scientific reports

Check for updates

OPEN

A multimodal Bayesian network for symptom-level depression and anxiety prediction from voice and speech data

Agnes Norbury[1], George Fairs[1], Alexandra L. Georgescu[1,2], Matthew M. Nour[3,4], Emilia Molimpakis[1] & Stefano Goria[1]✉

During psychiatric assessment, clinicians observe not only what patients report, but important nonverbal signs such as tone, speech rate, fluency, responsiveness, and body language. Weighing and integrating these different information sources is a challenging task and a good candidate for support by intelligence-driven tools—however this is yet to be realized in the clinic. Here, we argue that several important barriers to adoption can be addressed using Bayesian network modelling. To demonstrate this, we evaluate a model for depression and anxiety symptom prediction from voice and speech features in large-scale datasets (30,135 unique speakers). Alongside performance for conditions and symptoms (for depression, anxiety ROC-AUC = 0.842, 0.831 ECE = 0.018, 0.015; core individual symptom ROC-AUC > 0.74), we assess demographic fairness and investigate integration across and redundancy between different input modality types. Clinical usefulness metrics and acceptability to mental health service users are explored. When provided with sufficiently rich and large-scale multimodal data streams and specified to represent common mental conditions at the symptom rather than disorder level, such models are a principled approach for building robust assessment support tools: providing clinically-relevant outputs in a transparent and explainable format that is directly amenable to expert clinical supervision.

Psychiatric diagnosis represents one of medicine's most complex inferential challenges. During clinical assessment, psychiatrists must integrate multiple sources of information to arrive at diagnostic formulations that guide treatment decisions, in a process that amounts to inference to the best explanation given available evidence[1–3].

Psychiatry stands apart from other medical specialties in its reliance on subjective information sources. While neurologists integrate nerve conduction studies, brain imaging, and cerebrospinal fluid analysis with clinical presentation, psychiatric assessment depends primarily on clinical observation, patient self-report, and collateral information from family members or caregivers. The absence of established biological markers or objective diagnostic tests means that psychiatric diagnosis relies heavily on clinicians' abilities to synthesize complex, often ambiguous behavioural and self-report data[4].

The multimodal nature of psychiatric assessment reflects the complexity of mental health presentations. Experienced clinicians attend not only to reported symptoms, but to how these are communicated: including variations in tone of voice, speech rate, fluency, and responsiveness during conversational exchanges. Other important observations are body language, psychomotor activity, and cognitive processing as assessed through clinical interview and simple tasks[1,5]. These paralinguistic and behavioural cues—formalised in the "appearance and behaviour" sections of a typical mental state examination—are particularly valuable given that mental health conditions often impact executive functioning and insight[6].

Compounding this complexity, several factors influence the quality and consistency of assessment in a typical clinical setting. Clinicians often operate under significant time and capacity constraints[7,8]. The cognitive demands of gathering, weighing, and synthesizing multiple sources of uncertain information represent a substantial challenge even for experienced practitioners. Additionally, various forms of bias can affect which information is gathered, how it is weighted, and how diagnostic conclusions are reached—which can lead to inequities in both access to assessment and diagnostic outcomes[9–11].

[1]thymia Limited, London, UK. [2]Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. [3]Department of Psychiatry, University of Oxford, Oxford, UK. [4]Max Planck UCL Centre for Computational Psychiatry and Ageing, University College London, London, UK. ✉email: stefano@thymia.ai

We propose that digital tools can augment psychiatric assessment by quantifying the paralinguistic, cognitive, and behavioural signals that clinicians observe but cannot systematically measure. Rather than replacing clinical judgment, these tools offer a form of "digital phenotyping" that makes explicit and measurable the characteristics that experienced clinicians intuitively recognise as diagnostically relevant. Such tools could bring psychiatry closer to the quantified precision medicine approaches available in other medical specialties, whilst also also supporting richer symptom monitoring beyond the clinical encounter. To date, intelligence-driven approaches have shown promise in predicting disorder status from voice, speech, movement, and physiological data—however, their clinical translation remains limited[12]. Adoption of these approaches in the clinic has been hindered by a focus on binary condition classification rather than symptom- or sign-level assessment, a reliance on small sample sizes that compromises generalizability, poor transparency or explainability of underlying models, and lack of integration or congruence with existing clinical workflows[13–15].

Here, we argue that Bayesian network models with symptom-level estimation capacities are a promising approach for addressing these short-comings. These represent both a principled method for modelling the process by which clinicians must integrate across multiple noisy information sources during psychiatric assessment[16,17] and a natural way of accounting for the comorbidity and symptom co-occurrence patterns that are the norm rather than the exception in mental health[18]. From a clinical point of view, such models are able to deal well with heterogeneity within diagnostic categories[19], provide the granular information required for treatment planning and progress monitoring[20–23], and enable direct validation against clinical impressions and patient experience. From a modelling perspective, individual outputs are made more robust and explainable through integration of data-driven patterns with expert-derived knowledge, relative resilience to missing or noisy individual inputs, and transparent probabilistic reasoning. Crucially, Bayesian networks also support direct intervention in model predictions by supervising clinicians—for example as more information becomes available about the context of specific symptoms during evaluation—maintaining the primacy of expert judgement (from both clinicians and patients) during clinical decision-making[24].

To demonstrate these advantages in practice, we present a novel Bayesian network-based assessment for common mental health conditions (depression and anxiety) and their associated symptoms. We describe a model built using multimodal voice and speech features—although our approach would be naturally extendable to accommodate other relevant information sources such as cognitive test or passive-sensing data (see "Discussion"). Critically, this implementation leveraged multiple speech samples from over 30,000 unique participants, representing, to our knowledge, the largest dataset of its kind in psychiatric digital phenotyping research. This scale enabled application of rigorous best practices during both model development and evaluation. Specifically, sufficiently large samples were available for robust development and output calibration training, as well as independent test evaluation with appropriate numbers of observations across relevant demographic and clinical subgroups to support comprehensive fairness testing. This is vital for addressing both the documented brittleness of speech-based mental health prediction models trained on small, homogeneous samples[15,25], and concerns about generalizability and fairness in digital phenotyping tools more generally[13]. Alongside assessment of the true multimodal integration capacities of this model, we evaluate other characteristics of key translational relevance including clinical usefulness metrics and factors governing acceptability to mental health service users.

Results

A multimodal Bayesian Network for depression and anxiety prediction from speech and voice data

An overview of the model is provided in Fig. 1. Briefly, paralinguistic (acoustic and timing) and linguistic features were extracted from two different speech activities (reading out loud, and talking about recent mood). These were compressed into feature-type and symptom-specific representations using a set of surrogate models (neural networks trained to predict individual symptom presence from different subsets of paralinguistic and

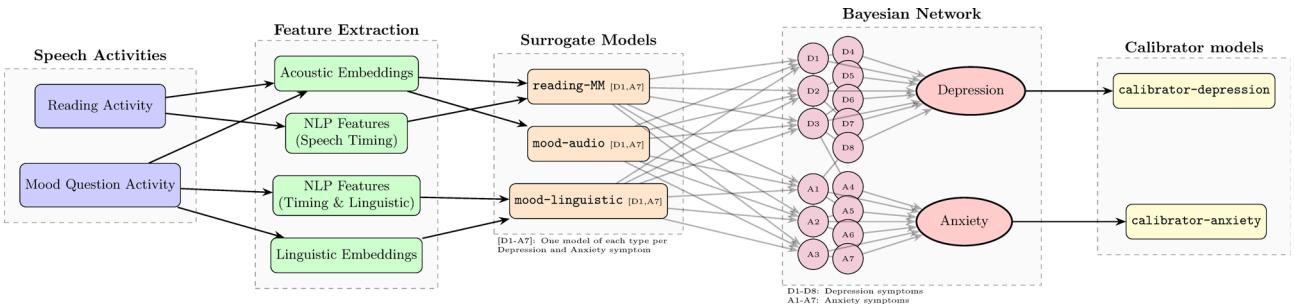

Fig. 1. Model overview. Speech activity data (reading out loud and answering a question about recent mood) are used to generate acoustic embeddings, speech timing, and linguistic feature sets (semantic embeddings and Natural Language Processing features), which are fed into relevant surrogate models to generate multiple predictions for each individual depression and anxiety symptom (for details, see Fig. S1). Symptom-level predictions are passed to a Bayesian Network, which specifies mapping weights of surrogate predictions to symptom severity estimates, inter-symptom relationships, and symptom severity to overall condition probabilities (simplified sketch of network architecture; for details see Fig. S2). Finally, condition probabilities are passed through a calibration layer to ensure meaningful output scores.

		Development (Uncalibrated)	Test (Calibrated)
Depression	ROC-AUC	0.837	0.842
	ECE	0.060	0.018
Anxiety	ROC-AUC	0.836	0.831
	ECE	0.090	0.015

Table 1. Model discrimination and calibration performance for overall condition probabilities for depression and anxiety. The development set test split (depression, anxiety $N = 2251, 2191$) was unseen during surrogate model training but used for development of the final Bayesian Network model architecture. The held-out test set (depression, anxiety $N = 1489, 1477$) was completely unseen during model training and development, with disorder probabilities calibrated using calibrator models trained on a further held-out calibration set. *ROC-AUC*, Area Under the Receiver-Operator Curve; *ECE*, Expected Calibration Error.

