Improving the Quality of Clinical Assessments in Talking Therapies Through Machine Learning

Limbic
Research team
London, UK
research@limbic.ai

July 1, 2024

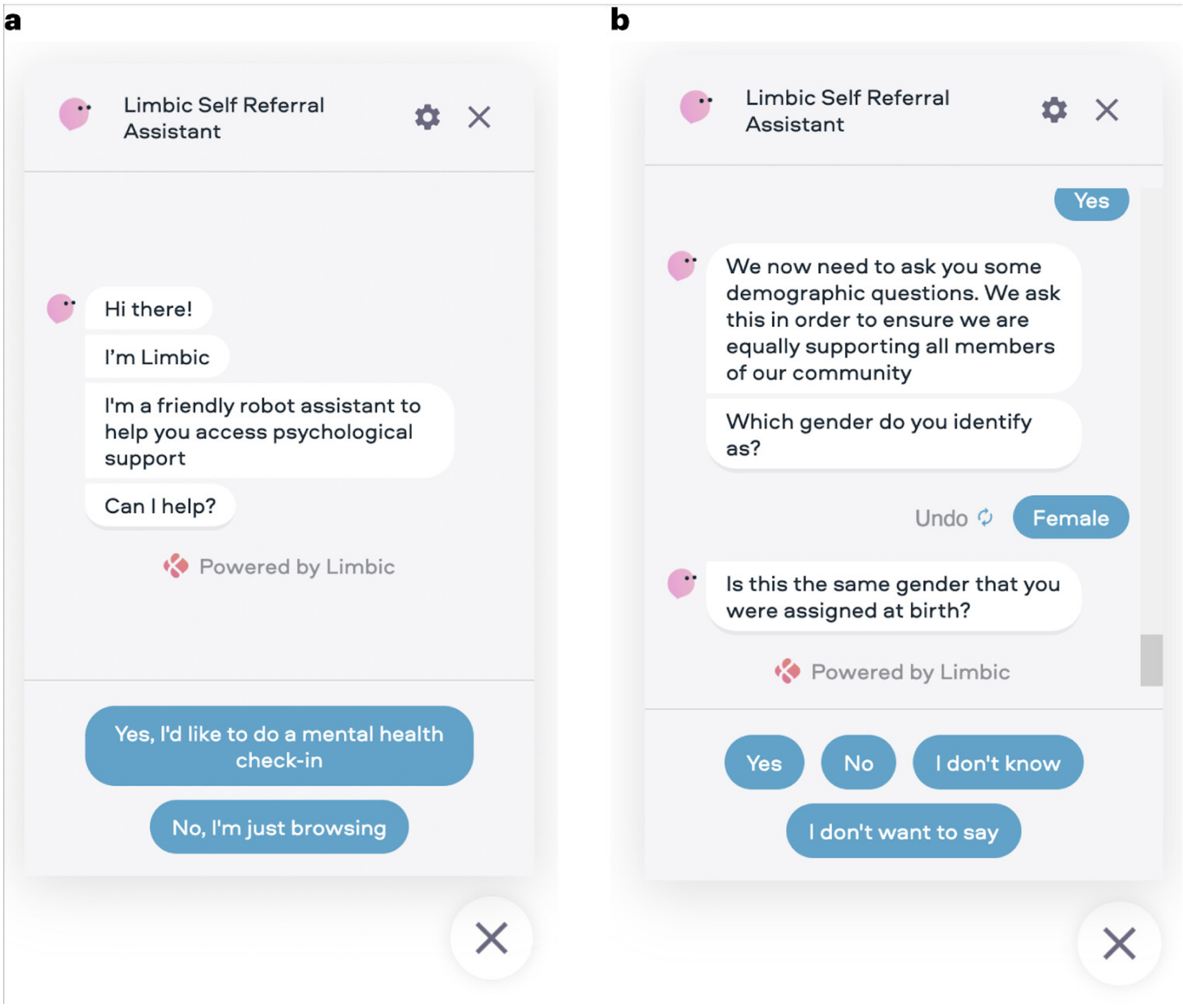
Summary

We developed a Clinical AI Assessment Assistant called Limbic Access, equipped with a clinical brain that uses specialised clinical machine learning models to inform clinical assessments and improve diagnostic quality. Limbic Access uses a clinically validated machine learning (ML) model to identify a ranked consideration set of presenting problems ordered by likelihood, from which additional anxiety disorder specific outcome measures (ADSMs) are adaptively administered. Additional clinical logic is applied to re-rank the consideration set into primary and secondary presenting problems and are used by clinicians to inform clinical assessment. In 3 different studies, with a total of over 20,000 patients, we show that our ML model identifies the appropriate ADSMs with an accuracy of over 93% across the 8 most common diagnostic categories, matching the performance of trained clinicians. Further, we show that screening for presenting problems by Limbic Access' is substantially more accurate than the usage of standardized screening questions from the Talking Therapy manual, which represents the gold-standard in the absence of ML. By accurately identifying presenting problems and adaptively administering personalised outcome measures, Limbic Access has the potential to save clinicians time and support the quality of assessments, as well as offer patients an enhanced referral experience and better outcomes - highlighting the important role that technological solutions can play in improving clinical efficiencies.

1 Method

1.1 Clinical AI Assessment Assistant

We developed Limbic Access, a web-hosted chatbot designed to facilitate self-referral into a given mental health service, support data collection for service provider staff, and provide insights to clinicians ahead of clinical assessments. The Limbic Access chatbot sits on the relevant service provider's website and "pops up" on the desired web-page, providing an interactive and engaging medium to support the user making a referral to the given service. (Figure 1)



Hi there!

I'm Limbic

I'm a friendly robot assistant to help you access psychological support

Can I help?

We now need to ask you some demographic questions. We ask this in order to ensure we are equally supporting all members of our community

Which gender do you identify as?

Is this the same gender that you were assigned at birth?

