Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

One item per disorder: a cross-diagnostic optimization framework for simultaneous prediction of six psychiatric disorders

Myna Lim
Korea Advanced Institute of Science and Technology
Sooyeon Suh
Sungshin Women's University
Jinhyeon Byeon
Sungshin Women's University
Daeil Jang
National Institute for Mathematical Sciences
Seockhoon Chung
Asan Medical Center
Jae Kyoung Kim
jaekkim@kaist.ac.kr

Institute for Basic Science

Article

Keywords:

Posted Date: April 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9356755/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

One item per disorder: a cross-diagnostic optimization framework for simultaneous prediction of six psychiatric disorders

Myna Lim[1,2,†], Sooyeon Suh[3,4,†], Jinhyeon Byeon[3], Daeil Jang[5*], Seockhoon Chung[6,7*], Jae Kyoung Kim[1,2,8,9*]

[1] Graduate School of Data Science, KAIST, Daejeon 34141, Republic of Korea

[2] Biomedical Mathematics Group, Institute for Basic Science, Daejeon 34126, Republic of Korea

[3] Department of Psychology, Sungshin University, Seoul 02844, Republic of Korea

[4] Human-centered AI Institute, Sungshin University, Seoul 02844, Republic of Korea

[5] Innovation Center for Industrial Mathematics, National Institute for Mathematical Sciences, Seongnam 13449, Republic of Korea

[6] Department of Psychiatry, Asan Medical Center, University of Ulsan College of Medicine, Seoul 05505, Korea

[7] Life Care Center for Cancer Patient, Asan Medical Center Cancer Institute, Asan Medical Center, Seoul 05505, Korea

[8] Department of Mathematical Sciences, KAIST, Daejeon 34141, Republic of Korea

[9] Department of Medicine, College of Medicine, Korea University, Seoul 02841, Republic of Korea

[†]M.L. and S.S. contributed equally

[*]Correspondence:

Daeil Jang, PhD, Innovation Center for Industrial Mathematics, National Institute for Mathematical Sciences, Seongnam, 13449, Republic of Korea.

Tel +82-31-759-9419, E-mail: torusdi@nims.re.kr

Seockhoon Chung, MD, PhD, Department of Psychiatry, Asan Medical Center, University of Ulsan College of Medicine, 86 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.

Tel +82-2-3010-3411, E-mail: schung@amc.seoul.kr

Jaekyoung Kim, PhD, Department of Mathematical Sciences, KAIST, 291 Daehak-ro Yuseong-gu, Daejeon 34141 Republic of Korea.

Tel + 82-42-350-2702~4, E-mail: jaekkim@kaist.ac.kr

Abstract

Digital mental health systems increasingly rely on multimodal, AI-driven approaches, yet remain constrained by lengthy, disorder-specific questionnaires. This burden is amplified in patients with overlapping symptoms, creating a measurement bottleneck that limits scalable multi-disorder screening. To address this, we developed a cross-diagnostic item selection framework that identifies one representative item from each of six psychiatric instruments—PHQ-9, GAD-7, PCL-5, AUDIT, PDSS, and DSI-SS—via multi-objective optimization balancing predictive utility and redundancy. This reduces assessment length from 57 to 6 items (~90% reduction). In a general population sample (N = 1,600), a CatBoost multi-output model using these six items plus age and sex achieved high performance on an independent lockbox dataset (AUROC: 0.909–0.977; AUPRC: 0.829–0.903). Cross-diagnostic integration improved performance for depression, anxiety, PTSD, and alcohol use, while panic and suicidality remained anchor-dominant, consistent with SHAP and GGM analyses. This compact framework enables scalable, interpretable, multi-condition screening.

Introduction

Mental illnesses are a leading contributor to the global burden of disease[1], yet scalable and efficient mental health screening remains a persistent clinical challenge. In real-world settings, patients rarely present with symptoms confined to a single diagnostic category; instead, they frequently exhibit complex, overlapping profiles spanning depression, anxiety, trauma-related symptoms, and substance use[2,3]. Despite the high prevalence of comorbidities, time constraints and complaint-focused assessment practices often fail to capture co-occurring conditions. These limitations highlight the need for screening tools capable of efficiently evaluating multiple psychiatric domains simultaneously[4,5,6].

In response, the field of digital health has expanded mental health risk prediction across diverse input modalities. Video-based multimodal models integrate text, speech, and facial expressions to estimate the severity of depression and anxiety[7,8]. Furthermore, wearable-based studies have demonstrated the feasibility of predicting mood and depressive risk from physiological signals such as sleep, activity, and heart rate[9,10,11,12]. These approaches emphasize automation, passive monitoring, and scalability, demonstrating strong predictive performance across several psychiatric outcomes.

Despite rapid advances in digital mental health technologies, the measurement framework used to define clinical outcomes has remained largely unchanged. Most prediction models continue to rely on traditional questionnaire-derived scores as their primary targets, meaning that while input modalities have diversified, the underlying measurement layer remains anchored to conventional survey instruments[13,14]. For example, video-based models commonly predict PHQ-9 or GAD-7 scores[7,15], and wearable-based approaches similarly depend on questionnaire-derived labels[16,17]. Consequently, digital health models remain structurally dependent on repeated administration of lengthy questionnaires, leaving the survey-burden bottleneck largely unresolved [18,19]. This imbalance suggests that progress in digital mental health has been driven primarily by expansion of input modalities rather than redesign of the measurement structure itself.

These limitations become more pronounced in multi-disorder and longitudinal assessment settings. Comprehensive instruments such as Minnesota Multiphasic Personality Inventory-2 (MMPI-2)[20] contain more than 500 items and are impractical for repeated use in digital health environments. As additional psychiatric conditions are included, the burden of administering multiple disorder-specific instruments increases proportionally[21,22]. Moreover, overlapping symptom dimensions are often assessed redundantly across separate questionnaires[7,23,24,25,26], contributing to user fatigue and data attrition in longitudinal studies[27,28]. This burden also constrains the integration of digital biomarkers, as continuously collected behavioral signals must be aligned with temporally synchronized clinical labels that are difficult to obtain at scale. Taken together, these challenges indicate that scalable multi-disorder digital mental health screening requires innovation not only in input modalities but also in the measurement architecture used to define clinical outcomes.