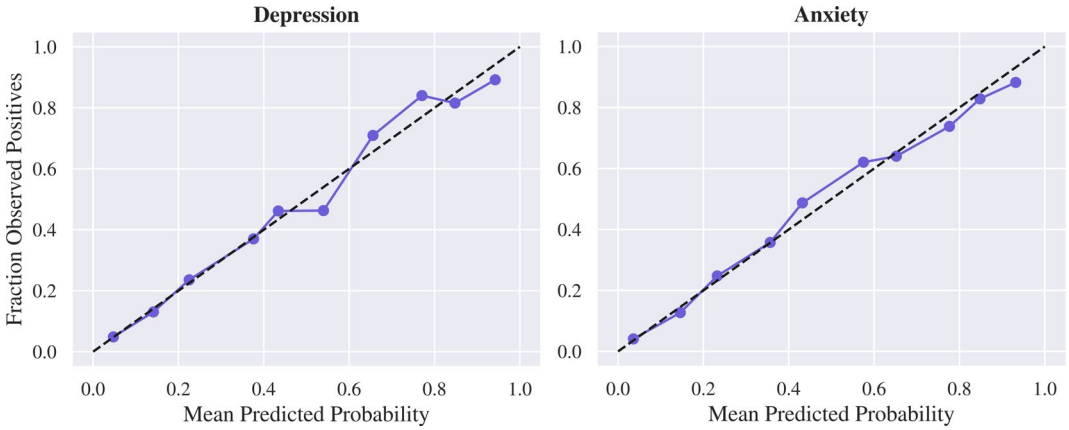

Fig. 2. Model calibration for overall depression and anxiety and status. Plots represent predicted probability score ranges *vs* observed positive case rate values across the full range of output condition probabilities in the held-out test set. The dotted line at $y = x$ represents performance of a perfectly calibrated model.

linguistic speech features). A joint Bayesian Network for depression and anxiety symptoms, alongside overall condition probabilities, was specified using insights from the clinical literature and parameterized using a dataset of surrogate model predictions, symptom and condition level ground truth values. This allowed the network to learn in parallel posterior probability distributions for the strength of relationships between different data types and true symptom severity scores, relationships between different symptoms, and relationships between symptom severity states and overall condition probabilities. In a final step, calibrator models were trained to ensure that output condition probability scores were well-aligned to observed case frequencies for depression and anxiety, respectively.

Development data for surrogate models and Bayesian Network training and evaluation were available from $N = 21,379$ participants. A separate calibration set from $N = 6,325$ participants was available for training of the calibrator models. Final performance evaluation was carried out in a held-out test set ($N = 2,431$ participants), which was completely unseen during model training and development. For a description of participants in each dataset, please see Table S1.

Model performance for conditions

Discrimination and calibration performance for overall condition status (depression and anxiety) are described in Table 1. These two metrics assess important independent model qualities: that the model is able to distinguish well between users with different condition statuses, and that output probabilities are meaningful (i.e., that a 75% probability score actually corresponds to 75% positive condition rate)[26,27]. Performance results in the held-out test should be interpreted as representing generalized estimates of model performance, with results from the development set test split (data unseen during surrogate model training but available during Bayesian Network development) provided for reference and as an indicator of performance stability.

Although specific benchmarks for interpreting discrimination and calibration performance measures are somewhat arbitrary and inconsistently applied in the literature, as a general rule-of-thumb ROC-AUC values above 0.80 and ECE values below 0.05 are often considered to be properties of clinically-useful prediction models[26,28]. Full calibration curve plots for condition probabilities (based on calibrated predictions in the held-out test set) are shown in Fig. 2.

Model performance for individual symptoms

Discrimination performance for individual depression and anxiety symptoms is described in Table 2. These metrics quantify the ability of the model to discriminate between cases of significant symptom presence *vs* absence (based on the threshold of symptom experience on "half or more days" *vs* "less than half of days" over the last 2 weeks, as described in DSM criteria for depressive and anxiety disorders, and applied in previously-published algorithms for converting combined symptom severity/frequency ratings on self-reported questionnaires to DSM-like criteria[29]).

Classification performance for individual symptoms was generally in the fair-to-good range (> 0.70 ROC-AUC[26]), with stronger performance (~ 0.75 ROC-AUC) observed for "core" symptoms (anhedonia, low mood, anxiety/nervousness, and uncontrollable worry). There was evidence that some symptoms may may benefit from augmentation of the network with non-speech data sources: for example cognitive task data may provide additional information about concentration problems and restlessness symptoms (see "Discussion").

Dealing with clinical heterogeneity

Given the heterogeneity of common mental health conditions in terms of observed symptom profiles (particularly for depression[19]), we examined robustness of discrimination performance to different condition definitions and presentation types (where possible, using previously-established algorithms for defining these from self-report data).

For DSM-like Major Depressive Disorder, defined as 5 or more depressive symptoms having been present at least "more than half the days" in the past 2 weeks, 1 of which must be depressed mood or anhedonia[29]), ROC-AUC values were 0.836, 0.852 in the development and test sets, respectively.

For DSM-like Other Depression, defined as 4 or more depressive symptoms having been present at least more than half the days in the past 2 weeks[29], ROC-AUCs were 0.814, 0.829, respectively.

For DSM-like Generalized Anxiety Disorder, no previously-published algorithm exists, so this was derived from DSM-5 criteria (which specify 5 or more anxiety symptoms having been present at least "more than half the days" in the past 2 weeks, 1 of which must be feeling nervous, anxious or edge or uncontrollable worry), ROC-AUCs were 0.819, 0.835, respectively.

This reveals that, whilst performance is slightly superior for typical depression presentations (involving significant alterations in mood and motivation levels), it remains within strong ranges for less typical presentations (e.g., those more dominated by somatic changes). In general, a model which is able to maximise performance across different possible individual symptoms should be better placed to deal with heterogenous presentations than one which simply predicts overall condition absence or presence according to sum severity cutoff values.

Condition severity

The Bayesian network analysed here represents symptoms not just in terms of binary presence or absence, but specifies posterior probability distributions across 4 possible severity categories. These can be used to calculate the expected severity level for each symptom, which can then be summed across conditions to output overall disorder severity predictions alongside condition probabilities. To validate these severity estimates, predicted values were compared to observed PHQ-8 and GAD-7 total scores (which similarly represent overall symptom load based on number of observed symptoms and their severity). Condition severity predictions were found to be strongly associated[30] with these scores: for depression, Pearson's $r = 0.526, 0.551$; for anxiety, $r = 0.514, 0.513$ in the development test split, and held out test set, respectively.

We also assessed whether severity estimates were related to other important patient-reported outcomes, in particular the impact of experience of mental health symptoms on quality of life and psychosocial functioning. Mental-Health related Quality of Life[31] scores were available in a subset of development set test split participants (*N*=804), and observed to be correlated with predicted depression severity at $r = -0.49$ and with anxiety severity at $r = -0.50$. In the held-out test set (*N*=2413), PHQ psychosocial functioning (level of impairment in work, home, and social life[29]) and the CDC's healthy days (number of recent days where poor physical or mental health limited usual activities[32]) measures were available: and were related to model-predicted depression and anxiety severity at $r = 0.47, 0.47$ and $r = 0.44, 0.42$, respectively.

	Depression symptoms							
	Anhedonia	**Low mood**	**Sleep**	**Low energy**	**Appetite**	**Worthlessness**	**Concentration**	**Psychomotor**
Development	0.732	0.785	0.709	0.734	0.697	0.761	0.696	0.721
Test	0.741	0.801	0.706	0.734	0.690	0.757	0.667	0.714
	Anxiety symptoms							
	Nervousness	**Uncontrollable worry**	**Excessive worry**	**Trouble relaxing**	**Restlessness**	**Irritability**	**Dread**	
Development	0.761	0.751	0.747	0.724	0.696	0.718	0.705	
Test	0.749	0.746	0.736	0.739	0.680	0.706	0.721	

Table 2. Model performance for individual depression and anxiety symptoms. Values represent ROC-AUC scores for symptom presence (for half of days or more) *vs* absence (less than half of days), in the development set test split (*N* = 3372, 3378 for depression, anxiety symptoms) and the fully held-out test set (*N* = 2426, 2424 for depression, anxiety symptoms).

Fairness

A key requirement for digital diagnostic support tools is to ameliorate (and not exacerbate) existing structural inequalities in diagnostic practices and healthcare access[33]. We therefore robustly assessed model performance for potential differences in performance for members of different demographic groups. Group-level discriminative performance (ROC-AUC values) and between-group differences in calibration metrics (ECE), combined calibration-accuracy measures (Brier scores), and outcome fairness (equalized odds ratios, which assess potential differences in true positive, true negative, false positive, and false negative rates between groups) are reported in Table 3.

Within all examined subgroups, ROC-AUC for both depression and anxiety remained ≥ 0.80, indicating maintenance of clinically-appropriate performance levels for specific groups. This also implies that model performance is unlikely to be significantly driven by between-group differences in voice or speech features, which in combination with differential rates of experience of common mental health problems between some groups has previously been shown to inflate performance estimates for some speech-based models[34].

There was some evidence of differences in probability calibration (Brier scores) between birth sex, gender and chronic health condition groups—although these differences were all within the approximately 20% range often used as a rule-of-thumb for fairness assessment (although this is somewhat a blunt heuristic which should take into account the context in which a particular tool will be used[35]). These could be addressed in the future using demographic-group specific calibration strategies. Minimal differences in calibration accuracy were observed for other tested groups including race/ethnicity, accent, and testing device type.