Figure 1: **Illustrations of the personalized self-referral chatbot on a NHS Talking Therapies service provider website** a, Illustration of the AI-enabled self-referral chatbot's initial message upon accessing the service's website. The message is customized to the specific service. b, Illustration of the chatbot collecting demographic information.

As part of the chatbot's configurable conversational flow, Limbic Access asks a number of questions and collects patient inputs in response. These include:

- **Free text**: A description of the patient's main problems in their own words and the goal they wish to achieve in seeking care.
- **Standardised questionnaire responses**: These are multiple choice answers corresponding to the 'minimum dataset' collected across all NHS Talking Therapies services, which includes the Patient Health Questionnaire (PHQ9), the Generalized Anxiety Disorder Questionnaire (GAD7), the Work and Social Adjustment Scale (WSAS) and the IAPT Phobia Scales.

- **Behavioral indicators**: This includes passive data recorded from users interacting with the chatbot, including typing speed, response times, number of deleted characters in free-text answers and number of changes for multiple choice answers.
- **Demographic variables**: These are "Yes/No" or multiple choice responses to demographic questions such as age, gender and so on. We use these primarily to assess algorithmic bias in our machine-learning model.

1.2 Machine Learning Model

Here we describe the purpose, architecture, training and evaluation of our machine learning (ML) model. The ML model is a gradient boosted classifier that operates on a set of preprocessed input vectors (rather than directly on the raw inputs). The preprocessing steps involve an ensemble of different pretrained models (including neural networks and gradient boosting algorithms) that transform the raw inputs (free text, questionnaire scores, behavioral indicators and demographic variables) into a set of numerical vectors. The model transforms these pre-processed inputs into a set of probabilities across a **Ranked Consideration Set** over the following 8 diagnostic categories. The ranking is sorted in descending order of probability.

- Depression
- GAD
- Social Anxiety
- PTSD
- Health Anxiety
- Panic Disorder
- OCD
- Phobia

Limbic Access then administers additional outcome measures, known as Anxiety Disorder Specific Measures (ADSMs), corresponding to the top 2 diagnostic categories in the Ranked Consideration set. The ADSMs are seamlessly presented during the chatbot flo w to the user, and appear as supplementary additional outcome measures to complete. The following ADSMs are useful as more sensitive measures for the purpose of detecting the following specific anxiety disorders. Here, we present the ADSMs utilised in the Limbic Access standard implementation.

