The corresponding author can provide clarification of the dataset used for the study but for access to the data, contact the Western Australia Department of Health at DataServ@health.wa.gov.au.

Received: 17 May 2024; Accepted: 29 August 2024
Published online: 04 September 2024

References

1. National Study of Mental Health and Wellbeing. https://www.abs.gov.au/statistics/health/mental-health/national-study-mental-health-and-wellbeing/latest-release (2023).
2. Villaume, S. C., Chen, S. & Adam, E. K. Age disparities in prevalence of anxiety and depression among US adults during the COVID-19 pandemic. *JAMA Netw. Open* **6**(11), e2345073 (2023).
3. Australian Institute of Health and Welfare. *Medicare-subsidised mental health-specific services.* https://www.aihw.gov.au/mental-health/topic-areas/medicare-subsidised-services (2023).
4. Castillo, E. G. *et al.* Community interventions to promote mental health and social equity. *Curr. Psychiatry Rep.* **21**, 1–14. https://doi.org/10.1007/11920-019-1017-0 (2019).
5. Australian institute of Health and Welfare. *Community Services—Mental health AIHW*. https://www.aihw.gov.au/mental-health/topic-areas/community-services (2023).
6. McMahon, F. J. Prediction of treatment outcomes in psychiatry—Where do we stand?. *Dialogues Clin. Neurosci.* **16**(4), 455–464 (2014).
7. Chekroud, A. M. *et al.* The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry* **20**(2), 154–170 (2021).
8. Eilertsen, S. E. H. & Eilertsen, T. H. Why is it so hard to identify (consistent) predictors of treatment outcome in psychotherapy? Clinical and research perspectives. *BMC Psychol.* **11**(1), 198 (2023).
9. Nemesure, M. D., Heinz, M. V., Huang, R. & Jacobson, N. C. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci. Rep.* **11**(1), 1980 (2021).
10. Stanojevic, M., Norris, L. A., Kendall, P. C. & Obradovic, Z. Predicting anxiety treatment outcomes with machine learning. In *2022 21st IEEE International Conference on Machine Learning and Applications* (*ICMLA*) 957–962 (IEEE, 2022).
11. Hornstein, S., Forman-Hoffman, V., Nazander, A., Ranta, K. & Hilbert, K. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach. *Digit. Health* **7**, 20552076211060659 (2021).
12. Erceg-Hurn, D. M., Campbell, B. N. & McEvoy, P. M. What explains the failure to identify replicable moderators of symptom change in social anxiety disorder?. *J. Anxiety Disord.* **94**, 102676 (2023).
13. Meehl, P. E. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence (1954).
14. Dawes, R. M., Faust, D. & Meehl, P. E. Clinical versus actuarial judgment. *Science* **243**(4899), 1668–1674 (1989).
15. Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L. & Latzman, R. D. Why ineffective psychotherapies appear to work: A taxonomy of causes of spurious therapeutic effectiveness. *Perspect. Psychol. Sci.* **9**(4), 355–387 (2014).
16. Mululo, S. C. C., Menezes, G. B. D., Vigne, P. & Fontenelle, L. F. A review on predictors of treatment outcome in social anxiety disorder. *Braz. J. Psychiatry* **34**, 92–100 (2012).
17. Ang, Y. S. & Pizzagalli, D. A. Predictors of treatment outcome in adolescent depression. *Curr. Treat. Options Psychiatry* **8**, 18–28 (2021).
18. Lee, C. M. Y. *et al.* Patterns of mental service utilisation: A population-based linkage of over 17 years of health administrative records. *Community Ment. Health J.* https://doi.org/10.1007/s10597-024-01300-8 (2024).
19. National Centre for Classification in Health. *The International Statistical Classification of Diseases and Related Health Problems, Australian Modification (ICD-10-AM)* 10th edn. (Independent Hospital Pricing Authority, 2017).
20. Kessler, R. C. M. D. & Mroczek, D. *An Update of the Development of Mental Health Screening Scales for the US National Health Interview Study* (University of Michigan, Survey Research Center of the Institute for Social Research, 1992).