Importantly, differences in calibration properties were not in most cases associated with strong differences in outcome fairness, as assessed by differences in equalized odds ratios. This measures a model's risk for misclassifcation bias, or that members of different groups will experience differential allocation of treatment outcomes. For this metric, most group differences were in excellent (< 0.05) or good (< 0.10) ranges (representing $< 5\%$ or $< 10\%$ differences between groups), with only one tested difference exceeding the 20% threshold. Differences in equalized odds ratios for anxiety between birth sex groups were found to be driven by higher true positive rates for females than males, which may be related to substantially higher base rates of anxiety for women *vs* men in our training data. This difference warrants further attention and may benefit from mitigation strategies such as sex-based stratification of network training data and/or sex-specific calibration strategies for anxiety. Of note, equalized odds ratio measures are dependent on the specific threshold used for classification of cases *vs* non-cases in the analysis, which can also be adjusted depending on required use-case characteristics.

		Depression					Anxiety				
	Group	*N*	ROC-AUC	ECE difference	Brier score difference	Equalized odds ratio difference	*N*	ROC-AUC	ECE difference	Brier score difference	Equalized odds ratio difference
Age	< 35	720	0.826	0.006	0.033	0.099	705	0.814	0.002	0.042	0.047
	≥ 35	769	0.848				772	0.836			
Birth sex	Female	974	0.838	0.036	0.037	0.111	972	0.825	0.017	0.044	0.251
	Male	515	0.828				505	0.800			
Gender identity	Woman, non-binary	975	0.834	0.038	0.043	0.067	974	0.820	0.022	0.052	0.177
	Man	514	0.842				503	0.812			
Race/Ethnicity	White	1086	0.850	0.055	0.009	0.024	1068	0.843	0.061	0.005	0.060
	Other	404	0.840				409	0.814			
Accent	UK	617	0.835	0.016	0.018	0.085	609	0.821	0.015	0.030	0.167
	US, other	862	0.846				860	0.831			
Chronic health condition	No	894	0.839	0.029	0.050	0.038	887	0.835	0.024	0.054	0.050
	Yes	595	0.840				590	0.814			
Device type	Laptop	876	0.831	0.006	0.008	0.054	880	0.839	0.003	0.028	0.106
	Smartphone, tablet	607	0.856				591	0.819			

Table 3. Differences in discrimination and calibration performance for anxiety and depression across different demographic groups. Results are for calibrated condition probabilities in the test set. *ROC-AUC*, single group discrimination performance for each condition; *ECE Difference*, difference in Expected Calibration Error between groups; *Brier Score Difference*, difference in Brier scores (a combined discrimination and calibration metric that measures accuracy of probabilistic predictions) between groups; *Equalized Odds Ratio Difference*, difference in a combined measure of true positive, true negative, false positive, and false negative rates between groups. A 0.05 difference in Brier score means predictions differ by approximately $\sqrt{0.05} = 0.22$ on a probability scale; equalized odds ratio differences can be directed interpreted on a probability scale.

Model properties

Relationship between conditions

Depression and anxiety often co-occur. For example, lifetime prevalence estimates for proportion of people with a depressive disorder who also meet diagnostic criteria for an anxiety disorder range from 49% to 81%; with similar rates for meeting depressive disorder criteria in people with an anxiety diagnosis (47% to 88%)[18,36].

A joint model of depression and anxiety should take into account this comorbidity, whilst also allowing for condition-specific sensitivity in predictions (cases where anxiety is present but not depression, and vice versa). When marginalizing over all other model states, the posterior conditional probability for the presence of anxiety, given presence of depression in our network was 0.409, and the probability for the presence of depression, given the presence of anxiety, was 0.424. This degree of separation was influenced by cross-correlation of condition states in our training data, given the presence of regularizing Bayesian priors (see Methods). A strength of the Bayesian Network framework is that it allows for enforcement of different conditional probability estimates between conditions: either by applying different prior settings during training, or intervening directly on posterior estimates, if this is desired for different populations or implementation settings.

Multimodal integration at the symptom level

We have argued that Bayesian Networks represent a principled method for integrating information about mental health symptoms across different data sources: by identifying which inputs carry greater signal for different symptoms, and integrating across these in an appropriate way to construct more accurate output estimates. Examining characteristics of our network inputs and inspecting the conditional posterior probability distributions it constructs from these during training allows us to see how well this works in practice.

Different surrogate model performance for individual depression and anxiety symptoms is described in Table S2. For every symptom, network performance is superior to that of any individual surrogate: suggesting the model successfully arbitrates between and integrates across the multiple noisy observable inputs it receives. For example, for sleep problems, surrogate model predictions with discriminative performances of 0.62, 0.66, and 0.68 (`reading-MM`, `mood-audio`, and `mood-linguistic`, respectively) are weighted and combined by the network to produce a final performance of 0.71 ROC-AUC (Table 2). Figure 3 shows that the network assigns different posterior probability weights to different sleep symptom severity states for each surrogate model input: with higher posterior probabilities assigned to more discriminative surrogates for that symptom.

Relative contribution and redundancy between different input types

Table S3 shows performance of the Bayesian Network when queried only with evidence from each surrogate model type (reading activity paralinguistic feature, mood activity paralinguistic feature, and mood activity linguistic feature predictions). Discriminative performance (ROC-AUC) for conditions was superior for `mood-audio` and `mood-linguistic` model predictions, compared to `reading-MM` predictions alone (although all were inferior to querying with the full set of surrogate types). Raw calibration performance (ECE) was superior when providing either paralinguistic compared to linguistic only model predictions.

It is important to note that the mood activity linguistic features may contain a natural semantic overlap with some PHQ/GAD items, as during this activity participants were explicitly asked to describe their mood over the past two weeks. Despite this advantage, we observed that providing the network with predictions from both paralinguistic surrogate types (`reading-MM` and `mood-audio`) resulted in matched discriminative performance, and improved calibration performance, to `mood-linguistic` predictions only for depression (both ROC-AUC$= 0.783$; for anxiety ROC-AUC was slightly inferior at 0.783 *vs* 0.805).

This indicates evidence of built-in redundancy—or that good signal is available from multiple input data types (here, both paralinguistic and linguistic speech features). This is a desirable model property, as it helps build robustness to noise or fragility in any individual data type. For example, whilst linguistic-only models can show strong average discriminative performance for depression status when elicited semantic content directly

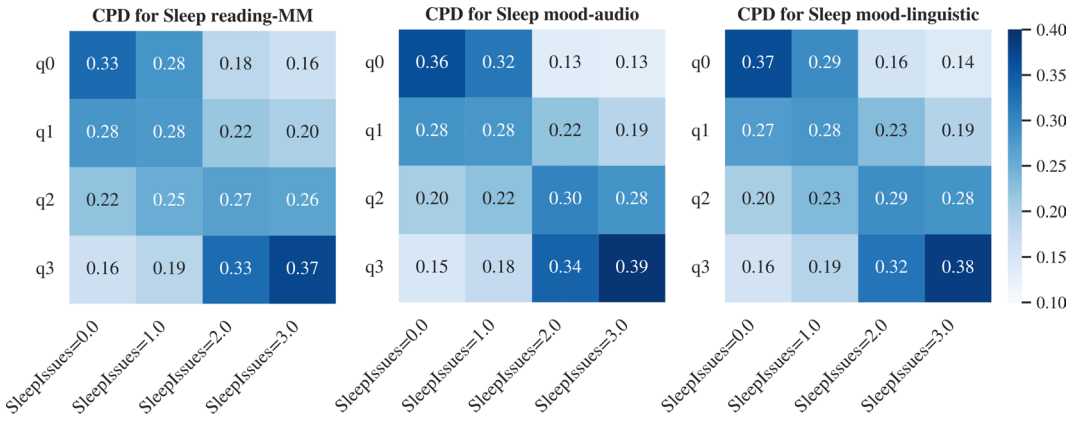

Fig. 3. Multimodal integration. Example posterior Conditional Probability Distributions (CPDs) for sleep symptom severity states (*SleepIssues = 0–3*) for different surrogate model types (*q0* = lowest quartile, *q3* = highest quartile of predicted symptom probability categories for each model).

relates to indicative symptoms, they are vulnerable to failure if particular individuals choose or neglect to disclose relevant semantic cues. Of note, our chosen network architecture enforced a design constraint whereby every symptom must be informed by at least one paralinguistic input, preventing over-reliance on linguistic features that may be fragile to individual disclosure patterns, cultural differences in expression, or vocabulary choices. Given a key feature of Bayesian Network models is resilience to missing information, this redundancy also allows the model to still produce meaningful outputs (albeit at slightly degraded performance levels) if, for example a user or clinician decides that a particular input activity type should not be used for inference.

Intervening in model predictions: clinician-in-the-loop

Other key properties afforded by the Bayesian Network architecture are explainability: or explicit reporting of the contribution of individual symptom severity estimates to overall condition probabilities, and modifiability: or the ability of a supervising clinician to intervene directly in model predictions on the basis of follow-up discussions with their client. A vignette of this iterative update process, which allows the clinician and client to collaboratively update predictions based on more in-depth evaluation of individual problem areas is described below and in Fig. 4.

Intervention by do-operation vignette. Sleep disturbances and low energy are common in people experiencing both depression and anxiety, and are therefore often influential symptoms in condition network models. However, they can also be related to external life factors that in some cases may mean either clients or consulting clinicians do not consider it appropriate to evaluate them as mental health-related symptoms. For example, a user may complete an initial screening assessment which indicates moderate to high probabilities of depression and anxiety. On reviewing the results, the clinician is able to see that this is mainly driven by high severity estimates for sleep and fatigue symptoms. When discussing this result with the client, they discover that they have been sleeping very poorly over the past week or so due to caring for an unwell dependent. The clinician is able to incorporate this information into model predictions by isolating the sleep symptom node from the rest of the network (in Bayesian Networks, this can be achieved straightforwardly using a causal inference method known as a do-operation). This removes the influence of this symptom in an updated set of symptom-level and overall condition probabilities.