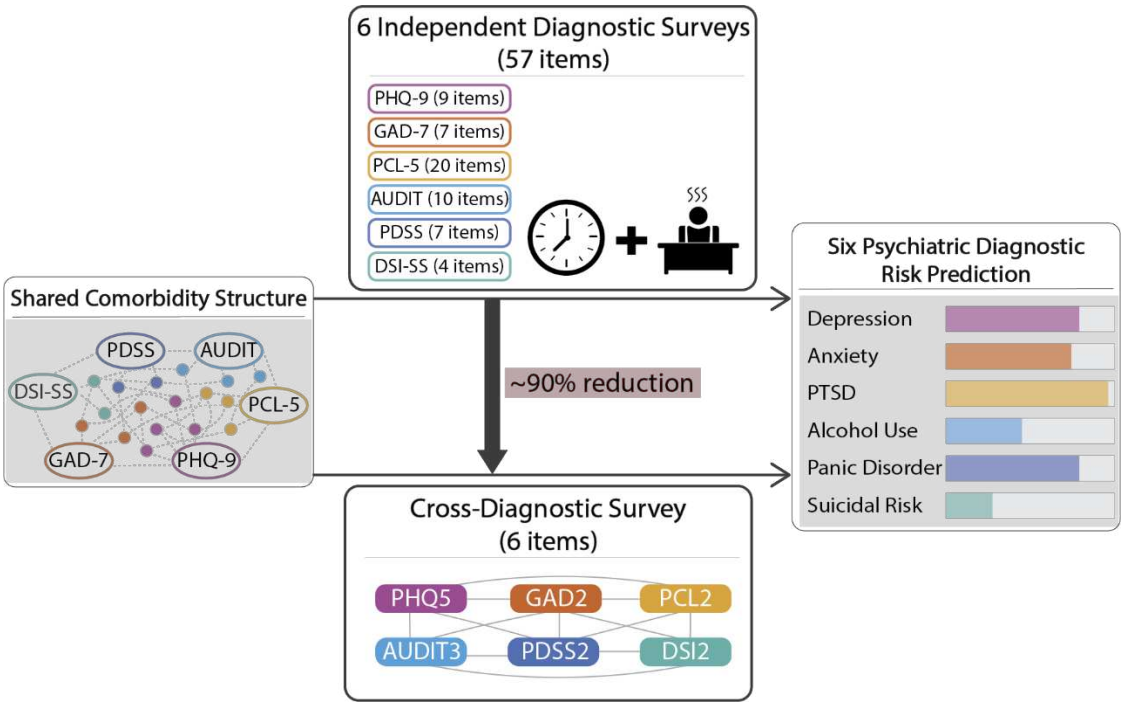
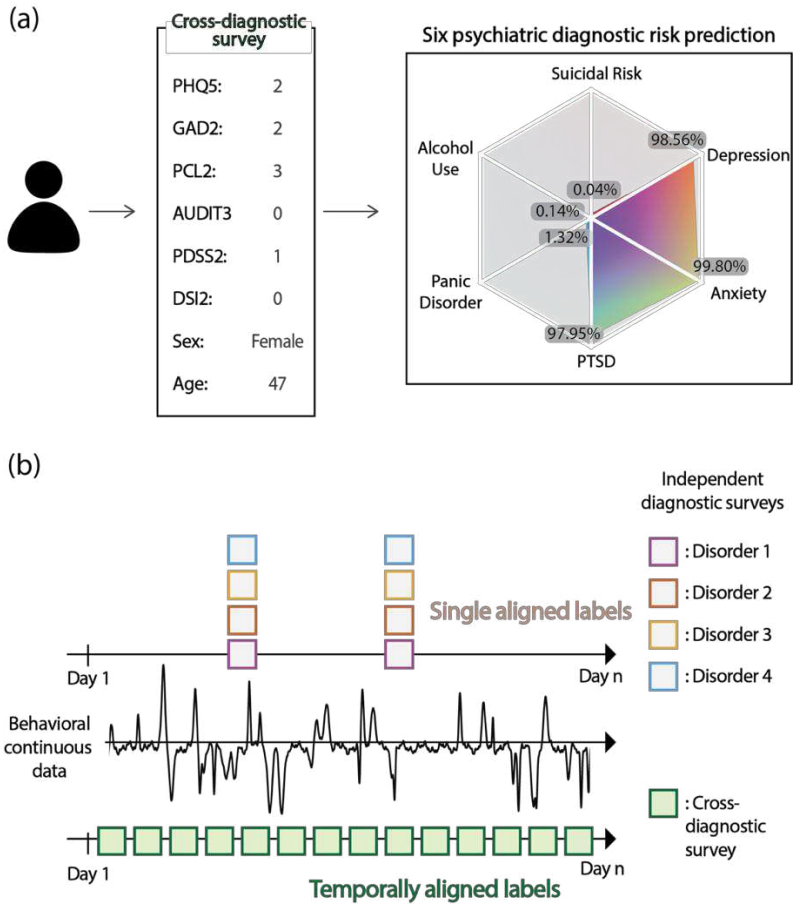


Fig. 1 Cross-diagnostic item selection framework for integrated multi-diagnostic prediction. Conventional psychiatric screening relies on six independent diagnostic surveys (PHQ-9, GAD-7, PCL-5, AUDIT, PDSS, DSI-SS), requiring a total of 57 items and treating each disorder separately despite substantial shared comorbidity and overlapping symptom structure (top and left). In contrast, the proposed framework leverages this shared structure to select one representative item per instrument (PHQ5, GAD2, PCL7, AUDIT3, PDSS4, DSI2), compressing the assessment from 57 to 6 items (~90% reduction) (bottom). These cross-domain items are integrated into a single predictive model that simultaneously estimates risk across six psychiatric outcomes (right), enabling a unified multi-disorder risk profile with substantially reduced assessment burden.

Rather than introducing additional input modalities, the present study redefines the measurement framework of psychiatric disorder prediction from a cross-diagnostic perspective. We developed a cross-diagnostic survey comprising one representative item from each of six major instruments—depression (PHQ-9), anxiety (GAD-7), trauma-related symptoms (PCL-5), alcohol use (AUDIT), panic (PDSS), and suicide risk (DSI-SS)—compressing 57 items into 6 items, representing an approximately 90% structural reduction (Fig. 1). This selection was guided by a multi-objective join optimization framework that jointly considered predictive utility and informational redundancy, enabling integrative compression that reflects shared symptom architecture rather than simple disorder-specific short forms. Using these six items together with age and sex, our multi-output prediction model achieved high discriminative performance in an independent lockbox dataset, with AUROC and AUPRC values ranging from 0.90 to 0.97 and 0.82 to 0.90, respectively, across all disorders. This compact structure enables simultaneous cross-diagnostic risk estimation from only six responses, allowing rapid screening in clinical settings (Fig. 2a). Beyond reducing assessment burden, it also enables temporally aligned multi-disorder labeling, addressing a fundamental bottleneck in digital health monitoring (Fig. 2b). Ultimately, the proposed framework extends beyond survey reduction, serving as a lightweight "gold standard" for scalable training and validation of wearable and multimodal AI models.



104

105 **Fig. 2 Cross-diagnostic risk profiling and temporally aligned labeling using the six-item framework. (a)** The six-
106 item cross-diagnostic survey enables rapid estimation of individual-level risk probabilities across multiple psychiatric
107 conditions. Responses to six representative items (PHQ5, GAD2, PCL2, AUDIT3, PDSS2, and DSI2), together with age
108 and sex, are used to generate simultaneous probability estimates for six psychiatric conditions (depression, anxiety, PTSD,
109 panic disorder, alcohol use), producing an integrated multi-disorder risk profile. **(b)** Conventional disorder-specific
110 assessments generate sparse, single0timepoint labels that are misallied with continuous behavioral data streams. In
111 contrast, the compact cross-diagnostic survey can be administered more frequently, enabling temporally aligned labels
112 that better correspond to high-resolution behavioral signals in digital health monitoring settings.

Results

Data description

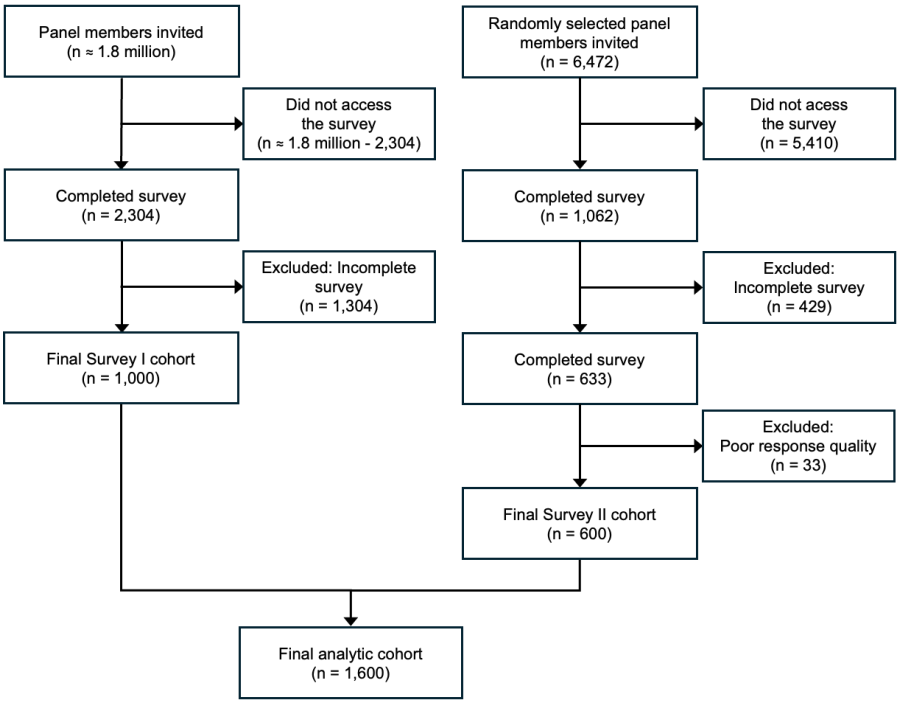


Fig 3: Flow diagram detailing subject inclusion. For Survey I, 2,304 individuals accessed the survey, and 1,000 participants were included after excluding incomplete responses. For Survey II, 1,062 individuals accessed the survey, and 600 participants were included after excluding incomplete and low-quality responses. The final analytic cohort comprised 1,600 participants.

