Supplementary Material

Supplementary Material for *A multimodal Bayesian Network for symptom-level depression and anxiety prediction from voice and speech data* by Agnes Norbury, George Fairs, Alexandra L. Georgescu, Matthew M. Nour, Emilia Molimpakis, and Stefano Goria (2025).

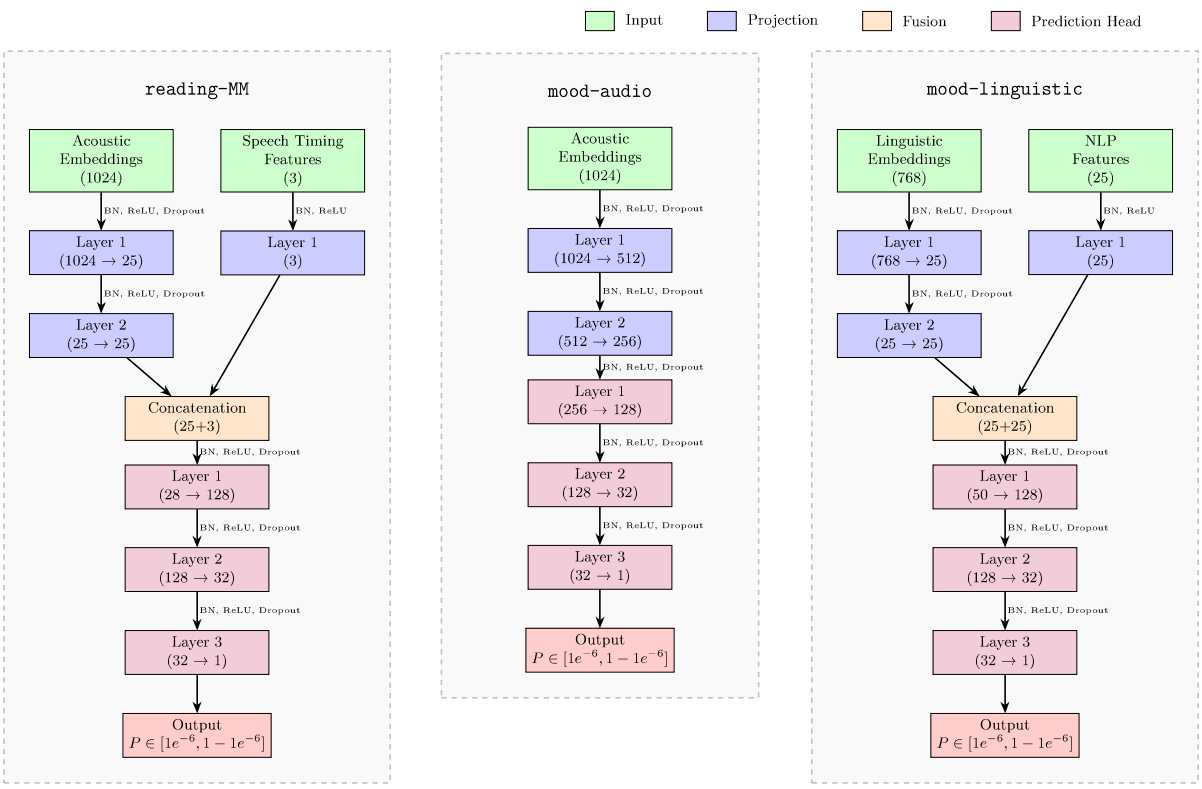

Figure S1. Surrogate model architecture. Architecture of the three surrogate model types. All surrogate models were feedforward neural networks. *BN*, batch normalization.