Clinical usefulness and acceptability

An assessment support tool of the type described here will only useful if it can both provide actionable information and is acceptable to end-users.

Screening tool usefulness: prevalence-dependent performance metrics

In order to assess the potential usefulness of model outputs in a real-world clinical setting, we supplemented our previous performance analyses by calculating prevalence-dependent metrics for calibrated condition predictions in the held-out test set. The case rates for anxiety and depression in our test data were $\sim 30\%$: a reasonable

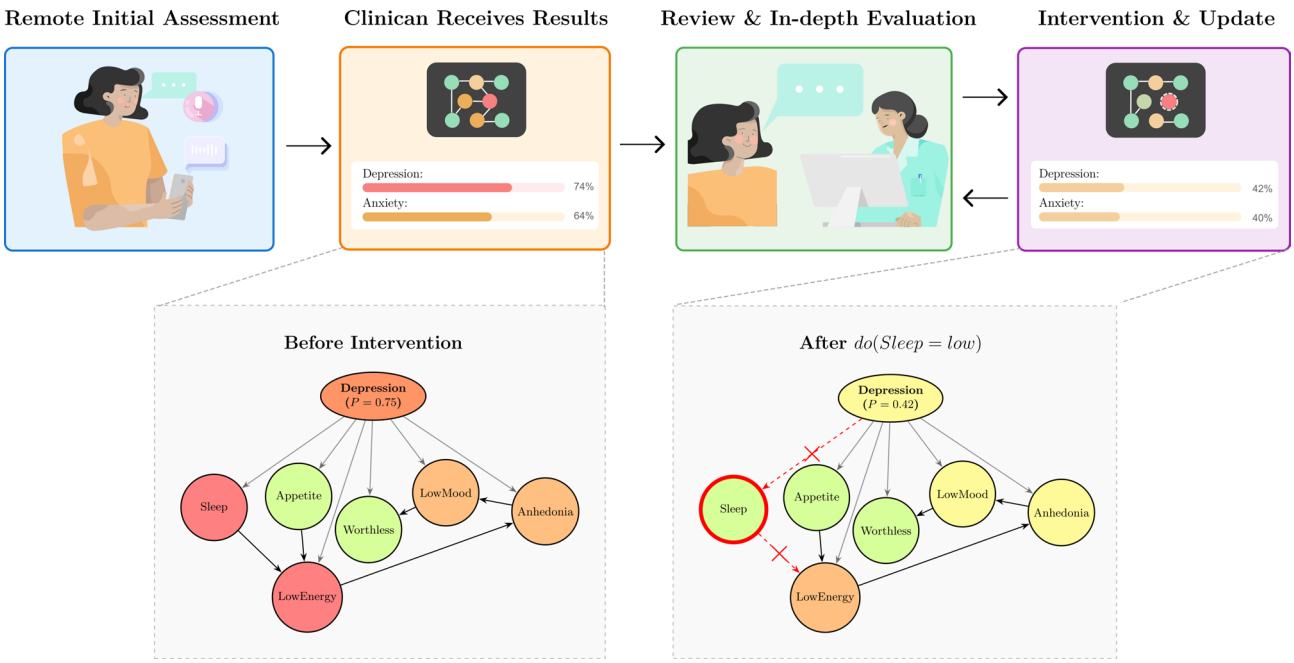

Fig. 4. Intervening on network predictions. Example of direct clinician intervention in model predictions based on follow-up discussions with a patient or client, using do-operation. For accompanying vignette, please see main text. Insets show a toy example for a subset of Bayesian Network depression symptom nodes, illustrating the effect of isolating sleep symptom predictions from the network after this has been evaluated as better explained by contextual rather than mental health-related factors.

	PPV	NPV	LR+	LR-
Depression	0.69	0.83	4.7	0.46
Anxiety	0.71	0.80	5.4	0.55

Table 4. Prevalence-dependent performance metrics for overall condition predictions. *PPV*, Positive Predictive Value (probability that a person from this population with a positive test result has a condition); *NPV*, Negative Predictive Value (probability that a person from this population with a negative test result does not have the condition); *LR+*, Positive Likelihood Ratio (change in odds of having a condition after receiving positive test result); *LR-* Negative Likelihood Ratio (change in odds of having a condition after receiving a negative test result).

approximation for primary care settings in the UK, where it has been estimated that 4 in 10 consultations involve a mental health element[37]. For calculating prevalence-dependent metrics analysis, we used a threshold of 0.5 for classifying predicted probabilities as positive *vs* negative cases—different thresholds may be appropriate for different use cases, based on the relative costs and benefits of false positive *vs* false negative results (Table 4).

Positive predictive value (PPV) is the proportion of true positives of all predicted positive cases. A PPV of ~ 0.70 suggests that , at a threshold of 0.5, the model produces some false positive results. However, this is usually considered acceptable for initial risk screening or population monitoring applications, where the cost of missing cases typically outweighs that of false alarms. Higher negative predictive value (NPV) scores suggest that the model is reliable at ruling out conditions when the ground truth is negative. Positive Likelihood Ratio (LR+) scores of ~ 5 can be considered strong: meaning that a user is 5*x* more likely to have a condition after receiving a positive test result. Negative Likelihood Ratios (LR-) are moderate: with scores of ~ 0.5 implying half the chance of having a condition after a negative test result. This means that a negative test result is not definitive in ruling out a condition (at least, in a population with a relatively high baseline prevalence for common mental health conditions).

Overall, these scores suggest good properties for an initial screening, triaging, or population-monitoring tool: with positive results requiring clinical follow-up for confirmation. Provision of well-calibrated condition probability scores allows stakeholders to select thresholds for further investigation most appropriate for their particular context (for example, via decision curve analysis).

Service user consultation and comparison to care as usual

Based on previous descriptions of the experiences of clinicians and clients with standard monitoring approaches embedded in the UK NHS Talking Therapies program[38–41], we asked a sample of participants with lived or living experience of both common mental health problems and treatment within the UK health system to provide feedback on a proposed speech-based tool for screening and monitoring of depression and anxiety symptoms, in comparison to care-as-usual with standard questionnaires.

Survey participants ($N = 230$) are described in Table S4. All participants had a diagnosis of a mental health condition (the most common being anxiety and depressive disorders), and had tried to access mental health support via the UK public health system (NHS) in the past 12 months. 96% of participants reported that they were somewhat or very familiar with mental health screening and monitoring via standardized questionnaires.

On average, survey respondents endorsed (agreed with) a range of opinions regarding strengths and weaknesses of the different assessment measure types (Fig. S4). For standardized questionnaires, the most strongly endorsed strengths were *being able to take time filling them out (without feeling pressured)*, and the *predictability of content*. Consistent with previous qualitative research in UK mental health service users, the most commonly endorsed weaknesses were being *unable to add nuance or explain the context behind answers*, and *completing them feeling like an impersonal tick-box exercise*. Participants who were very familiar with standard screening measures (filling them out regularly) showed a tendency towards more negative overall views of them, compared to participants who had only completed them a few times (Mann-Whitney U=6980, p=0.05, rank-biserial correlation effect size=-0.146 on a 7-point preference scale). For voice-based screening tools (such as the one described here), the most strongly endorsed strengths were *being able to capture non-verbal as well as semantic state information*, and *not needing to mentally average the intensity and frequency of symptoms over the past weeks* into a single rating number. More commonly endorsed concerns were *tool reliability*, the *possibility of technical issues affecting results*, and *having enough time to give considered spoken responses*. In this sample, average endorsement of potential strengths was numerically higher than for these concerns.

Overall, initial responses to multimodal voice and speech-based symptom monitoring approaches were positive (71% of respondents were excited by or somewhat interested in the idea), with significant proportions of respondents either unsure (19%) or somewhat or very concerned about the idea (6 and 4%, respectively). Qualitative analysis of themes identified in open-ended feedback regarding features required for trustworthy implementation of voice-based mental health screening technologies is available in Table S5. The most commonly raised themes were privacy, security, and control over use of speech recording data; provision of sufficient evidence of tool accuracy and reliability, including for diverse demographic groups; and endorsement by healthcare professionals or public health bodies. 7% of participants reported a general reluctance to use any form of AI-based screening tool, on the basis of their current experience with such technologies.

These results provide a limited view of overall appetite for and concerns around implementation of speech and voice-based screening technologies, given that they are from a small, self-selected sample of research participation platform users, which was predominantly White and generally comfortable with technology use

(see Table S4). However, they do help identify some important considerations for any potential implementation of these methods in clinical practice. Specifically, that key to the realization of the potential benefits of these methods are that users retain control over how their voice data is stored and used, and how and when it is shared with their clinical team, that they should supplement rather than displace clinical contact time, and that traditional options remain available as an alternative for users with legitimate concerns that such methods will work for them[33].