A total of 1,000 participants were included in Survey I. Survey invitations were distributed to approximately 1.8 million registered panel members, of whom 2,304 accessed the survey and 1,000 completed it. All completed responses were anonymized and included in the final dataset (Fig. 3, left).

For Survey II, invitation emails were sent to 6,472 randomly selected panel members from the same population pool. Among them, 1,062 accessed the survey and 633 completed responses were obtained. After applying predefined data quality filters, 600 responses were retained for analysis (Fig. 3, right).

Among the collected sample, approximately 54.3-56.4% met the clinical cut-offs for at least one of the six assessments. Psychiatric symptom prevalence in both samples indicated similar prevalence of depression, anxiety, suicidality, panic disorder, PTSD, and alcohol use (Table 1). The proportion of the sample exceeding two or more clinical cut-off scores—reflecting potential multi-disorder diagnosis—was 31.6 - 33.2%.

Six cross-diagnostic survey items preserve multi-disorder predictive performance

The cross-diagnostic feature selection framework reduced 57 questionnaire items across six psychiatric instruments to a minimal set of six representative items—one per survey—corresponding to an approximately 90% reduction in assessment length. The resulting shortened instrument consisted of PHQ5, GAD2, PCL2, AUDIT3, PDSS2, and DSI2, each representing the most informative item from its respective questionnaire. To evaluate whether this substantial dimensionality reduction preserved clinically meaningful predictive information, we assessed discriminative performance on the fully held-out lockbox dataset using area under

the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC) for each disorder (Fig. 4a).

Using only these six items, together with age and sex as covariates, a CatBoost multi-output classifier achieved high discriminative performance across all six outcomes. Lockbox AUROC values were 0.963 (PHQ-9), 0.922 (GAD-7), 0.909 (PCL-5), 0.942 (AUDIT), 0.977 (PDSS), and 0.970 (DSI-SS), with corresponding AUPRC values of 0.878, 0.894, 0.865, 0.829, 0.903, and 0.829, respectively.

Despite the substantial reduction in item count, predictive performance remained robust and clinically meaningful across all disorders. These findings indicate that cross-diagnostic predictive signal is not diffusely distributed across all questionnaire items. Rather, it is concentrated within a compact, cross-diagnostic informative core set. This suggests that multi-disorder psychiatric risk profiling can be achieved without requiring multiple full-length assessments.

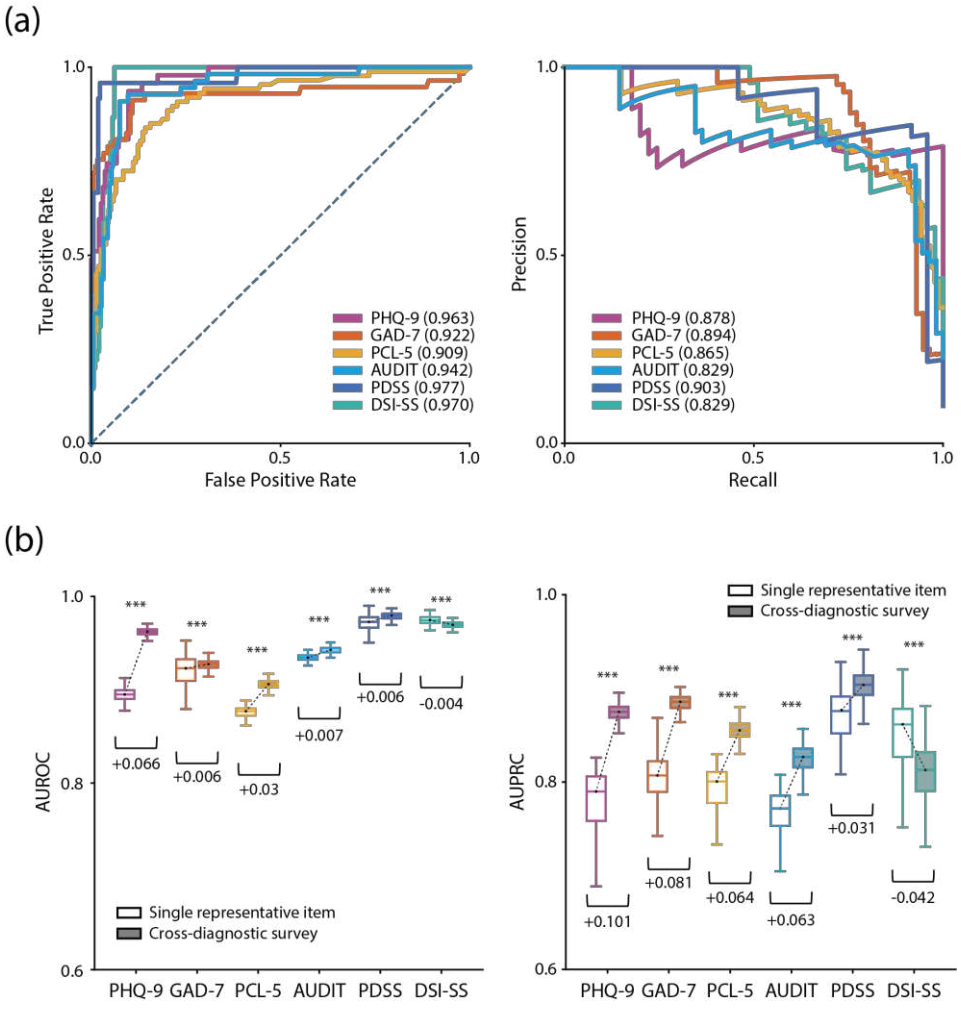


Fig. 4: The cross-diagnostic survey model achieves high predictive performance across six psychiatric outcomes Predictive performance of cross-diagnostic survey models and comparison with single-item models. **(a)** Receiver operating characteristic (ROC) curves and precision–recall (PR) curves for the six-item cross-diagnostic survey model evaluated on the independent lockbox dataset. The model achieved high discriminative performance across all six outcomes. **(b)** Performance comparison between disorder-specific single-item models (+age and sex) and the six-item cross-diagnostic survey model (+age and sex) on the lockbox dataset across 300 bootstrap iterations. Boxplots show the distribution of AUROC (left) and AUPRC (right) estimates for each disorder. Points indicate mean performance across bootstrap samples, with lines connecting paired estimates from single representative item and cross-diagnostic survey models. Annotated values represent the performance difference (Δ) between the two models. Statistical significance was assessed using two-sided t-tests across bootstrap samples (***P < 0.001).

Cross-diagnostic survey items provide disorder-specific additive predictive signal

We next evaluated whether representative items from other disorders provide additional predictive information beyond that captured by a target disorder's own representative item. Specifically, we tested whether predictive information relevant to a target disorder is partially embedded in representative items from other psychiatric domains.

To quantify cross-diagnostic information, we compared disorder-specific single representative item models (target representative item + age + sex) with cross-diagnostic survey models (all six representative items + age + sex). Performance was evaluated on the independent lockbox dataset using AUROC and AUPRC, with estimates derived from 300 bootstrap iterations. Differences between models were assessed using paired two-sided t-tests across bootstrap iterations (Fig. 4b).