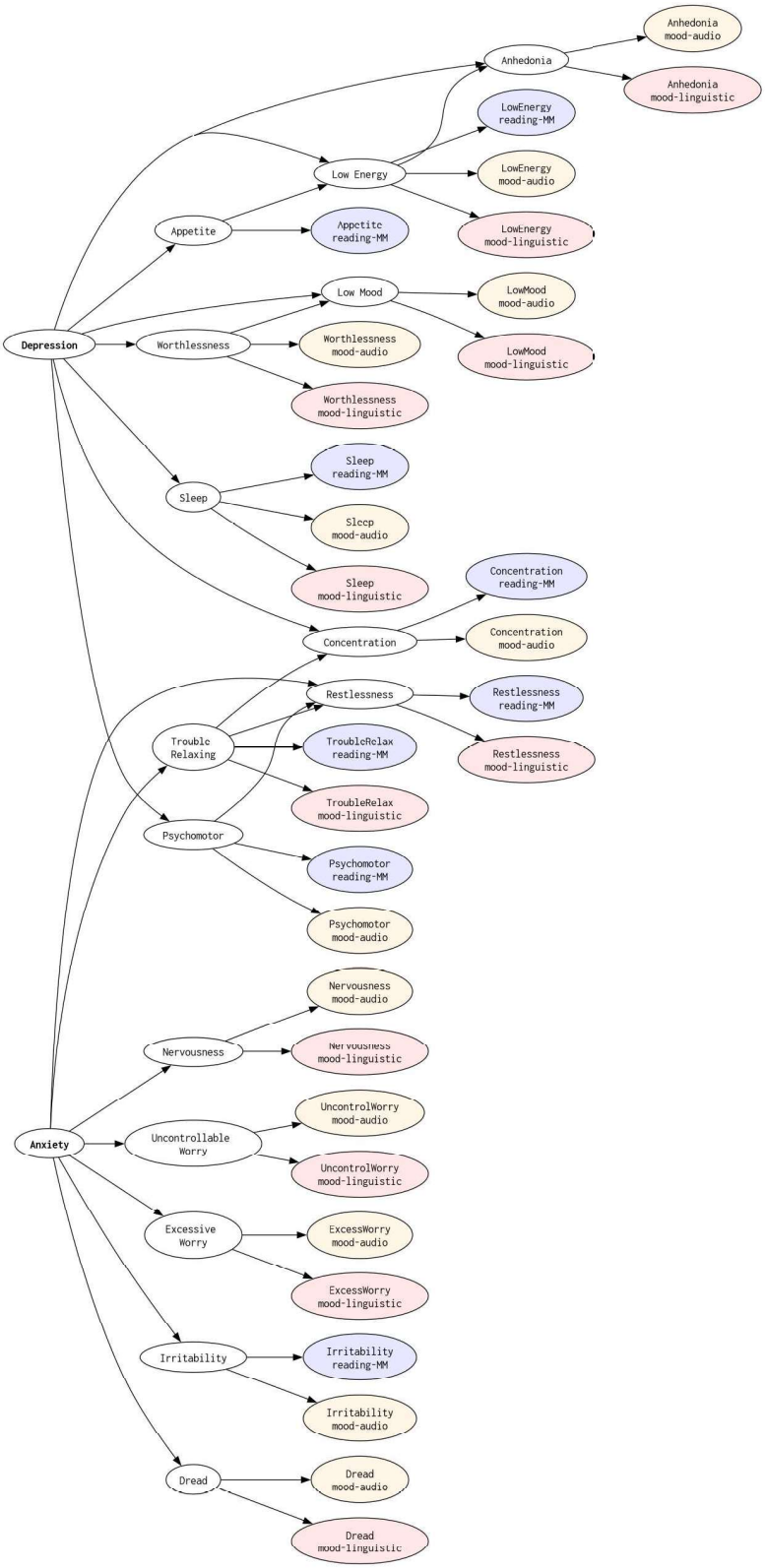

Figure S2. Bayesian Network structure. A joint network of both overall depression and anxiety status, and individual symptom severity states was constructed. Arrows represent the direction of causality implied by the specified direct edges (connections between nodes). Filled nodes represent observed evidence used at inference.