- Social Anxiety: Social Phobia Inventory (SPIN)
- PTSD: PTSD Checklist for DSM-5 (PCL5)
- Health Anxiety: Short Health Anxiety Inventory (SHAI-18)
- Panic Disorder: Panic Disorder Severity Scale (PDSS)
- OCD: Obsessive-Compulsive Inventory revised (OCI-R)
- Phobia: Severity Measure for Specific Phobia (SMSP)

With additional data collected from the ADSMs, Limbic Access uses deterministic clinical logic to rank the consideration set into **Primary and Secondary Presenting Problems** which are presented to the clinician. This additional clinical logic analyses the patients scores on all administered questionnaires. The purpose of these ranked presenting problems is to enable and assist clinicians to make more accurate decisions around diagnosis and treatment pathways. Importantly, the output of the machine learning model is not used to deliver a diagnosis, rather it is used to select appropriate ADSMs which in combination with the clinical logic define a consideration set of Primary and Secondary Problem Descriptors.

1.2.1 Model training and testing

The labels used to train the model are actual diagnoses assigned by clinicians to patients that were referred through Limbic Access. Therefore, for each patient, we have a set of input data, and a human clinician-assigned diagnostic label assigned to them during the treatment in NHS Talking Therapy services. We trained the model using a total of 18,278 datapoints across 4 different NHS services from users who gave consent to the usage of this data. We trained and tested the model using 10-fold stratified cross validation with a multi-class log loss function. The training data for each fold was further split into a training set (90% of data) and a validation set (10% of data) which was used for early stopping to avoid over-fitting, and a single split was used to tune hyper-parameters. Within the training set (but not the validation or test set) we over sampled the less common diagnostic categories to match the count of the most common diagnosis (i.e. depression), to ensure that the algorithm would not over-optimise for the most common mental health diagnoses and neglect less common ones.

We refer to this dataset as the **historical dataset**, since this dataset was collected prior to and for the purposes of model training. A subset of this dataset is our ”test” data, and is expected to follow roughly the same distribution as the training data, and so the measurement of performance on this test data is the industry standard.

The most rigorous test of model performance and generalizability is to collect a **prospective dataset**, data that was collected after the ML model was trained and its weights were frozen. In a second study we tested these generalization capabilities in the most rigorous way with a prospective study including a total of 2,557 data points. Here the model made predictions about the consideration set, but these predictions were not used in any way to administer ADSMs - it was running in ”shadow” mode in our product. Out of these patients, 773 patients had finished their therapy so that an end of treatment diagnosis was available, hence we compared agreement of the model’s prediction with the final diagnosis to the agreement between therapists’ diagnosis at assessment and the final diagnosis, as a stringent measure of the model’s performance. Such a ”prospective” evaluation tests the model’s performance on new data, in order to ensure that the model generalizes well in the actual real-world application.

Finally, we also report the performance of the model on a **live dataset** with 890 datapoints i.e. data that was collected after the fully certified Medical Device Class II model was deployed and live in production, which means that the model’s predicted presenting problems were being used to administer ADSMs. This labelled dataset lets us compare the actual accuracy of the model’s predictions to the final clinical diagnoses, allowing us to evaluate the model’s real world performance.

1.2.2 Evaluation metric

Since the model is not aiming to select the single most likely presenting problem, standard evaluation metrics such as an ordinary F1-score are not well suited for evaluating the model performance.

The main evaluation metric of interest is the accuracy with which the algorithm would administer the relevant ADSM for a diagnosis, if that diagnosis was present. Formally that means, this is the percentage of times with which the actual diagnosis is within the top two problems in Ranked Consideration Set selected and ranked by our machine learning model. This metric is equivalent to a recall score and similar scores have been used in comparable settings of evaluating the performance of mental health symptom checkers (e.g. Hennemann et al. [2022]).

2 Results

2.1 Performance on historical data

On historical data, **the model achieved an overall accuracy of 93.5%** (CI=[93.1%, 93.9%]) in identifying the correct diagnosis for the 8 most common mental health disorders. As outlined above, this is the percent of times the top 2 problems in the ML model’s ranked consideration set contained

Diagnosis	Number of diagnoses	ML-model accuracy over folds
Depression	8966	mean=95.9%; (min: 94.4%, max=97.2%)
Generalised anxiety disorder	4974	mean=96.8% (min: 95.7%, max=98.4%)
Social phobia	872	mean=88.0% (min: 83.9%, max=91.9%)
PTSD	834	mean=83.0% (min: 77.3%, max=86.7%)
OCD	416	mean=73.1% (min: 63.4%, max=82.9%)
Panic disorder	338	mean=74.0% (min: 58.8%, max=88.2%)
Health anxiety	308	mean=76.0% (min: 64.5%, max=90.3%)
Specific phobia	109	mean=46.8% (min: 27.2%, max=72.7%)