Discussion

Here, we evaluated a novel approach to supporting assessment for common mental health problems, using a Bayesian network model built from multimodal voice and speech features to explicitly represent individual depression and anxiety symptoms. Analysis of model properties showed that the network was successful in arbitrating between and integrating across multiple noisy input data sources for each symptom—resulting in excellent discrimination and calibration properties for both conditions, and good performance for the majority of individual symptoms. Providing symptom-level outputs is crucial for clinical adoption of assessment support tools in mental health care, as symptoms, not diagnoses often guide clinical practice in psychiatry[20]. As symptoms both occur across diagnostic boundaries and differ from each other in terms of their impact on general functioning, relationships to life events, and responsiveness to different treatments options—being able to monitor individual symptom trajectories over time is vital for effective treatment management[21–23]. A desirable feature of intelligence-driven psychiatric support tools is that they directly quantify the contribution of different symptom severity estimates to overall condition risk in a modifiable way: in order to allow supervising clinicians (in dialogue with their clients) to maintain decision-making autonomy and preserve the therapeutic relationship from threats associated with fully automated or opaque assessment methods[14]. In our implementation, capacity for providing accurate symptom-level readouts was supported by the use of a layer of surrogate models to compress rich feature input spaces to dimensions suitable for triangulation by a Bayesian network arbitrator—a scalable method that is naturally extendable to other relevant sources of information (for example cognitive test features for concentration or motivational problems or passive-sensor data for somatic symptoms such as restlessness and sleep disturbance[42]). The underlying Bayesian network architecture intrinsically supports direct clinical intervention at the individual symptom level, for example by isolating those better explained by contextual life factors than current mental wellbeing from network predictions.

Underwriting the potential success of intelligence-driven tools in healthcare is the availability of data with sufficient depth, breadth, and diversity to both maximize generalizability and support rigorous testing of important properties such as calibration accuracy and demographic fairness. Here, we were able to make use of one of largest to our knowledge consented and labelled mental health voice datasets (representing over 30,000 unique speakers, around a third of whom had a mental health diagnosis), to carry out robust development, calibration, and independent evaluation testing[15]. This degree of scale is likely to be necessary for satisfying recent best practice guidelines for health-related prediction models[43] as well as emerging regulatory standards for AI-based tools (for example, the recent ISO/IEC 42001 framework for AI governance[44]).

Other strengths of our results include resilience of overall condition predictions to heterogeneity in symptom presentations, and demonstrated redundancy across different speech and voice input modalities. This is an important feature of multimodal models: as evidence of relevant signal across different input types (here, paralinguistic *vs* linguistic feature predictions) reduces exposure to potential fragilities in any individual data source (for example, high background noise for acoustic features, or unwillingness to disclose sensitive information, insight problems, or cross-cultural expression differences for linguistic features). Critically, in-depth demographic fairness assessment revealed good-to-excellent properties in terms of potential for outcome allocation bias across age, gender, and race/ethnicity groups: with some evidence that demographic-specific calibration strategies may be useful for mitigating sex-based differences for anxiety. Performance was also found to be well-matched for individuals with and without comorbid chronic health conditions which are likely to impact voice (most commonly in our participants, asthma or other respiratory disorders and cardiovascular conditions; Table S1), and across recording device types (mobile devices *vs* laptops). Finally, condition severity estimates (calculated as the sum of predicted severities across individual symptoms for each disorder) were found to be strongly associated with important patient-reported outcome measures (quality of life and psychosocial functioning scores)—indicating validity of predictions in terms of symptom impact on functioning.

The model presented here also has some important limitations. In order to facilitate scale, datasets used self-reported information for current experience of depression and anxiety symptoms, as well as mental health diagnostic history. Future validation efforts incorporating and testing against clinician-provided diagnostic labels and symptom estimates would therefore be valuable. Further, additional testing of output robustness to a wider range of accent types, and in non-first language speakers, would be important prior to deployment in real-world clinical settings, where language familiarity and proficiency may be pre-existing barriers to access. Finally, it is vital that the introduction of new digital support tools occurs within the context of repeated acceptability and usefulness testing with the people who will use them, which here includes both healthcare practitioners and mental health service users[45]. Whilst both existing literature and our service user survey results revealed that, for some people, new assessment tools may help address perceived shortcomings in current screening and monitoring practices—implementations should be mindful of user requirements for trustworthy systems. For example, whilst time to complete is a factor affecting the acceptability of standard monitoring measures (particularly if it is felt to detract from within-appointment care time[39,41]), our survey results indicated that for speech and voice-based tools, concerns that limiting sample recording time might restrict expression may outweigh those around time-efficiency. Overall, assurances as to data security, privacy, and use-control, augmentation rather than displacement of clinical contact time, and maintenance of alternative monitoring

options have been identified as key factors governing the acceptability of new assessment technologies in mental health services[33,46].

In conclusion, we have argued that Bayesian network models offer a principled, explainable, and intervenable method for multi-source information integration which can be used to support clinical decision-making during assessment for common mental health problems. Our implementation demonstrates that with sufficiently large sample sizes, these models can achieve robust performance whilst maintaining transparent probabilistic reasoning that allows clinicians to both inspect dependencies between symptoms and conditions and incorporate their own insights into individual case structure. Appropriate scale also enables comprehensive fairness testing for different demographic groups—which in combination with other model properties may help address several key previously-identified barriers to implementing digital phenotyping methods in the clinic[13]. Bridging the gap between traditional psychiatric assessment and data-driven approaches while preserving clinical autonomy represents a viable path towards precision psychiatry that enhances rather than replaces expert clinical judgment[47].

Methods

Ethical approval and informed consent

The ethical review process for this study was led by an independent research ethics expert (Dr. David Carpenter), on behalf of the Association of Research Managers and Administrators (ARMA, https://arma.ac.uk/), with a favourable opinion granted on September 2, 2024. All participants gave written informed consent and were compensated for their time. All study procedures were performed in accordance with the Declaration of Helsinki.

Data collection

Participants were recruited using an online research participation platform (Prolific) and required to be 18 years or older, speak English as a first language, and resident in the US or UK. During recruitment, quotas were applied to ensure sufficient sampling of individuals of experience with significant mental health challenges (previous condition diagnosis), and at least nationally representative levels of race/ethnicity diversity.

Speech activity and self-report data were collected using the thymia research platform[48]. Participants completed two types of speech activity, both around 1 minute in length:

1. **Reading out loud task**. Participant read out loud a standard text commonly used as a speech elicitation task due to its phonetic range (the Aesop fable "The North Wind and the Sun"[49]).
2. **Answering a question task**. Participants were asked to describe what their mood had been like over the past two weeks, speaking at their usual volume and pace of speech.

Self-report measures included validated questionnaire measures of depression and anxiety symptoms (the PHQ-8[50] and GAD-7[51]), mental health history (including past and current psychiatric diagnoses), and demographic information.

Dataset construction

For corpus construction, speech activity recordings were passed to an in-house pre-processing pipeline that involved resampling to a 16kHz rate and transcript generation using automated speech-to-text tools (Deepgram). For acoustic feature generation, silences were trimmed to produce audio segments of durations in the range $10 - 20s$. Segments without sufficient speech to reach this threshold, or where recording quality was not sufficient to support automated transcript generation, were excluded from analysis. To balance data quality against model generalizability to real-world settings, minimal data cleaning was applied. Specifically, we further excluded samples from participants with very short completion times for PHQ-8 or GAD-7 measures ($< 1.5s$ per item, a simple indicator of potential inattentive responding[52]) and/or mean speech-to-text transcription confidence of < 0.80 ($\sim 10\%$ of processed speech data).

Following these steps, a development dataset was constructed, consisting of 39,571 total speech activities from N = 21,379 users (see Table S1). Two further datasets, a calibration set (11,334 observed speech activities from N = 6,325 users) and held-out test set (4,866 observations from N=2,431 users) were drawn from data collected after the development set was constructed and model structure was determined. No users overlapped between datasets, and the held-out test set was completely unseen during model development.

Analysis

Surrogate models

Since it would be computationally intractable to train a Bayesian network with the high-dimensional feature arrays used to represent speech acoustic and linguistic features, surrogate models were used to compress features into efficient representations of the signal contained in each feature type for each network symptom. Maintaining a level of unimodality at the surrogate representation level allowed delegation of integration of these different signals for each symptom to the Bayesian Network (see below).

To maximise signal strength, surrogate models were trained on binary classification targets derived from item-level PHQ-8 and GAD-7 response data. Individual symptom severity scores were binarized to categories representing significant symptom presence (experienced on half or more days over the past 2 weeks) *vs* mild or absent levels (experienced on less than half of days)—in line with previously-established algorithms for converting PHQ data to DSM-like symptom presence criteria[29]. For each symptom, three types of surrogate models were trained and evaluated (45 models total):

1. `reading-MM` models. Symptom predictions from combined acoustic and paralinguistic natural language processing (NLP) features from reading task data. Acoustic features were extracted from trimmed audio segments using an in-house speech model based on `TRILLsson5`[53]. Paralinguistic NLP features included speech rate, time to first utterance, and pause durations.
2. `mood-audio` models. Symptom predictions from the same acoustic features extracted from trimmed mood question activity audio segments.
3. `mood-linguistic` models. Symptom predictions from a linguistic feature set extracted from full mood question activity transcripts. Linguistic features were semantic embeddings from `ModernBERT-base`[54] plus an extended set of NLP features describing parts of speech such as first person pronoun use, generated using `spaCy`[55].

Model architecture. Surrogate models were feedforward multilayer neural networks (Fig. S1).

- For `reading-MM` models, acoustic embeddings and paralinguistic NLP (speech timing) features were initially processed separately. Acoustic embeddings were projected into a common projection dimension space using a two-layer network with batch normalization, dropout, and ReLU activation functions between layers ($1024 \rightarrow 25 \rightarrow 25$; where 25 is a common projection dimension across models based on the length of the full NLP feature set). The smaller speech timing feature set was projected using a simpler single-layer network that preserved dimensionality, with batch normalization and ReLU. The two projected representations were then concatenated in an additive fusion step.
- `mood-audio` models were two-layer networks with batch normalization, ReLU activation, and dropout between layers ($1024 \rightarrow 512 \rightarrow 256$).
- For `mood-linguistic` models, linguistic embeddings and traditional NLP features were initially processed separately. Linguistic embeddings were projected using a two-layer network with batch normalization, ReLU, and dropout ($768 \rightarrow 25 \rightarrow 25$). NLP features were projected using a single layer network that preserved dimensionality, with batch normalization and ReLU. The two projected representations were then concatenated in an additive fusion step.