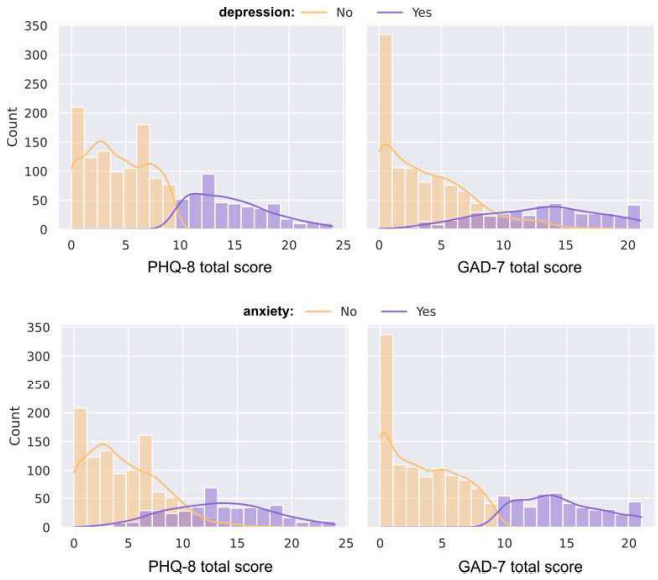

Figure S3. Distribution of PHQ-8 and GAD-7 total scores in test set data, by condition.

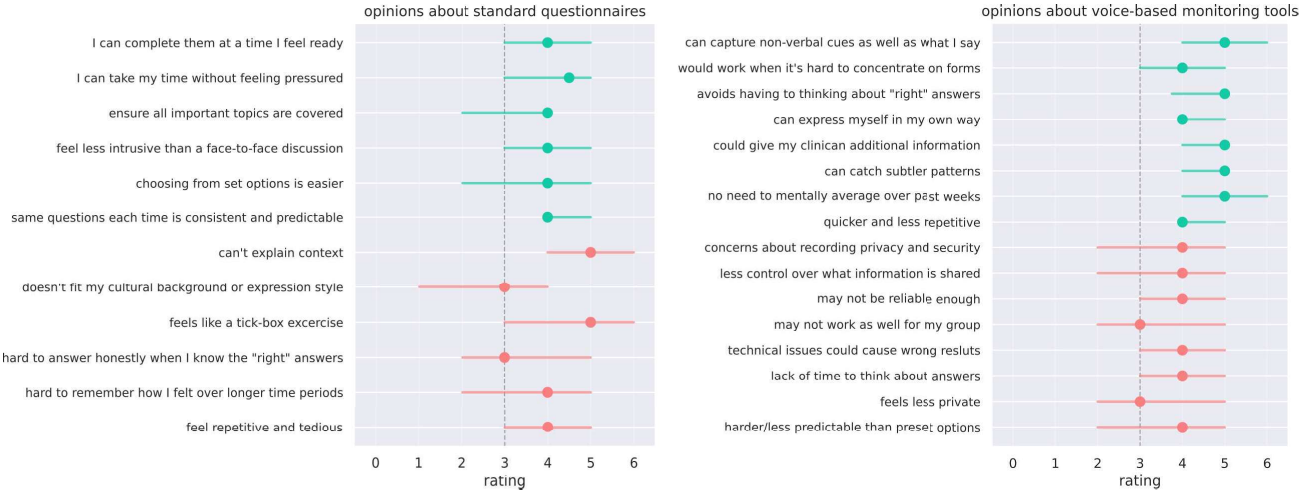

Figure S4. Mental health service user views of strengths and weaknesses of standard monitoring questionnaires and a hypothetical speech-based assessment tool. Summary of responses from *N*=230 participants with a mental health diagnosis and experience of screening and monitoring practices in UK mental health services. For each statement, survey participants were asked to indicated how much they agreed or disagreed on a 7-point likert scale (0-completely disagree, 6-completely agree). Plotted values represent interquartile ranges and medians across participant ratings.