Cross-diagnostic survey integration yielded heterogeneous performance changes across disorders. For PHQ-9 (depression), cross-diagnostic survey modeling produced the largest gains, with AUPRC increasing from 0.772 to 0.873 (Δ = +0.101) and AUROC from 0.895 to 0.961 (Δ = +0.066). GAD-7 (anxiety) showed moderate AUPRC improvement (0.802 to 0.883; Δ = +0.081), with minimal AUROC change (Δ = +0.006). PCL-5 (PTSD) and AUDIT (alcohol use) also demonstrated improvement (AUPRC Δ = +0.064 and +0.063, respectively; AUROC Δ = +0.030 and < +0.01). PDSS (panic disorder) showed comparatively smaller gain (AUPRC Δ = +0.031; AUROC Δ < +0.01), while DSI-SS (suicidality) showed minimal AUROC change and a reduction in AUPRC (0.852 to 0.810; Δ = −0.042). All differences were statistically significant across bootstrap iterations (all $P < 0.001$).

These results demonstrate that cross-diagnostic representative items provide statistically significant and non-redundant predictive information across disorders. Mood-, anxiety-, trauma-, and alcohol use related outcomes (PHQ-9, GAD-7, PCL-5, AUDIT) exhibited greater cross-diagnostic integration. In contrast, panic and suicidality outcomes (PDSS, DSI-SS) were predominately driven by disorder-specific signal, with limited benefit from cross-diagnostic integration.

Disorders with larger cross-diagnostic survey gains exhibit distributed cross-diagnostic attribution patterns

To understand the observed differences in multi-output performance, we analyzed feature attribution patterns for each outcome predicted by the six-item model. SHAP values were computed and normalized for each disorder outcome, enabling comparison of relative feature contributions and quantification of the extent to which each prediction relied on its disorder-specific anchor items versus cross-diagnostic survey items. (Fig. 5a).

For PHQ-9 prediction, although PHQ5 served as the dominant anchor feature (normalized to 1.00), substantial additional contributions were observed from GAD2 (0.73) and DSI2 (0.53), indicating meaningful cross-diagnostic influence (Fig. 5a, top left). For GAD-7 prediction, PHQ5 contributed 0.25 relative to the anchor GAD2 (1.00) (Fig. 5a, top middle). Similarly, for PCL-5 prediction, GAD2 contributed 0.32 alongside the anchor PCL2 (1.00) (Fig. 5a, top right). Overall, PHQ-9, GAD-7, and PCL-5 exhibited distributed SHAP contribution patterns, with predictive signals extending beyond their own representative item.

In contrast, PDSS and DSI-SS displayed more concentrated attribution structures (Fig. 5a, bottom). For these outcomes, the disorder-specific representative item accounted for the majority of predictive contribution, with minimal input from cross-diagnostic features. Contributions from non-anchor items were substantially lower than those observed for PHQ-9, GAD-7, and PCL-5.

Notably, AUDIT did not follow this pattern. Despite demonstrating measurable performance gains in the multi-input setting (Fig. 4b), its SHAP attribution remained largely concentrated on the anchor features, showing limited cross-diagnostic contribution (Fig. 5a). This mismatch between performance improvement

and attribution pattern suggests that SHAP alone may not fully capture the mechanism of cross-diagnostic integration completely.

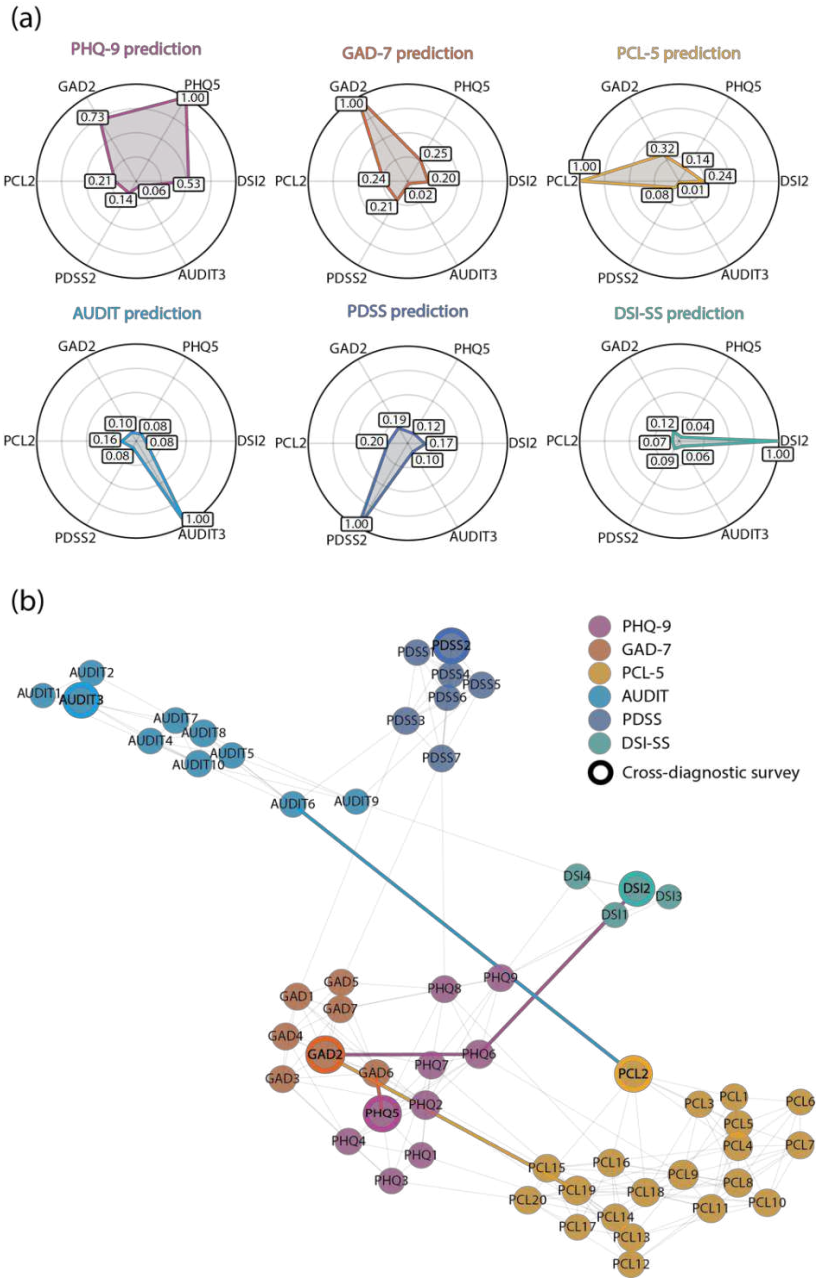


Fig. 5: SHAP attribution profiles and item-level network structure of selected representative items. (a) SHAP values for each disorder predicted by the six-item cross-diagnostic survey model, normalized to the disorder-specific representative item (anchor = 1.00). PHQ-9 prediction shows cross-diagnostic contributions from GAD2 (0.73) and DSI2 (0.53); GAD-7 prediction shows contribution from PHQ5 (0.25); PCL-5 prediction shows contribution from GAD2 (0.32); and AUDIT prediction shows contribution from PCL2 (0.16). In contrast, PDSS and DSI-SS exhibit more concentrated attribution patterns, with prediction primarily driven by their respective representative items. **(b)** GGM based on partial correlations among all survey items. Nodes represent individual items (colored by survey), and edges indicate non-zero partial correlations. Bold edges highlight cross-diagnostic connections involving selected representative items that contribute to cross-domain prediction and reflect the structural pathways through which information is shared across disorders, including GAD2–PHQ6, DSI2–PHQ6, PHQ5–GAD6, GAD2-PCL19, and PCL2-AUDIT6. PDSS- and DSI-related items form comparatively segregated subnetworks with fewer cross-instrument connections.