Table 1: Cross-validation performance on historical dataset

the true diagnostic label given by the clinician. Moreover, it achieved good accuracy for each of the individual diagnostic categories as well, indicating that it has good performance. (Table 1)

2.2 Performance on prospective data

The model achieved a similar level of accuracy on our prospective evaluation - **achieving an overall accuracy of 94.2%** (CI=[93.3%, 95.1%]) for detecting the 8 most common mental health problems. Similarly to the test and training dataset, this accuracy did hold for each of the relevant diagnoses (Table 2).

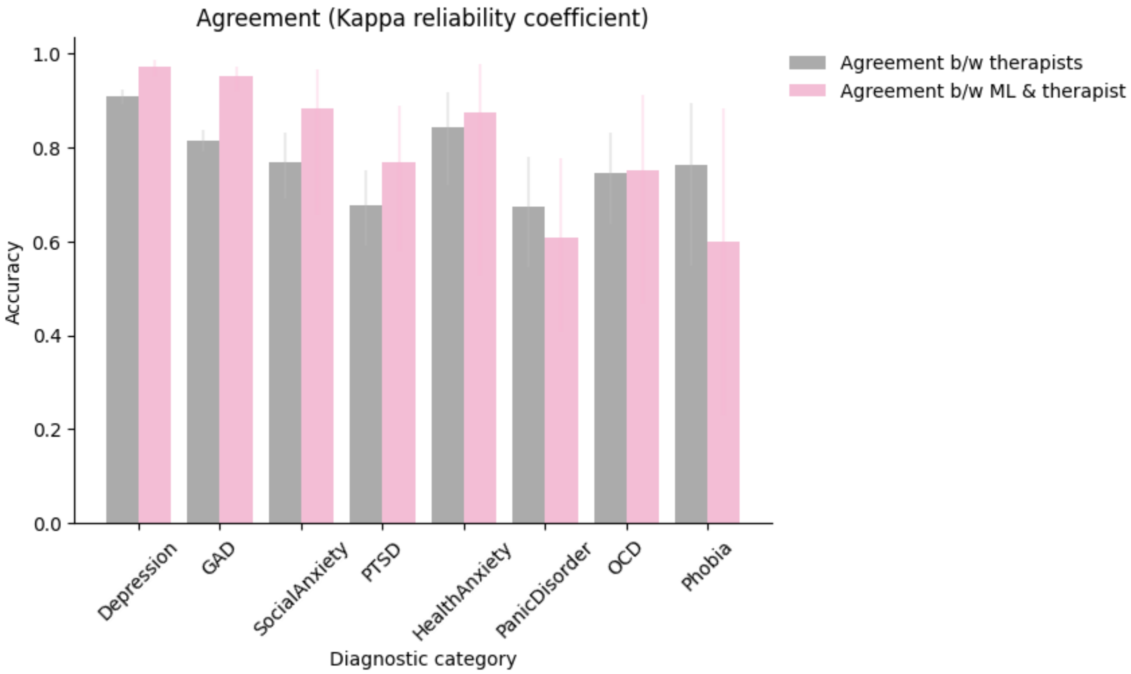


Figure 2: **Comparison of model performance to clinician reliability on prospective data** In pink, we present the agreement between our ML-model predictions and diagnoses assigned by therapists during treatment. In grey, we present the agreement between the therapists' diagnosis at assessment and the diagnosis at the end of therapy. Agreement was calculated as a Kappa reliability coefficient, and the error bars represent the 95% confidence intervals

Impressively, the overall agreement between the model's prediction and the final diagnosis was 93.7%, while the agreement between the human diagnosis at clinical assessment and the end of treatment diagnosis was only 85.1% of all cases (See Table 3, Figure 2 for agreement on individual categories). This evaluation is the most rigorous approach for testing model performance and it is noticeable that even in this setting our model performed with excellent accuracy. This effectively gives us a real