For all surrogate model types, projections were passed to a prediction head that was a 3-layer network with a single output dimension, with batch normalization, ReLU, and dropout between each layer (combined projection dimension $\rightarrow$ 128 $\rightarrow$ 32 $\rightarrow$ 1). Final outputs were sigmoid-activated probabilities clamped to [0.000001, 0.999999].

Hyperparameter tuning and training. Surrogate models were trained using a nested cross-validation procedure in the development dataset, which was partitioned into training-validation and test splits in a 80:20 ratio.

Model hyperparameters were tuned via stratified K-fold cross-validation in the training-validation split using Bayesian optimization via `Optuna`[56] (k = 4, with 50 Optuna trials per fold). Specifically, mean ROC-AUC across inner folds was optimized via tuning of learning rate, batch size, dropout rate, and weight decay.

Final models were then trained using optimal hyperparameter configurations in all development training-validation split data. Training used an AdamW optimizer with weight decay for regularization, binary cross-entropy loss, learning rate scheduling via ReduceLROnPlateau, early stopping based on validation loss and gradient clipping to mitigate against exploding gradients. Models were selected based on best ROC-AUC in validation data, and performance was then evaluated in the development test split.

Bayesian network modelling

The Bayesian network described here consists of nodes representing overall condition status for depression and anxiety, and nodes representing individual depression and anxiety symptom severity states. The network takes the form of a directed acyclic graph, in which condition probabilities influence symptom severity states, and symptom states may influence both other symptoms levels and the values of observable nodes (here: different speech-feature derived surrogate model predictions for each symptom). During training, we parameterize the network by providing it with sets of observed values for all network nodes (conditions, symptoms, and surrogate predictions), from which it learns conditional probability distributions for relationships between different nodes states (for example, the distribution across restlessness symptom severity categories, given different levels of psychomotor issues and trouble relaxing). At inference, we can query the network with evidence derived from speech activity data (surrogate model outputs), to generate predictions for both symptom and overall condition states.

Network structure. Given purely data-driven approaches to specifying network structure are known to be fragile (particularly for networks with larger numbers of nodes), we used a hybrid literature and data-informed approach to structure specification.

Specifically, for specifying the inter-symptom network component, we leaned upon previously published large-sample joint network modelling analyses of depression and anxiety symptoms (predominantly measured using PHQ and GAD scales;[57–60]). From these studies, the most commonly identified and strongest intra-condition symptom connections (representing undirected partial correlations across participants at the same measurement occasion) for depression were between:

- Low mood and anhedonia
- Sleep isues and low energy
- Low mood and worthlessness issues
- Psychomotor issues and concentration issues

- Anhedonia and low energy
- Low energy and appetite issues
- Low mood and low energy

And for anxiety were between:

- Excessive worry and uncontrollable worry
- Trouble relaxing and restlessness
- Uncontrollable worry and nervousness
- Nervousness and trouble relaxing
- Nervousness and dread

The most commonly identified between-condition symptom associations ("bridge symptoms") were between psychomotor symptoms and restlessness, and concentration issues and trouble relaxing.

Inferring within-participant directed effects between symptoms requires intensive longitudinal data. Where possible, inter-symptom edge directions were informed by results temporal network analyses, which give insight into Granger-causal relationships between symptoms (i.e., whether increases or decreases in symptom A tends to precede increases or decreases in symptom B, between successive study time-points). Specifically, increases in self-reported low energy have been seen to precede increases in anhedonic symptoms[61,62], and worthlessness issues have been observed to precede increases in low mood[61].

Key differences between these analyses and our setting are that (1) in order to perform inference on conditions as well as symptoms, the network contains direct mappings from constituent symptoms to condition nodes and (2) that symptom estimates are informed by observable inputs (surrogate model predictions) that, although separately trained for each symptom, together hold information about the correlation structure across symptoms (for example, participants with higher excessive worry predictions will also tend to have higher uncontrollable worry predictions, since these are correlated at the ground truth level). During development testing it was found that indirect association via common association with parent depression or anxiety nodes and relationships between surrogate model predictions was sufficient to preserve many of the literature-identified inter-symptom associations in model predictions, without the need for specifying direct inter-symptom edges in the network. The retained set of sufficient inter-symptom network edges were:

- Low energy $\rightarrow$ anhedonia
- Worthlessness issues $\rightarrow$ low mood
- Appetite issues $\rightarrow$ low energy
- Trouble relaxing $\rightarrow$ restlessness
- Psychomotor issues $\rightarrow$ restlessness
- Trouble relaxing $\rightarrow$ concentration issues

To define surrogate prediction to symptom node edges, we started with a fully connected network (with all surrogates informing all relevant symptom estimates), then pruned connections according to relative informativeness (according to both overall performance strength and cross-symptom specificity)—with the constraint that each symptom must be informed by at least one two surrogate observations, with one being an acoustic features model (i.e., no symptoms were informed only by linguistic features). This process was intended to encourage the network to make use of symptom-specific information, rather than leaning on shared variance across symptoms. Effects of pruning explicit inter-symptom and symptom-observable edges on model performance and symptom prediction covariance were evaluated in the development set test split, with optimal network configuration chosen to optimize symptom and condition level performance (ROC-AUC) whilst retaining cross-symptom prediction covariance similar to the ground truth correlation structure.

For a diagram of the final network structure, including all retained symptom-surrogate edge connections, see Fig. S2.

In order to support efficient estimation and exact inference methods, the model was specified as a Discrete Bayesian Network, using `pgmpy`[63].

Data processing and target definitions. Ground truth values for symptom severity levels already existed in the form of discrete categories (PHQ-8 and GAD-7 item scores). Surrogate model predictions were continuous probabilities for symptom presence and were therefore discretized prior to entering the network using quartile transforms (with cut boundaries defined from observed distribution of each model's predictions in the development set).

For condition status, we used a compound target definition based on established cut-off scores for PHQ-8 and GAD-7 measures[50,51] in conjunction with self-reported mental health condition diagnosis information. Specifically, we defined depression/anxiety as being present when current PHQ/GAD total scores were ≥ 10 and a participant reported a previous mental health condition diagnosis, and absent when PHQ/GAD total scores were < 10 and a participant reported no mental health condition diagnosis. This definition was chosen on the basis of previous experiments in our datasets that have shown improved performance for simple acoustic speech models when training on this definition compared to simple binary split approaches: whether testing on either these or binary split target definitions. This implies that there is an advantage at training to using target definitions which (indirectly) incorporate external clinical judgement. We verified that use of this definition does not artificially decrease problem difficulty relative to binary split approaches by confirming that target categories retain good coverage across possible PHQ-8 and GAD-7 severity ranges (see Fig. S3). Additionally, we assessed performance robustness to alternative condition definitions which did not require presence or absence of a self-reported diagnosis (see Results).

Parameter estimation. Network parameters (posterior conditional probability distributions between state values) were estimated from development set training-validation split observations of ground truth and surrogate model prediction values using Bayesian estimation methods.

Specifically, we used Bayesian Dirichlet equivalent uniform priors, which act to regularize the posterior conditional probability distributions, and allow for some posterior weight to remain on combinations of state values which are rare or unseen in training data (guarding against over-fitting). This form of prior is specified as pseudo counts which are equivalent to having observed N uniform samples of each network variable state and parent configuration before the estimation data is observed (with observed data added to the pseudo count priors before normalization to produce posterior values). In `pgmpy`, equivalent uniform priors are specified by providing an equivalent sample size parameter, with the prior pseudo counts added to each node state combination calculated as $equivalentsamplesize/(nodecardinality * product(parentcardinalities))$.

During network development, we explored the effects of different degrees of prior regularization strength (equivalent sample size values) on ROC-AUC performance and correlation between output predictions for overall condition probabilities. This was motivated by the fact that depression and anxiety status were strongly correlated across individuals in our network training data, but a desired model property for generalization purposes is to ensure condition-specific sensitivity in predictions (allowing for cases where anxiety is present but not depression, and vice versa). Tested equivalent sample size parameter values effectively explored adding an additional 5-10% prior counts to single condition presence configurations, with final models using $equivalentsamplesize = 8000$ (for our network specification and estimation dataset size this is equivalent to adding 250 or 7.5% additional observations to each possible depression/anxiety node state configuration).

Inference procedure. Inference was performed by querying the parameterized model with unseen test data surrogate model predictions, using Variable Elimination (an exact inference method[64]).

Calibration

Given the importance of accurate probability estimates to clinical usefulness, we used an additional data set (matched to the development set in terms of stratification for age, birth sex, race/ethnicity, and mental health diagnosis status) to train calibrator models for refining overall depression and anxiety predictions. Specifically, we trained a collection ($N = 10$) of bagged `sklearn IsotonicRegression` estimators. Initially, 5-fold cross-validation was performed within the calibration set, in order to assess the stability of both standard calibration metrics (ECE, MCE, and Brier score) and discriminative performance (ROC-AUC) in calibrated predictions. With stable performance achieved, a final collection of calibrator models was trained using the entire calibration set.

Evaluation

Final performance estimates are based on results obtained in the held-out test set (data unseen during model development), with results from development testing provided for reference. In order to provide a fair assessment of the network's multimodal capabilities, only samples from users with surrogate model predictions from all three model types were included. For conditions, performance is reported for all users where ground truth depression/anxiety status was defined (for each condition separately). For symptoms, performance is reported for all test data (for the exact N included in each analysis see Results).