Item-level network structure mirrors attribution and performance differences

Although SHAP attribution patterns explained disorder-specific differences in cross-diagnostic survey performance, they did not fully account for all observed performance patterns, particularly for AUDIT. To evaluate whether attribution heterogeneity was grounded in underlying item-level structure, we constructed a GGM based on partial correlations across all survey items (Fig. 5b).

The resulting network revealed distinct cross-diagnostic connectivity patterns consistent with the attribution findings. For PHQ-9, GAD2 and DSI2 were directly connected to PHQ6, a central node within the PHQ-9 subnetwork. Although PHQ5 was the selected representative item, these cross-diagnostic items were structurally adjacent to the PHQ-9 core network, providing indirect access to central PHQ-9 nodes. This aligns with their elevated SHAP contributions, and the substantial cross-diagnostic survey performance gain observed for PHQ-9 prediction.

For GAD-7, PHQ5 was directly connected to GAD6, indicating structural adjacency between PHQ-9 and GAD-7 subnetworks. This cross-network linkage corresponds to the prominent SHAP contribution of PHQ5 and the distributed attribution pattern underlying GAD-7 prediction.

For PCL-5, GAD2 formed a direct connection to PCL19, linking the PCL-5 domain to the broader cross-diagnostic network despite PCL2 being relatively peripheral. This bridging configuration corresponds to the notable SHAP contribution of GAD2 in PCL-5 prediction and observed cross-diagnostic survey performance gain.

For AUDIT, PCL2 was directly connected to AUDIT6, indicating a structural linkage between trauma-related and alcohol use domains. Notably, this cross-domain connection is not strongly reflected in SHAP attribution patterns, suggesting that structurally mediated predictive information may not always manifest as large feature-level contributions. This provides a plausible explanation for the discrepancy between attribution results and performance improvement observed for AUDIT.

In contrast, PDSS and DSI-SS formed comparatively segregated subnetworks with fewer direct cross-instrument connections. These items exhibited limited structural adjacency to other disorder domains, consistent with their concentrated SHAP attribution profiles and relatively modest cross-diagnostic survey gains.

Together, these findings suggest that while SHAP captures part of the cross-diagnostic mechanism, item-level network structure provides a more complete explanation of performance heterogeneity. In particular, GGM analysis compensates for limitations of attribution-based interpretation by revealing latent structural pathways through which cross-diagnostic information may contribute to prediction.

Discussion

In this study, we addressed the structural bottleneck of cross-diagnostic mental health assessment by redesigning the measurement architecture while maintaining robust predictive performance. By compressing 57 items from six psychiatric instruments into 6 representative items (~90% reduction), our framework preserved cross-diagnostic predictive performance, with AUROC values ranging from 0.90 to 0.97. These findings empirically mitigate the long-assumed trade-off between assessment burden and predictive accuracy, offering an immediately deployable solution for scalable digital monitoring. Importantly, the proposed framework does not simply shorten existing questionnaires but reconceptualizes the measurement structure to a transdiagnostic perspective by capturing shared predictive information across disorders[29,30,31,32,33]. The preservation of performance under such substantial compression suggests that multi-condition risk signals are

not evenly distributed across questionnaire items but instead concentrated within a limited set of symptom dimensions. Notably, even within this general population sample, 31.6-33.2% of participants met clinical cut-offs for two or more disorders (Table 1), highlighting the high prevalence of cross-diagnostic symptom burden. In clinical populations, where psychiatric comorbidity is typically higher, the need for efficient multi-condition screening frameworks may be even more pronounced.

The significance of this structural redesign also lies in the interpretability it affords for understanding how cross-diagnostic risk signals emerge within psychiatric symptom structures. In much of the current digital health literature, predictive improvements have been pursued primarily through increasingly complex input modalities—such as video, speech analysis, and wearable sensors[7,9,10,34,35,36,37,38,39]. Although these multimodal approaches can achieve strong predictive performance, their high-dimensional behavioral inputs often have unclear relationships to specific clinical symptoms[40,41]. In contrast, the present framework generates predictions from a compact set of clinically interpretable symptom dimensions. This enables direct examination of feature contributions and facilitates interpretation of model outputs within established psychiatric constructs, linking predictive signals to the broader symptom architecture underlying comorbidity.

Further insights into the structural role of the selected representative symptoms emerged from integrating SHAP attribution with a GGM network analysis, revealing how the model utilizes intrinsic symptom topology [42,43]. Across internalizing disorders, cross-diagnostic items occupying bridging positions in the GGM network emerged as key predictive contributors. For instance, GAD2 (uncontrollable worry) and DSI2 (suicidal ideation) were directly connected to the central depressive node PHQ6 (suicidal thoughts / self-harm ideation) (Fig. 6b). Consistent with this structural adjacency, depression prediction relied not only on its anchor feature (PHQ5; appetite changes) but also on cross-diagnostic inputs from anxiety and suicidality features (Fig. 6a, top left). Similar patterns were observed across other domains: PHQ5 (appetite changes) connects to GAD6 (irritability) (Fig. 6b), explaining why anxiety prediction incorporates depressive signals (Fig. 6a, top middle); likewise, GAD2 (uncontrollable worry) bridges to PCL19 (difficulty concentrating) (Fig. 6b), contributing to trauma-related prediction (Fig. 6a, top right). Notably, this cross-dianostic bridging extends beyond internalizing domains, as PCL2 was directly connected to AUDIT6, indicating a structural linkage between trauma-related and alcohol use domains. This connection aligned with the observed contribution of PCL2 in AUDIT prediction, supporting the inclusion of AUDIT among disorders exhibiting cross-diagnostic integration. These findings align with transdiagnostic frameworks such as the Hierarchical Taxonomy of Psychopathology (HiTOP), which proposes the internalizing spectrum as a higher-order dimension that encompasses distress and fear-related syndromes[4]. From a clinical perspective, symptoms representing central transdiagnostic processes—such as uncontrollable worry—may function as efficient indicators of vulnerability across multiple psychiatric conditions and represent potential targets for both screening and intervention.

In contrast, panic disorder (PDSS) and suicidal ideation (DSI-SS) exhibited more segregated structural patterns (Fig. 6a, bottom). These conditions showed weaker connectivity within the broader symptom network (Fig. 6b), indicating greater disorder-specificity in their symptom organization[3]. This pattern aligns with the HiTOP framework, which distinguishes between a shared "distress" dimension and more specialized sub-factors[4]. While depression and anxiety typically load strongly on the shared distress dimension[44], alcohol use is often conceptualized within the externalizing behavior[45], and panic disorder may reflect a more specific fear-related sub-factor characterized by acute autonomic arousal[46]. Clinically, this suggests that although a small set of symptom indicators can effectively capture transdiagnostic psychological distress, high-risk outcomes such as suicidality and alcohol misuse may still require disorder-specific assessment modules to preserve diagnostic sensitivity[47]. These findings highlight the importance of balancing transdiagnostic screening efficiency with disorder-specific precision.

Beyond interpretability, the proposed framework has immediate practical implications for real-world screening. Because the model requires only six brief questionnaire items along with age and sex, multi-disorder risk estimates can be generated almost instantly without specialized equipment or preprocessing. This enables simultaneous screening of probabilistic risk estimates across depression, anxiety, PTSD, panic disorder, alcohol use, and suicidality, producing an integrated cross-diagnostic risk profile (Fig. 2a). This capability is particularly valuable in time-constrained settings, including primary care, emergency screening,

and digital intake systems[48]. Unlike multimodal approaches that depend on video recording, speech transcription, or wearable sensor integration, the present framework can be deployed through a simple questionnaire interface, allowing immediate integration into clinical and digital workflows.