Diagnosis	Number of diagnoses	ML-model average accuracy
Depression	1,432	97.5%
Generalised anxiety disorder	672	95.7%
Social phobia	145	91.7%
PTSD	132	83.3%
OCD	61	82.0%
Panic disorder	46	67.4%
Health anxiety	53	77.4%
Specific phobia	16	50%

Table 2: Performance on prospective dataset

Diagnosis	Model v final diag	Clin. assessment v final diag
Depression	97.3% (N=368)	90.8% (N=1513)
Generalised anxiety disorder	95.2% (N=289)	81.5% (N=1038)
Social phobia	88.2% (N=17)	76.9% (N=134)
PTSD	76.9% (N=26)	66.7% (N=127)
OCD	75% (N=12)	74.7% (N=75)
Panic disorder	60.8% (N=23)	67.2% (N=58)
Health anxiety	87.5% (N=8)	84.3% (N=51)
Specific phobia	60% (N=5)	76.1% (N=21)

Table 3: Agreement with final diagnosis on prospective dataset

measure of the model's performance for generalising in the context it will actually be used in and showcases human level performance.

2.3 Performance on live data

Finally, we examined the model's performance while it was in use in production. Unlike the previous two datasets, in the live dataset the model's consideration set predictions were used to administer ADSMs. Moreover, we used our Clinical Logic to rank the problems in the consideration set, utilising additional data gathered from the ADSMs as Primary and Secondary Presenting Problems.

Thus, this dataset offers us the opportunity to examine the accuracy of the model's real world predictions about the top presenting problems, and see how it measures up against 1) the Talking Therapy screening questions (typically manually administered by clinicians during patient assessments), and 2) actual patient scores from the model-chosen ADSMs, which determine if the predicted presenting problems are indeed above caseness cutoffs.

We found that compared to a clinical logic purely relying on screening questions, the Limbic Access predictions detected specific anxiety disorders more often in patients - particularly evident in PTSD, Health Anxiety, Panic Disorder and OCD (Figure 3). Importantly, this translated to a much higher accuracy for accurately identifying specific anxiety disorders (Table 4, Figure 4), suggesting that our model's increased detection was accurate and well-calibrated. This feature is in line with NHS Talking Therapies' objective of combating the under-diagnosis of specific anxiety disorders, and suggests that our model could help clinicians make higher quality assessments.

Moreover, when we compared the accuracy of presenting problems chosen by the model to the accuracy of presenting problems whose questionnaire scores were above clinical caseness cutoffs - we found that there was negligible loss of accuracy (Figure 5). In other words, the Limbic Access system consisting of a ML-model to select personalised ADSMs in combination with a clinical logic which uses ADSM cut-offs to determine a consideration set of Primary and Secondary Presenting Problems is a highly accurate way of detecting the presence of the 8 most common mental health conditions.
Overall, on the live dataset the **Limbic Access correctly detected 92.47% of diagnoses**, which is comparable to the performance on historical and prospective datasets.