Following recent reporting recommendations for predictive AI models in medicine[27], model performance for conditions is described using both discrimination (ROC-AUC) and calibration (ECE) metrics (using M = 10 bins), with full calibration plots provided. These specific metrics are recommended based on assessment of statistical properness, and whether or not they incorporate factors related to misclassification cost that are more appropriately dealt with by formal decision analysis and clinical utility calculations. For example, whilst it is sometimes argued that F1 and related metrics should be provided in cases of class imbalance, the guidelines state that these are inappropriate for clinical predictive models since these both ignore true negative rates (which are important for medical applications) and conflate classification performance with clinical utility. This is because the extent to which an outcome is imbalanced (an epidemiological feature of the data) is not mathematically proportional to the extent to which misclassification costs are imbalanced (a clinical characteristic related to the specific medical decisions supported by the model).

For fairness assessment, we report within-group discriminative performance, and group differences in ECE, Brier scores (the mean squared distance between observed and predicted outcomes, or degree of alignment between predicted and observed outcome probabilities), and equalized odds ratios (as implemented in `fairlearn`[65]), which assesses whether different group members experience differential outcome allocation in terms of true and false positive rates.

We additionally report prevalence-dependent performance metrics for our test population, calculated using an example threshold of 0.5 for calibrated condition probabilities. Positive predictive values represent the proportion of positive predictions that are true positives, and negative predictive values the proportion of negative predictions that are true negatives. Positive and negative likelihood ratios describe the change in odds of having a condition after receiving a positive and negative test result, respectively, for a given population.

Service user consultation

For the service user consultation survey, we recruited a sample of N = 230 individuals from Prolific, with the following eligibility criteria: (1) UK residence; (2) aged 18 years or older; (3) self-reported current diagnosis of a mental health condition; (4) attempted to access NHS mental health support within the past 12 months; and (5) currently receiving or awaiting treatment (psychological therapy, medication, brain-stimulation treatments, hospital/residential programs, or self-management). Quota sampling ensured equal representation by birth sex and age ($\leq$40 vs > 40 years).

Participants provided their views about both standardized questionnaire measures and a new hypothetical voice-based mental health screening tool using likert ratings and open-ended free text responses. The voice-based mental health screening tool was described as a computer program to which short speech samples could be provided, that can take into account both the content of their words and other speech qualities (like tone, pace, or pauses) to provide estimates of their current mental health status to their healthcare provider ahead of an appointment.

Rating scale responses were summarised using medians and interquartile ranges. Free-text responses were analysed using thematic analysis adapted from Braun and Clarke[66] with additional systematic coding procedures[67]. The first author (AN) familiarised herself with responses and inductively developed a coding framework capturing distinct patterns in participant feedback. Each response was coded for presence/absence of themes, with multiple themes possible per response. Recognising the applied nature of this research and need for reproducibility, we incorporated inter-coder reliability testing with a second independent coder (AG). Cohen's Kappa indicated substantial agreement across all categories (mean $\kappa = 0.83$). Theme prevalence was then determined through a frequency count to identify the most salient concerns among participants.

Data availability

The datasets generated during and analysed during the current study are not publicly available due to lack of consent for public sharing of raw data, which due to the nature of speech data would compromise participant privacy. They may be made available on reasonable request to the corresponding author for non-commercial research purposes related to detecting or monitoring mental health and wellbeing (the purposes for which consent for research data sharing was obtained), subject to completion of a Data Sharing Agreement with thymia Ltd.

Code availability

Supporting analysis code is available to reviewers and editors on request during submission.

Received: 23 September 2025; Accepted: 17 December 2025
Published online: 06 February 2026

References

1. Kind, A. *How Does the Psychiatrist Know?: On the Epistemology of Psychiatric Diagnostic Reasoning* (Transcript, 2025).
2. Huda, A. S. The medical model and its application in mental health. *Int. Rev. Psychiatry* **33**, 463–470. https://doi.org/10.1080/09540261.2020.1845125 (2021).
3. Aftab, A., Banicki, K., Ruffalo, M. L. & Frances, A. Psychiatric diagnosis: a clinical guide to navigating diagnostic pluralism. *J. Nervous Mental Dis.* **212**, 445–454. https://doi.org/10.1097/NMD.0000000000001791 (2024).
4. Bhugra, D., Easter, A., Mallaris, Y. & Gupta, S. Clinical decision making in psychiatry by psychiatrists. *Acta Psychiatr. Scand.* **124**, 403–411. https://doi.org/10.1111/j.1600-0447.2011.01737.x (2011).
5. Casey, P. & Kelly, B. *Fish's Clinical Psychopathology* (Cambridge University Press, 2019).
6. Trzepacz, P. T. & Baker, R. W. *The Psychiatric Mental Status Examination* (Oxford University Press, 1993).
7. Royal College of Psychiatrists. Workforce strategy 2020-2023 (2020). https://www.rcpsych.ac.uk/improving-care/workforce/workforce-strategy.
8. Giotakos, O. Psychiatry in the real world. *Front. Psychiatry* **16**, 1452. https://doi.org/10.3389/fpsyt.2025.1616276 (2025).
9. Mouchabac, S. et al. Improving clinical decision-making in psychiatry: implementation of digital phenotyping could mitigate the influence of patient's and practitioner's individual cognitive biases. *Dialogues Clin. Neurosci.* **23**, 52–61. https://doi.org/10.1080/19585969.2022.2042165 (2021).
10. Bansal, N. et al. Understanding ethnic inequalities in mental healthcare in the UK: a meta-ethnography. *PLoS Med.* **19**, e1004139. https://doi.org/10.1371/journal.pmed.1004139 (2022).
11. Clery, E. et al. Mental health treatment and service use. In *Adult Psychiatric Morbidity Survey: Survey of Mental Health and Wellbeing, England, 2023/4* (NHS England, 2025). https://digital.nhs.uk/data-and-information/publications/statistical/adult-psychiatric-morbidity-survey/survey-of-mental-health-and-wellbeing-england-2023-24/mental-health-treatment-and-service-use.
12. Chia, A. Z. & Zhang, M. W. Digital phenotyping in psychiatry: a scoping review. *Technol. Health Care* **30**, 1331–1342. https://doi.org/10.3233/THC-213648 (2022).
13. Huckvale, K., Venkatesh, S. & Christensen, H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *Npj Digital Med.* **2**, 88. https://doi.org/10.1038/s41746-019-0166-1 (2019).
14. Martin, V. P. & Rouas, J.-L. Estimating symptoms and clinical signs instead of disorders: the path toward the clinical use of voice and speech biomarkers in psychiatry. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 10606–10610 (2024). https://doi.org/10.1109/ICASSP48485.2024.10445888.
15. Rutowski, T. et al. Toward corpus size requirements for training and evaluating depression risk models using spoken language. *arXiv* https://doi.org/10.48550/arXiv.2501.00617 (2024).
16. Kyrimi, E. et al. A comprehensive scoping review of bayesian networks in healthcare: past, present and future. *Artif. Intell. Med.* **117**, 102108. https://doi.org/10.1016/j.artmed.2021.102108 (2021).
17. Polotskaya, K. et al. Bayesian networks for the diagnosis and prognosis of diseases: a scoping review. *Mach. Learn. Knowl. Extract.* **6**, 1243–1262. https://doi.org/10.3390/make6020058 (2024).
18. McGrath, J. J. et al. Comorbidity within mental disorders: a comprehensive analysis based on 145 990 survey respondents from 27 countries. *Epidemiol. Psychiatr. Sci.* **29**, e153. https://doi.org/10.1017/S2045796020000633 (2020).
19. Forbes, M. K. et al. Elemental psychopathology: distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. *Psychol. Med.* **54**, 886–894. https://doi.org/10.1017/S0033291723002544 (2024).
20. Waszczuk, M. A. et al. What do clinicians treat: Diagnoses or symptoms? The incremental validity of a symptom-based, dimensional characterization of emotional disorders in predicting medication prescription patterns. *Compr. Psychiatry* **79**, 80–88. https://doi.org/10.1016/j.comppsych.2017.04.004 (2017).
21. Fried, E. I. & Nesse, R. M. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med.* **13**, 72. https://doi.org/10.1186/s12916-015-0325-4 (2015).
22. Fried, E. Moving forward: how depression heterogeneity hinders progress in treatment and research. *Expert Rev. Neurother.* **17**, 423–425. https://doi.org/10.1080/14737175.2017.1307737 (2017).