Importantly, this compact measurement structure also addresses a key limitation in how clinical labels are generated for digital health models. Although behavioral sensing technologies have rapidly advanced, most predictive systems still rely on traditional questionnaire scores as their primary clinical outcomes. In practice, these questionnaires are administered only intermittently[27,49], creating a temporal misalignment between continuously collected behavioral signals and sparsely measured clinical states. Moreover, disorder-specific instruments are often administered independently, making it difficult to obtain temporally aligned multi-diagnostic labels[50] (Fig. 2b, top). As additional psychiatric conditions are included, the cumulative burden of multiple instruments further increases, often leading to participant fatigue and reduced data quality in longitudinal monitoring settings[28]. By compressing several disorder-specific instruments into a unified cross-diagnostic structure, the proposed framework substantially reduces assessment burden while preserving clinically meaningful information. This enables more frequent and repeatable assessments, allowing clinical labels to be collected at higher temporal resolution and synchronously across multiple disorders (Fig. 2b, bottom).

These improvements in label frequency and temporal alignment have important implications for the development of digital mental health algorithms. When high-resolution behavioral data are paired with sparse or misaligned clinical labels, the supervisory signal for model training is limited[51]. The proposed framework enables synchronized and repeated collection of multi-condition labels, providing a more coherent representation of a patient's evolving mental health state. In turn, these labels can serve as stronger supervisory signals for training complex multimodal and wearable-based prediction models[52], supporting more scalable and clinically grounded digital mental health systems.

Despite these promising findings, several directions for future research remain important. First, validation in independent external cohorts is necessary to establish generalizability. Because the present study used Korean-language versions of the questionnaires in a Korean population, cross-cultural validation will be important. Second, the clinical utility of the six-item framework should be evaluated in real-world settings, including its impact on decision making and patient outcomes. Third, usability and sustainability in real-world digital environments—such as mobile health applications and digital therapeutics platforms —require further evaluation. Repeated administration of screening instruments may influence completion rates, response patterns, and missingness, all of which can affect the long-term viability of digital monitoring systems. Finally, longitudinal studies integrating this framework with multimodal data are needed to determine whether reduced-burden labeling can support larger-scale data collection and improve predictive modeling. Together, these efforts will clarify whether structurally simplified measurement architectures can enable scalable and clinically meaningful digital health monitoring in real-world settings.

Methods

Participants

We collected data from the Korean general population using anonymous online surveys administered by a professional survey company (EMBRAIN).

Survey I was conducted from December 9 to 12, 2024. A total of 1,000 participants were recruited using quota sampling, with approximately equal allocation across twelve demographic cells defined by sex and five age groups (18–29, 30s, 40s, 50s, and ≥60).

Survey II was conducted from October 23 to November 4, 2025. A target sample size of 600 participants was determined, with 50 respondents allocated to each of twelve demographic cells defined by sex and six age groups (18–29, 30s, 40s, 50s, 60s, and 70–79).

For Survey II, additional data quality control procedures were applied. A speed-check filter excluded the fastest 5% of respondents within each quota group to remove potential inattentive responses. Responses with an average item response time exceeding three times the overall mean were also excluded.

All participants provided informed agreement by selecting "yes" prior to beginning the survey.

The study protocols were reviewed and approved by the Institutional Review Board of Asan Medical Center (Survey I: 2024-1236; Survey II: 2025-1331), and the requirement for written informed consent was waived in accordance with institutional guidelines.

Measures

1) Patient Health Questionnaire-9 (PHQ-9)

Depressive symptoms were assessed using the Korean version of the Patient Health Questionnaire-9 (PHQ-9)[53,54], a 9-item self-report screening measure of depressive symptoms over the past two weeks. Items are rated on a 4-point scale (0 = not at all to 3 = nearly every day), yielding total scores from 0 to 27, with higher scores indicating greater symptom severity. A cut-off score of ≥9 was used to indicate clinically significant depressive symptoms, consistent with prior Korean validation[54].

2) Generalized Anxiety Disorder-7 (GAD-7)

Anxiety symptoms were measured with the Korean version of the Generalized Anxiety Disorder 7 (GAD-7)[55,56]. The GAD-7 consists of seven items assessing generalized anxiety symptoms experienced over the past two weeks. Each item is rated on a 4-point Likert scale (0 = not at all to 3 = nearly every day), with total scores ranging from 0 to 21. A cut-off score of ≥5 was used to define at least mild anxiety symptom severity[55].

3) PTSD Checklist for DSM-5 (PCL-5)

PTSD symptoms were assessed using the Korean version of the PTSD Checklist for DSM-5 (PCL-5)[57,58]. The PCL-5 is a 20-item self-report measure rated on a 5-point scale (0 = not at all to 4 = extremely), with total scores ranging from 0 to 80. A cut-off score of ≥28 was used to define clinically significant PTSD symptoms, consistent with prior recommendations[59].

4) Alcohol Use Disorder Identification Test (AUDIT)

Alcohol use problems were assessed using the Korean Version of Alcohol Use Disorders Identification Test (AUDIT)[60,61]. The AUDIT is a 10-item self-report measure designed to screen for hazardous and harmful alcohol consumption. Each item is rated on a 5-point scale (0-4), yielding total scores ranging from 0 to 40. A cut-off score of 14 for males and 10 for females was used to define problematic alcohol use.

5) The Self-Report version of the Panic Disorder Severity Scale (PDSS)

Panic disorder severity was assessed using the Korean version of the self-report version of the Panic Disorder Severity Scale (PDSS)[62]. This 7-item self-report measure rates the severity of panic disorder and associated impairment on a 5-point scale (0-4), with total scores ranging from 0 to 28. A cut-off score of ≥7 was used to define a positive screening result for panic disorder.

6) Depression Symptom Inventory - Suicidality Subscale (DSI-SS)

Suicide risk was measured using the Korean Depressive Symptom Inventory-Suicidality Subscale (DSI-SS)[63,64], a 4-item self-report screening tool derived from the Hopelessness Depression Symptom

Questionnaire. Items are rated on a 4-point scale (0-3), with total scores ranging from 0 to 12. A cut-off score of 4 was used to define elevated suicide risk, consistent with prior validation[64].

Multi-objective joint optimization framework

Instead of independently identifying optimal items for each disorder, we framed cross-diagnostic survey development as a constrained, multi-objective joint optimization problem to account for cross-diagnostic efficiencies. Unlike traditional heuristic methods that reduce scales independently, our framework simultaneously optimizes a minimal six-item symptom set—one item drawn from each diagnostic domain—that jointly maximizes cross-diagnostic predictive performance while minimizing informational redundancy across the latent psychopathology space.

Let $k \in \{1, \dots, 6\}$ index the target psychiatric disorders (PHQ-9, GAD-7, PCL-5, AUDIT, PDSS and DSI-SS) and $j \in S_k$ denote the candidate items belonging to the respective disorder k. We defined binary decision variables $z_{j,k} \in \{0,1\}$, where $z_{j,k} = 1$ indicates that item j is selected as the representative for disorder k. To ensure a structurally interpretable instrument, we imposed a strict constraint of selecting exactly one item per diagnostic category:

$$\sum_{j \in S_k} z_{j,k} = 1.$$

This constraint preserves broad domain coverage while reducing the total assessment burden from 57 to 6 items (Fig. 4a).

To balance predictive utility and informational diversity, the optimization problem is formulated as:

$$max \sum_{k=1}^{6} \sum_{j \in S_k} U_{j,k} z_{j,k} \; - \lambda \sum_{p<q} cos\,(X_p, X_q) x_p x_q.$$

The first term aggregates the predictive utility of the selected items across all six disorders, whereas the second term imposes a redundancy penalty to discourage the simultaneous selection of highly similar items.

To promote predictive utility, $U_{j,k}$ is defined as:

$$U_{j,k} = \frac{1}{2} AUROC\big(f\big(X_j, Age, Sex\big), Y_k\big) + \frac{1}{2} AUPRC\big(f\big(X_j, Age, Sex\big), Y_k\big),$$

where $f(\cdot)$ denotes a logistic regression model adjusted for demographic covariates. By combining AUROC (threshold-independent discrimination) with AURPC (positive-class sensitivity under class imbalance), this composite metric provides a robust estimate of standalone predictive value.