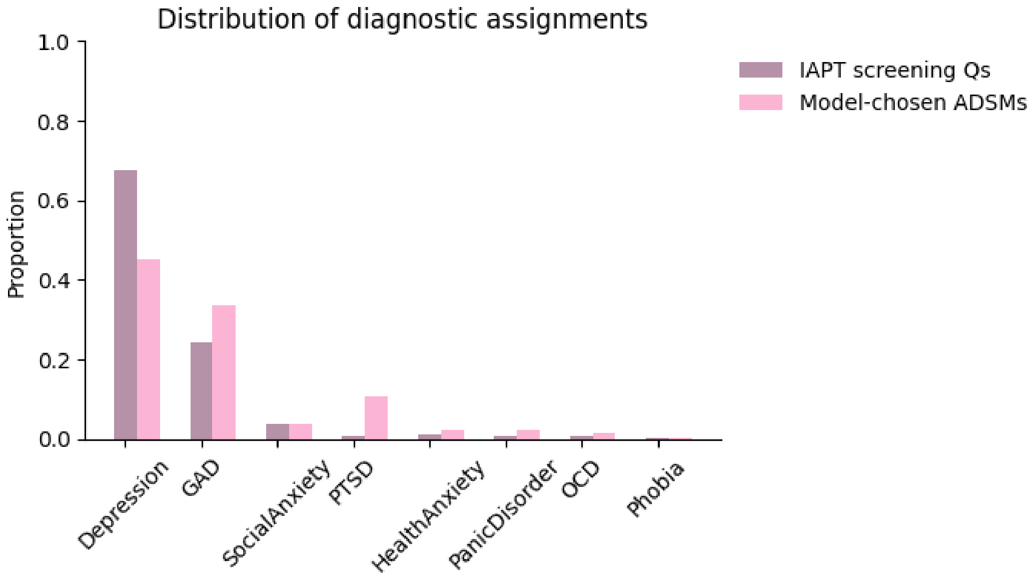


Figure 3: **Distribution of assignments** Distribution of assignments to different presenting problem when using the MD2 model that administers ADSMs (live dataset, pink) v.s. when using the IAPT screening questions (purple)

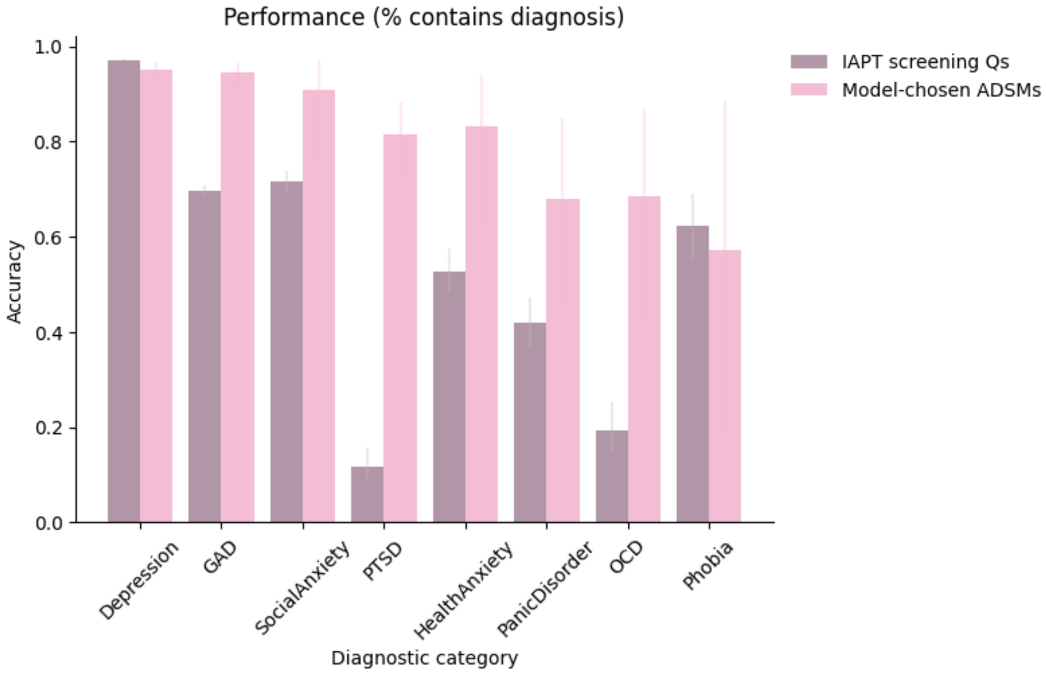


Figure 4: **Accuracy of assignments** Percentage that the final diagnosis is in the consideration set of the Limbic Access predictions (live dataset, pink) v.s. that of a clinical logic based on the IAPT screening questions (purple). The ML-model in Limbic Access vastly outperforms the screening questions in accuracy on specific anxiety disorders.

3 Discussion

In this study we describe a digital solution that we have developed to improve quality of clinical assessments and clinical efficiency, called Limbic Access. In addition to enabling self-referrals through a user-friendly interface for data collection and signposting, our tool uses a ML algorithm to identify

Diagnosis	Model chosen ADSMs	IAPT Screening Qs
Depression	95% (N=366)	97% (N=24972)
Generalised anxiety disorder	94.4% (N=273)	69.7% (N=9027)
Social phobia	90.9% (N = 30)	71.5% (N=1452)
PTSD	81.5% (N=88)	11.7% (N=322)
OCD	68.4% (N=13)	19.2% (N=204)
Panic disorder	68% (N=17)	41.9% (N=338)
Health anxiety	83.3% (N=20)	52.8% (N=451)
Specific phobia	57.1% (N=4)	62.3% (N=184)