21. Kessler, R. C. *et al.* Screening for serious mental illness in the general population. *Arch. Gen. Psychiatry* **60**(2), 184–189 (2003).
22. Andrews, G. & Slade, T. Interpreting scores on the Kessler Psychological Distress Scale (K10). *Aust. N. Z. J. Public Health* **25**, 494–497 (2001).
23. McEvoy, P. M. *et al.* Group metacognitive therapy for repetitive negative thinking in primary and non-primary generalized anxiety disorder: An effectiveness trial. *J. Affect. Disord.* **175**, 124–132 (2015).
24. Vallat, R. Pingouin: Statistics in Python. *J. Open Source Softw.* **3**(31), 1026. https://doi.org/10.21105/joss.01026 (2018).
25. Jacobson, N. S. & Truax, P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *J. Consult. Clin. Psychol.* **59**, 12–19 (1992).
26. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
27. Ali, M. PyCaret: An open source, low-code machine learning library in Python. https://www.pycaret.org (2020).
28. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **5**(9), 1315–1316 (2010).
29. Bishop, C. M. & Nasrabadi, N. M. *Pattern Recognition and Machine Learning* Vol. 4, 738 (Springer, 2006).
30. Strumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst. J.* https://doi.org/10.1007/s10115-013-0679-x (2014).
31. Lundberg, S. M., & Lee, S. I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (*NeurIPS*) (2017).
32. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**(2), 1157–1182 (2003).
33. Swets, J. A. Measuring the accuracy of diagnostic systems. *Science* **240**(4857), 1285–1293 (1988).
34. Streiner, D. L. & Cairney, J. What's under the ROC? An introduction to receiver operating characteristics curves. *Can. J. Psychiatry* **52**(2), 121–128 (2007).
35. Moore, D. S., Notz, W. & Fligner, M. A. *The Basic Practice of Statistics* (W.H. Freeman and Company, 2013).
36. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn. (Lawrence Erlbaum Associates, 1988).
37. Lambert, M. J. & Harmon, K. L. The merits of implementing routine outcome monitoring in clinical practice. *Clin. Psychol. Sci. Pract.* **25**(4), e12268 (2018).
38. Naragon-Gainey, K. Meta-analysis of the relations of anxiety sensitivity to the depressive and anxiety disorders. *Psychol. Bull.* **136**(1), 128 (2010).
39. McEvoy, P. M., Hyett, M. P., Shihata, S., Price, J. E. & Strachan, L. The impact of methodological and measurement factors on transdiagnostic associations with intolerance of uncertainty: A meta-analysis. *Clin. Psychol. Rev.* **73**, 101778 (2019).
40. Akbari, M., Seydavi, M., Hosseini, Z. S., Krafft, J. & Levin, M. E. Experiential avoidance in depression, anxiety, obsessive-compulsive related, and posttraumatic stress disorders: A comprehensive systematic review and meta-analysis. *J. Context. Behav. Sci.* **24**, 65–78 (2022).
41. Vaz, A. M., Ferreira, L. I., Gelso, C. & Janeiro, L. The sister concepts of working alliance and real relationship: A meta-analysis. *Counsel. Psychol. Q.* **37**(2), 247–268 (2024).
42. de Graaf, R., ten Have, M., Tuithof, M. & van Dorsselaer, S. First-incidence of DSM-IV mood, anxiety and substance use disorders and its determinants: Results from the Netherlands Mental Health Survey and Incidence Study-2. *J. Affect. Disord.* **149**(1–3), 100–107 (2013).
43. Sharma, S., Powers, A., Bradley, B. & Ressler, K. J. Gene × environment determinants of stress-and anxiety-related disorders. *Annu. Rev. Psychol.* **67**(1), 239–261 (2016).
44. Cuijpers, P. Targets and outcomes of psychotherapies for mental disorders: An overview. *World Psychiatry* **18**(3), 276–285 (2019).
45. Lundqvist, L. O. *et al.* Influence of mental health service provision on the perceived quality of life among psychiatric outpatients: Associations and mediating factors. *Front. Psychiatry* **14**, 1282466 (2024).
46. McAleavey, A. A., de Jong, K., Nissen-Lie, H. A., Boswell, J. F., Moltu, C. & Lutz, W. (2024). Routine outcome monitoring and clinical feedback in psychotherapy: Recent advances and future directions. *Administration and Policy in Mental Health and Mental Health Services Research*, 1–15.

Acknowledgements

This work was supported by the Digital Health Cooperative Research Centre (DHCRC) [DHCRC-0076]. DHCRC is funded under the Australian Commonwealth's Cooperative Research Centres (CRC) Program. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation for the manuscript. The authors wish to thank Justin Manuel from Western Australia Country Health Service for his ongoing contribution to the overall project and to the staff from the Department of Health WA's Data Linkage Services and the Hospital Morbidity Data Collection, Emergency Department Data Collection, and Mental Health Data Collection.

Author contributions

K.E.K.C analysed the data, conducted the experiments and drafted the manuscript. K.E.K.C, K.G.S C.M.Y.L, P.M.M., D.R, M.C conceived the design and P.M.M, K.G.S, D.R, M.C provided clinical advice for the project. K.S.B, P.M, D.R, S.R secured funding for the project. All authors contributed to the critical revision of the manuscript and approved the final version of the article to be published.

Competing interests

The authors declare no competing interests.

Ethics approval

This study was approved by the Department of Health Western Australia Human Research Ethics Committee (approval number: RGS0000004782) and the Curtin University Human Research Ethics Committee (approval number: HRE2022-0001) with a waiver of informed consent obtained from the Department of Health Western Australia Human Research Ethics Committee. All methods in this study were performed in accordance with the relevant guidelines and regulations.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-71557-2.

Correspondence and requests for materials should be addressed to K.E.K.C.