		Development (*N*=21379)	Calibration (*N*=6325)	Test (*N*=2431)
Age	Mean (SD)	37.6 (12.9)	37.4 (13.2)	37.1 (13.0)
	Range	18-89	18-88	18-80
Birth Sex	Female	13031	4143	1637
	Male	8290	2015	766
	Prefer not to say	44	31	7
	Intersex	14	7	3
Gender	Woman	12780	4091	1591
	Man	8250	1988	759
	Non-binary	303	93	56
	Prefer not to say	46	24	7
Race/Ethnicity	White	15544	4508	1771
	Black	3289	774	302
	Mixed	1076	388	150
	Asian	1054	385	141
	Other	416	141	49
Diagnosed Mental Health Condition	No	12387	3466	1307
	Yes	8597	2615	1057
	Prefer not to say	259	115	49
Mental Health Diagnosis	Depressive Disorder	5718	1709	698
	Anxiety Disorder	6460	1981	844
	Other	1471	407	146
PHQ-8 Total Score	Mean (SD)	7.9 (5.7)	7.6 (5.4)	7.7 (5.4)
	Range	0-24	0-24	0-24
GAD-7 Total Score	Mean (SD)	7.3 (5.7)	7.1 (5.6)	7.2 (5.6)
	Range	0-21	0-21	0-21
Chronic Physical Health Condition	None	12301	3663	1391
	Respiratory	4933	745	343
	Cardiovascular	3419	956	373
	Arthritis	1516	367	155
	Anaemia	1257	357	153
	Sleep Apnoea	1215	332	139
	Diabetes	1107	299	114
	Fibromyalgia	637	149	71
	Long COVID	625	137	55
	Chronic Fatigue	605	148	66
	Other	2445	659	290
Accent	US/Other	11565	3668	1444
	UK	9578	2488	957
Device Type	Laptop	12857	3567	1422
	Smartphone or Tablet	8272	2605	984

Table S1. Participant numbers and demographics by study dataset. Values represent *N* unless otherwise specified.

Condition	Symptom	Model	ROC-AUC
Depression	Anhedonia	mood-audio	0.674
		mood-linguistic	0.715
	Low Mood	mood-audio	0.712
		mood-linguistic	0.779
	Sleep	reading-MM	0.620
		mood-audio	0.662
		mood-linguistic	0.684
	Low Energy	reading-MM	0.634
		mood-audio	0.692
		mood-linguistic	0.724
	Appetite	reading-MM	0.620
	Worthlessness	mood-audio	0.691
		mood-linguistic	0.746
	Concentration	reading-MM	0.601
		mood-audio	0.649
	Psychomotor	reading-MM	0.638
		mood-audio	0.680
Anxiety	Nervousness	mood-audio	0.709
		mood-linguistic	0.742
	Uncontrollable Worry	mood-audio	0.695
		mood-linguistic	0.733
	Excessive Worry	mood-audio	0.692
		mood-linguistic	0.735
	Trouble Relaxing	reading-MM	0.607
		mood-linguistic	0.714
	Restlessness	reading-MM	0.624
		mood-linguistic	0.652
	Irritability	reading-MM	0.623
		mood-audio	0.677
	Dread	mood-audio	0.654
		mood-linguistic	0.682

Table S2. Discrimination performance for individual surrogate models. Results are reported for the development set test split where surrogate model performance was evaluated, for all surrogates used as inputs to the final (pruned) Bayesian network model.