Table 4: Performance of model-chosen ADSMs (live dataset) vs IAPT screening questions

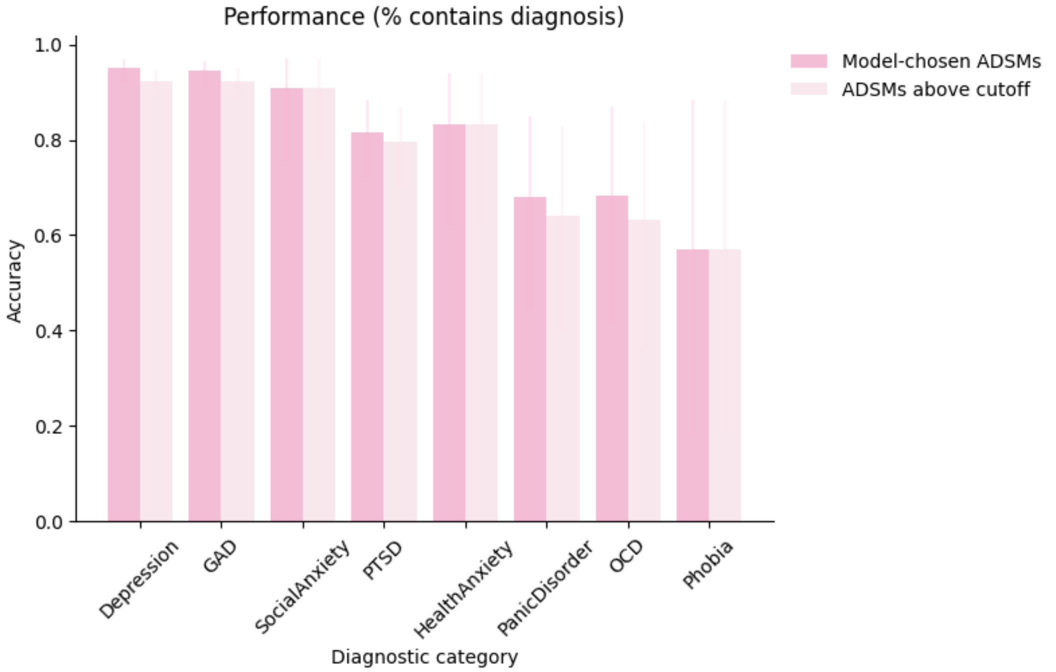


Figure 5: **Comparing predicted ADSMs vs actual caseness** Comparing the probability of the true diagnosis being contained in the model-predicted set of problem descriptors (pink) v.s. being contained in problem descriptors whose actual measured scores pass clinical caseness cutoffs (light pink). This suggests that the combination of ML-based administration of ADSMs and the usage of ADSM scores to identify Primary and Secondary Presenting Problems allows for highly accurate detection of the most common mental health diagnoses.

likely presenting problems based on patient inputs, and adaptively administers ADSMs that are personalised to the individual. The scores on these administered questionnaires are then processed with clinical logic to determine a consideration set of the most likely presenting problems which can inform the clinical assessment.

Importantly, the accuracy with which the algorithm selected ADSMs and the accuracy with which these ADSMs were used to determine the likely presenting problem are in range of human level performance. Across 3 studies, we found that the true diagnosis was contained within the top 2 items of the Ranked Consideration Set with an accuracy of around 93% - the model achieved an accuracy of 93.5% on historical data, 94.2% on prospective data (with an agreement of 93.7% with final diagnosis) and 92.47% on live data. Together, this evidence supports our approach of using the ML-model to administer personalized ADSMs and presenting clinicians with the entire set of model-informed Primary and Secondary Presenting Problems to assist in their assessments.

By appropriately administering ADSMs pre-assessment and accurately identifying presenting problems, Limbic Access has the potential to save clinicians time and support the quality of their assessments, as well as offer patients an enhanced referral experience and potentially better outcomes. Our tool highlights the role that technological solutions can play in improving clinical efficiencies.

References

S. Hennemann, S. Kuhn, M. Witthöft, and S. M. Jungmann. Diagnostic performance of an app-based symptom checker in mental disorders: comparative study in psychotherapy outpatients. *JMIR Mental Health*, 9(1):e32832, 2022.