23. Cross, S., Mellor-Clark, J. & Macdonald, J. Tracking responses to items in measures as a means of increasing therapeutic engagement in clients: a complementary clinical approach to tracking outcomes. *Clin. Psychol. Psychother.* **22**, 698–707. https://doi.org/10.1002/cpp.1929 (2015).
24. D'Alfonso, S., Coghlan, S., Schmidt, S. & Mangelsdorf, S. Ethical dimensions of digital phenotyping within the context of mental healthcare. *J. Technol. Behav. Sci.* **10**, 132–147. https://doi.org/10.1007/s41347-024-00423-9 (2025).
25. Berisha, V., Krantsevich, C., Stegmann, G., Hahn, S. & Liss, J. Are reported accuracies in the clinical speech machine learning literature overoptimistic? In *Interspeech 2022* 1440–1444 (ISCA, 2022). https://doi.org/10.21437/Interspeech.2022-691.
26. Hond, A. A. H. d., Steyerberg, E. W. & Calster, B. v. Interpreting area under the receiver operating characteristic curve. *Lancet Digital Health***4**, e853–e855. https://doi.org/10.1016/S2589-7500(22)00188-1 (2022).
27. Calster, B. V. et al. Performance evaluation of predictive AI models to support medical decisions: overview and guidance. *arXiv* https://doi.org/10.48550/arXiv.2412.10288 (2024).
28. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On Calibration of modern neural networks. *arXiv* https://doi.org/10.48550/arXiv.1706.04599 (2017).
29. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9. *J. Gen. Intern. Med.* **16**, 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x (2001).
30. Hemphill, J. F. Interpreting the magnitudes of correlation coefficients. *Am. Psychol.* **58**, 78–79. https://doi.org/10.1037/0003-066X.58.1.78 (2003).
31. van Krugten, F. C. W., Busschbach, J. J. V., Versteegh, M. M., Hakkaart-van Roijen, L. & Brouwer, W. B. F. The mental health quality of life questionnaire (MHQoL): development and first psychometric evaluation of a new measure to assess quality of life in people with mental health problems. *Qua. Life Res.* **31**, 633–643. https://doi.org/10.1007/s11136-021-02935-w (2022).
32. Moriarty, D. G., Zack, M. M. & Kobau, R. The centers for disease control and prevention's healthy days measures—population tracking of perceived physical and mental health over time. *Health Qual. Life Outcomes* **1**, 37. https://doi.org/10.1186/1477-7525-1-37 (2023).
33. NICE. Digital front door technologies to gather service user information for NHS talking therapies for anxiety and depression assessments: early value assessment (2025). https://www.nice.org.uk/guidance/hte30.
34. Goria, S., Polle, R., Fara, S. & Cummins, N. Revealing confounding biases: a novel benchmarking approach for aggregate-level performance metrics in health assessments. In *Interspeech 2024* 1440–1444 (ISCA, 2024). https://doi.org/10.21437/Interspeech.2024-1092.
35. Watkins, E. A., McKenna, M. & Chen, J. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. *arXiv* https://doi.org/10.48550/arXiv.2202.09519 (2022).
36. Jacobson, N. C. & Newman, M. G. Anxiety and depression as bidirectional risk factors for one another: a meta-analysis of longitudinal studies. *Psychol. Bull.* **143**, 1155–1200. https://doi.org/10.1037/bul0000111 (2017).
37. Mind. GP mental health training survey summary (2018). https://www.mind.org.uk/media-a/4414/gp-mh-2018-survey-summary.pdf.
38. Malpass, A. et al. Usefulness of PHQ-9 in primary care to determine meaningful symptoms of low mood: a qualitative study. *Br. J. Gen. Pract.* **66**, e78–e84. https://doi.org/10.3399/bjgp16X683473 (2016).
39. Bendall, C. & McGrath, L. Contending with the minimum data set: subjectivity, linearity and dividualising experiences in improving access to psychological therapies. *Health* **24**, 94–109. https://doi.org/10.1177/1363459318785718 (2020).
40. Ford, J., Thomas, F., Byng, R. & McCabe, R. Use of the patient health questionnaire (PHQ-9) in practice: interactions between patients and physicians. *Qual. Health Res.* **30**, 2146–2159. https://doi.org/10.1177/1049732320924625 (2020).
41. Faija, C. L. et al. Using routine outcome measures as clinical process tools: maximising the therapeutic yield in the IAPT programme when working remotely. *Psychol. Psychother. Theory Res. Pract.* **95**, 820–837. https://doi.org/10.1111/papt.12400 (2022).
42. Fara, S. et al. Bayesian networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data. In *Interspeech 2023* 1728–1732 (ISCA, 2023). https://doi.org/10.21437/Interspeech.2023-1709.
43. de Hond, A. A. H. et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *Npj Digital Med.* **5**, 2. https://doi.org/10.1038/s41746-021-00549-7 (2022).
44. International Organization for Standardization. Information technology - artificial intelligence—management system. International Standard ISO/IEC 42001:2023, ISO/IEC, Geneva, Switzerland (2023).
45. Timmons, A. C. et al. Bridging fair-aware artificial intelligence and co-creation for equitable mental healthcare. *Nat. Rev. Psychol.* **2025**, 1–15. https://doi.org/10.1038/s44159-025-00491-5 (2025).
46. Martinez-Martin, N., Greely, H. T. & Cho, M. K. Ethical development of digital phenotyping tools for mental health applications: Delphi study. *JMIR Mhealth Uhealth* **9**, e27343. https://doi.org/10.2196/27343 (2021).
47. Kabrel, N., Stade, E., Aru, J. & Eichstaedt, J. Current AI should extend (not replace) human care in mental health. *OSF* https://doi.org/10.31234/osf.io/e3qg7_v1 (2025).
48. Fara, S., Goria, S., Molimpakis, E. & Cummins, N. Speech and the n-back task as a lens into depression. how combining both may allow us to isolate different core symptoms of depression. In *Interspeech 2022* 1911–1915 (ISCA, 2022). https://doi.org/10.21437/Interspeech.2022-10393.
49. Deterding, D. The north wind versus a wolf: short texts for the description and measurement of english pronunciation. *J. Int. Phon. Assoc.* **36**, 187–196. https://doi.org/10.1017/S0025100306002544 (2006).
50. Kroenke, K. et al. The PHQ-8 as a measure of current depression in the general population. *J. Affect. Disord.* **114**, 163–173. https://doi.org/10.1016/j.jad.2008.06.026 (2009).
51. Spitzer, R. L., Kroenke, K., Williams, J. B. W. & Löwe, B. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch. Intern. Med.* **166**, 1092–1097. https://doi.org/10.1001/archinte.166.10.1092 (2006).
52. Ulitzsch, E., Shin, H. J. & Lüdtke, O. Accounting for careless and insufficient effort responding in large-scale survey data-development, evaluation, and application of a screen-time-based weighting procedure. *Behav. Res. Methods* **56**, 804–825. https://doi.org/10.3758/s13428-022-02053-6 (2024).
53. Shor, J. & Venugopalan, S. TRILLsson: distilled universal paralinguistic speech representations. In *Interspeech 2022* 356–360 (ISCA, 2022). https://doi.org/10.21437/Interspeech.2022-118.
54. Warner, B. et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv* https://doi.org/10.48550/arXiv.2412.13663 (2024).
55. Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. spaCy: industrial-strength natural language processing in Python (2020). https://doi.org/10.5281/zenodo.1212303.
56. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (2019). https://doi.org/10.1145/3292500.3330701.
57. Beard, C. et al. Network analysis of depression and anxiety symptom relations in a psychiatric sample. *Psychol. Med.* **46**, 3359–3369. https://doi.org/10.1017/S0033291716002300 (2016).
58. Kaiser, T., Herzog, P., Voderholzer, U. & Brakemeier, E.-L. Unraveling the comorbidity of depression and anxiety in a large inpatient sample: network analysis to examine bridge symptoms. *Depress. Anxiety* **38**, 307–317. https://doi.org/10.1002/da.23136 (2021).
59. Hoffart, A., Johnson, S. U. & Ebrahimi, O. V. The network of stress-related states and depression and anxiety symptoms during the COVID-19 lockdown. *J. Affect. Disord.* **294**, 671–678. https://doi.org/10.1016/j.jad.2021.07.019 (2021).

60. Cai, H. et al. A network model of depressive and anxiety symptoms: a statistical evaluation. *Mol. Psychiatry* **29**, 767–781. https://doi.org/10.1038/s41380-023-02369-5 (2024).
61. Ebrahimi, O. V., Burger, J., Hoffart, A. & Johnson, S. U. Within- and across-day patterns of interplay between depressive symptoms and related psychopathological processes: a dynamic network approach during the COVID-19 pandemic. *BMC Med.* **19**, 317. https://doi.org/10.1186/s12916-021-02179-y (2021).
62. Lunansky, G., Hoekstra, R. H. A. & Blanken, T. F. Disentangling the role of affect in the evolution of depressive complaints using complex dynamical networks. *Collabra: Psychology***9**, 74841. https://doi.org/10.1525/collabra.74841 (2023).
63. Ankan, A. & Textor, J. pgmpy: A python toolkit for bayesian networks (2024). http://jmlr.org/papers/v25/23-0487.html.
64. Zhang, N. L. & Poole, D. Exploiting causal independence in bayesian network inference. *J. Artif Intell. Res.* **5**, 301–328. https://doi.org/10.1613/jair.305 (1996).
65. Weerts, H. et al. Fairlearn: assessing and improving fairness of ai systems (2023). http://jmlr.org/papers/v24/23-0389.html.
66. Braun, V. & Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**, 77–101. https://doi.org/10.1191/1478088706qp063oa (2006).
67. Boyatzis, R. E. *Transforming Qualitative Information: Thematic Analysis and Code Development* (Sage Publications, 1998).

Author contributions

S.G. and E.M. conceived the experiment(s), A.N. and G.F. conducted the experiment(s), A.N., G.F. and A.L.G. analysed the results. All authors reviewed and contributed to the writing of the manuscript.

Funding

This research did not receive any external funding. A.N., G.F., and A.L.G. conducted this research as part of their employment at thymia Limited, which provided internal resources and support for the work.

Declarations

Competing interests

E.M. and S.G. are co-founders of thymia Ltd. A.N., G.F., and A.L.G. are employees of thymia Ltd. E.M., S.G., A.N., A.L.G. hold equity in the company. Thymia Ltd. develops and commercializes digital mental health assessment technologies. As equity holders, these authors may financially benefit if positive research findings in this field enhance the perceived value or commercial success of similar technologies developed by thymia Ltd. M.M.N. has acted as a paid consultant for thymia Ltd, which does not extend to involvement in the current work.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-33331-w.

Correspondence and requests for materials should be addressed to S.G.