The second term of the objective function penalizes similarity between selected items to promote informational diversity and maximize latent space coverage. Here, information redundancy is quantified using the cosine similarity between standardized response vectors X_p and X_q:

$$cos\,(X_p, X_q) = \frac{X_p^{\top} X_q}{\| X_p \| \| X_q \|}.$$

In this formulation, x_p indicates the selection of item p in the final composite set, and the hyperparameter $\lambda \geq 0$ controls the stringency of the redundancy penalization.

This regularization discourages the selection of highly correlated items and favors symptom dimensions that span distinct regions of the empirical symptom manifold. Crucially, this framework does not assume

independence between disorders. Instead, it explicitly leverages cross-domain correlations to identify a compact symptom basis that captures overlapping psychopathological signals, enabling a unified multi-disorder risk profile with minimal items.

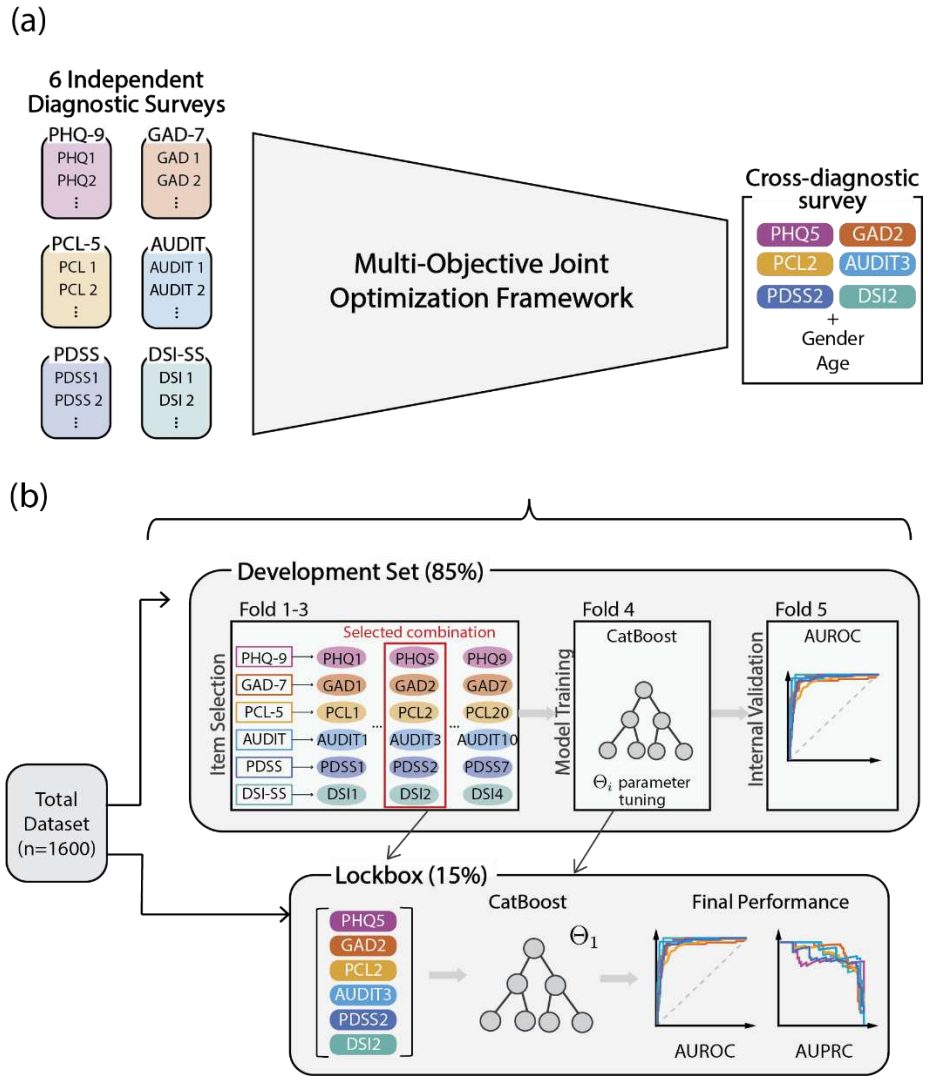


Fig 6: Multi-objective joint optimization framework for cross-diagnostic survey construction and validation. (a) A multi-objective joint optimization framework was used to select one representative item from each of six psychiatric instruments (PHQ-9, GAD-7, PCL-5, AUDIT, PDSS, DSI-SS), yielding a six-item cross-diagnostic survey. The framework leverages shared symptom structure across instruments to identify an optimal combination of representative items for integrated multi-diagnostic prediction. (b) The total dataset (n = 1600) was divided into a development set (85%) and an independent lockbox set (15%). Within the development set, folds 1–3 were used for item selection, fold 4 for CatBoost model training (parameters denoted as θ_i), and fold 5 for internal validation using AUROC. The finalized cross-diagnostic survey items and optimized model parameters (θ_i) were subsequently evaluated on the independent lockbox dataset to obtain final performance metrics (AUROC and AUPRC).

Data partitioning and validation strategy

The total dataset ($N = 1600$) was partitioned into a development set (85%) and an independent lockbox set (15%) using stratified sampling to preserve disorder prevalence.

The lockbox dataset was strictly isolated at the outset of the study and remained inaccessible during all stages of item selection, model training, and internal validation. This strict separation ensured that model development was conducted independently of the data used for final evaluation, thereby preserving the

integrity of generalization estimates.

Within the development set, we implemented a five-fold structure where each fold was assigned a predefined functional role to prevent information leakage (Fig. 4b):

- **Folds 1–3 (Item selection)**: These folds were utilized to perform joint optimization of the six-item composite, where candidate combinations were evaluated based on their aggregated predictive utility across all disorders.
- **Fold 4 (Model training)**: Using the selected feature set, a CatBoost classifier was trained[65], and its hyperparameters were optimized exclusively on this fold.
- **Fold 5 (Internal validation)**: The tuned model was evaluated on this fold to obtain an unbiased performance estimate within each development iteration

We incorporated 300 bootstrap iterations within the development pipeline to ensure robustness to sampling variability. In each iteration, the complete selection–training–validation procedure was repeated, and selection frequencies across candidate six-item sets were tracked. The most frequently selected configuration was chosen as the final cross-diagnostic survey set. This stability-driven approach reduces sensitivity to data partitioning and ensures that the selected feature set reflects a stable underlying signal rather than a sampling-induced artifact.

Independent lockbox evaluation and performance metrics

We conducted a comprehensive evaluation using the independent lockbox dataset to obtain an unbiased assessment of the finalized cross-diagnostic survey.

The evaluation was performed using the optimal 6-item combination (PHQ5, GAD2, PCL2, AUDIT3, PDSS2, and DSI2) identified through the stability analysis. For all predictive tasks, individual models incorporated these representative items along with age and sex as essential demographic covariates to ensure clinical relevance. The final predictive performance for each of the six psychiatric outcomes was quantified using the independent lockbox set, with discriminative accuracy and precision visualized as AUROC and AUPRC profiles (Fig. 6a). To ensure robust estimation, performance metrics were reported as means from 300 bootstrap repetitions.

A core objective of this study was to determine whether integrating cross-diagnostic signals provides an additive predictive advantage over disorder-specific screening. To test this, we compared two distinct modeling architectures (Fig. 6b):

- **Single representative item models**: Utilized only the designated representative item of the target disorder as the primary feature, supplemented by age and sex.
- **Cross-diagnostic survey models**: Utilized the complete set of all six representative items simultaneously, alongside age and sex. This comparative framework allowed us to quantify the performance gain—defined as the incremental improvement in AUROC and AUPRC—attributable to the integration of cross-diagnostic predictive information.

The discriminative ability of the models was evaluated through threshold-independent metrics, primarily focusing on the AUROC and the AUPRC. The latter was particularly emphasized to assess sensitivity in the context of varying disorder prevalence within the sample. Statistical significance for the performance differences between single representative item and cross-diagnostic architectures was determined by examining the non-overlap of bootstrap-derived the delta (Δ) values for both AUROC and AUPRC across all outcomes. Furthermore, a paired t-test was employed to statistically verify the performance gap between the

single representative item and cross-diagnostic survey models based on the bootstrap sampling distributions, ensuring that the observed predictive gains were statistically robust.