	reading-MM only	mood-audio only	mood-linguistic only	reading-MM & mood-audio
ROC-AUC Depression	0.717	0.772	0.783	0.783
ROC-AUC Anxiety	0.705	0.776	0.805	0.783
ECE Depression (Uncalibrated)	0.052	0.060	0.091	0.078
ECE Anxiety (Uncalibrated)	0.077	0.084	0.086	0.077
ROC-AUC Anhedonia	0.589*	0.681	0.725	0.680+
ROC-AUC Low Mood	0.619*	0.716	0.779	0.716+
ROC-AUC Sleep	0.627	0.669	0.693	0.674
ROC-AUC Low Energy	0.634	0.686	0.723	0.686
ROC-AUC Appetite	0.629	0.670*	0.689*	0.669+
ROC-AUC Worthlessness	0.627*	0.702	0.751	0.707+
ROC-AUC Concentration	0.620	0.660	0.719*	0.663
ROC-AUC Psychomotor	0.640	0.693	0.699*	0.691
ROC-AUC Nervousness	0.618*	0.708	0.746	0.706+
ROC-AUC Uncontrollable Worry	0.615*	0.702	0.738	0.700+
ROC-AUC Excessive Worry	0.621*	0.697	0.737	0.696+
ROC-AUC Trouble Relaxing	0.614	0.678*	0.715	0.665+
ROC-AUC Restlessness	0.625	0.658*	0.671	0.659+
ROC-AUC Irritability	0.631	0.689	0.725*	0.689
ROC-AUC Dread	0.599*	0.663	0.694	0.664+

Table S3. Performance for Bayesian Network when supplied only with different surrogate model prediction types.Results are reported for the development set test split where effects of Bayesian network architecture on performance were explored. Given in the final (pruned) network structure not all symptoms are informed by each surrogate type, some symptoms in each single-surrogate analysis are informed only by connections with other symptom estimates (*). For the combined paralinguistic surrogates results (reading-MM & mood-audio), + indicates symptoms only directly informed by one of the two paralinguistic surrogate types. Results are reported for the development set test split where effects of Bayesian Network structure on performance were evaluated.

Variable		Value
Age	Mean (SD)	41.0 (12.0)
	Range	19-69
Birth Sex	Female	115
	Male	115
Race/Ethnicity	White	206
	Mixed	8
	Black	7
	Asian	7
	Other	2
Mental Health Diagnosis	Anxiety	179
	Depression	172
	PTSD	31
	OCD	27
	Eating Disorder	16
	Bipolar Disorder	14
	Other	38
Treatment Experience	Medication	145
	Psychological therapy	101
	Own care	66
	Hospital and residential	6
	Other	3
Length of Experience	More than 10 years	163
	6-10 years	33
	3-5 years	22
	1-2 years	8
	Less than 1 year	3
	Prefer not to say	1
Questionnaire Familiarity	Very familiar	105
	Somewhat familiar	116
	Not very familiar	9
Comfort with Technology	Very comfortable	158
	Somewhat comfortable	58
	Moderately comfortable	11
	Not very comfortable	3

Table S4. Description of participants for the stakeholder consultation survey. All participants were based in the UK, had a current mental health diagnosis, and reported that they had tried to access mental health support via the NHS in the past 12 months (total N=230). Values represent N unless otherwise specified (diagnosis and treatment experience categories are non mutually-exclusive).

Theme	*N* Mentions
Strong privacy policy; effective cybersecurity; demonstrated compliance with relevant data protection legislation: including full information on who has access to the data and what purposes it will be used for	127
Evidence of tool accuracy and reliability (e.g., published studies and clinical trials); including evidence of fair performance for people from my specific demographic group	81
Recommendation by care provider (e.g., GP) or health system (e.g., NHS)	18
At this moment in time, nothing would make me trust or be comfortable with this kind of tool (AI scepticism)	15
Recommendations from other end-users (e.g., testimonials, reviews)	14
Option to test out and self-verify results (check if these align with my and/or my clinician's impressions)	10
Easy to use interface; well-tested, with technical support available if needed	7
Assurance that results will always be interpreted in conjunction with input from my clinician (and not replace human clinical interactions)	5
Ability to control which speech recordings are analysed or passed to clinicians; option to re-record responses if I want	5
Ethical implementation; safeguards against misuse; accountability for errors	4
Assurance that I will be able to see my results and challenge them if I don't agree	4

Table S5. Qualitative analysis of features required for trustworthy implementation of a speech-based mental health assessment tool according to a UK-based sample of mental health service users. Thematic analysis of feedback provided by *N*=230 respondents, alongside number of individual participants referencing each theme (for themes mentioned by 2 or more participants).