Model interpretability using SHAP values

We employed SHAP values to determine whether cross-diagnostic survey performance gains reflected specific contribution structures and calculated a normalized contribution ratio ($s_{d,i}$) for each selected item:

$$s_{d,i} = \frac{\sum_n \mid SHAP_{d,i}(n) \mid}{\sum_j \sum_n \mid SHAP_{d,j}(n) \mid}.$$

This metric allowed us to distinguish between single-item dominance, where the signal is concentrated in a primary anchor item, and distributed cross-contribution, where the predictive signal is shared across multiple domains (Fig. 6a).

Inter-item network analysis via Gaussian Graphical Modeling

We constructed a Gaussian Graphical Model (GGM) using the full 57 questionnaire items survey pool to further characterize whether the observed heterogeneity in predictive performance and SHAP attribution patterns were grounded in underlying symptom structure. While SHAP quantifies feature contributions within the predictive model, the GGM provides a complementary structure-level view of conditional relationships among items.

The GGM was estimated using partial correlations, which quantify conditional dependence relationships between items while controlling for all remaining variables. Let S denote the empirical covariance matrix of the standardized item responses. The precision matrix Θ (inverse covariance matrix) was estimated using the graphical lasso:

$$\hat{\Theta} = \arg\min_{\Theta \succ 0} (-\log\det\Theta + tr(S\Theta) + \rho\|\Theta\|_1),$$

where ρ is a regularization parameter controlling sparsity. The optimal value of ρ was selected using the Extended Bayesian Information Criterion (EBIC).

In this topological framework, each node represents an individual 57 questionnaire items. The presence of an edge corresponds to non-zero off-diagonal entries in the estimated precision matrix Θ, representing the partial correlation coefficient between two specific items. These edges effectively encode conditional independence relationships, indicating that the associated symptoms maintain a significant relationship even after adjusting for the presence of all other symptoms in the network. For visualization purposes, only edges with absolute partial correlation magnitude greater than 0.03 were displayed (Fig. 6b). This threshold was applied to enhance interpretability and reduce visual clutter while preserving the core structural topology of the network.

We conducted a topological analysis to examine whether predictive integration was grounded in the relational architecture of survey items. First, for each disorder outcome, we identified cross-diagnostic representative items exhibiting elevated SHAP contribution values in the cross-diagnostic survey model. Second, within the GGM, we evaluated two key structural properties:

- **Structural adjacency:** Whether cross-diagnostic items with high SHAP contributions were directly connected (i.e., shared an edge) with central nodes within the target disorder's subnetwork.

- **Connectivity patterns:** Whether disorders demonstrating larger cross-diagnostic survey performance

gains occupied more interconnected or bridge-like positions in the global symptom network.

By comparing cross-diagnostic survey performance gains, SHAP attribution distributions, and GGM connectivity patterns, we evaluated whether the predictive integration observed in the model was supported by the underlying structural interdependence of psychiatric symptoms.

The multi-objective joint optimization framework was implemented in Python. Training and evaluation of the CatBoost models was performed using the "catboost" package for Python. Hyperparameter tuning for the CatBoost models was conducted using the "bayes_opt" package for Python. SHAP value analysis was performed using the "shap" package for Python. GGM analysis was performed using the "GraphicalLasso" implementation from the scikit-learn Python library.

Data availability

The data that support the findings of this study are available from EMBRAIN, but restrictions apply to the availability of these data, which were used under licence for the current study and so are not publicly available. The access to the data was facilitated by one of the corresponding authors of this study - S.S, J.B., and S.C. - who served as a principal investigator of the EMBRAIN study.

Code availability

The trained CatBoost models and code for implementing these models will be provided upon acceptance.

References

1 Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry* **9**, 137–150 (2022). https://doi.org/10.1016/S2215-0366(21)00395-3

2 Kessler, R. C., Chiu, W. T., Demler, O. & Walters, E. E. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* **62**, 617–627 (2005).

3 Caspi, A. *et al.* The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clinical psychological science* **2**, 119–137 (2014).

4 Kotov, R. *et al.* The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology* **126**, 454 (2017).

5 Thornicroft, G. *et al.* Evidence for effective interventions to reduce mental-health-related stigma and discrimination. *The Lancet* **387**, 1123–1132 (2016).

6 Patel, V. *et al.* The Lancet Commission on global mental health and sustainable development. *The lancet* **392**, 1553–1598 (2018).

7 Grimm, B., Yilmam, P., Talbot, B. & Larsen, L. Multimodal machine learning for video based single question mental health assessment. *npj Digital Medicine* (2025).

8 Pampouchidou, A. *et al.* in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge.* 27–34.

9 Lee, M. P. *et al.* The real-world association between digital markers of circadian disruption and mental health risks. *NPJ digital medicine* **7**, 355 (2024).

10 Lim, D. *et al.* Accurately predicting mood episodes in mood disorder patients using wearable sleep and circadian rhythm features. *NPJ Digital Medicine* **7**, 324 (2024).

11 Song, Y. M. *et al.* Causal dynamics of sleep, circadian rhythm, and mood symptoms in patients with major depression and bipolar disorder: insights from longitudinal wearable device data. *EBioMedicine* **103** (2024).

12 Rykov, Y., Thach, T.-Q., Bojic, I., Christopoulos, G. & Car, J. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. *JMIR mHealth and uHealth* **9**, e24872 (2021).

13 Insel, T. R. Digital phenotyping: technology for a new science of behavior. *Jama* **318**, 1215–1216 (2017).

14 Schueller, S. M. & Torous, J. Scaling evidence-based treatments through digital mental health. *American Psychologist* **75**, 1093 (2020).

15 Li, Y. *et al.* Predicting depression by using a novel deep learning model and video-audio-text multimodal data. *Frontiers in Psychiatry* **Volume 16 - 2025** (2025). https://doi.org/10.3389/fpsyt.2025.1602650

16 Mohr, D. C. *et al.* IntelliCare: an eclectic, skills-based app suite for the treatment of depression and anxiety. *Journal of medical Internet research* **19**, e10 (2017).

17 Wang, R. *et al.* in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* 3–14 (Association for Computing Machinery, Seattle, Washington, 2014).

18 Bufano, P., Laurino, M., Said, S., Tognetti, A. & Menicucci, D. Digital phenotyping for monitoring mental disorders: systematic review. *Journal of medical Internet research* **25**, e46778 (2023).

19 Outcomes assessment in the SPRINT multicenter tibial fracture trial: Adjudication committee size has trivial effect on trial results. *Journal of Clinical Epidemiology* **64**, 1023–1033 (2011). https://doi.org/https://doi.org/10.1016/j.jclinepi.2010.12.007

20 Ben-Porath, Y. & Tellegen, A. Minnesota multiphasic personality inventory-2-restructured form. *Assessment* (2008).

21 Tate, B., Wright, J., Scott, E. & Robinson, S. Dynamic parameter morphogenesis for adaptable task specialization in large language models to self-directed learning. (2024).

22 Borghouts, J. *et al.* Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. *Journal of medical Internet research* **23**, e24387 (2021).

23 Stamatis, C. A. *et al.* Differential temporal utility of passively sensed smartphone features for depression and anxiety symptom prediction: a longitudinal cohort study. *npj Mental Health Research* **3**, 1 (2024). https://doi.org/10.1038/s44184-023-00041-y

24 Rykov, Y. G., Ng, K. P., Patterson, M. D., Gangwar, B. A. & Kandiah, N. Predicting the severity of mood and neuropsychiatric symptoms from digital biomarkers using wearable physiological

613 data and deep learning. *Computers in Biology and Medicine* **180**, 108959 (2024).
614 https://doi.org/https://doi.org/10.1016/j.compbiomed.2024.108959

615 25 Shin, Y.-B. *et al.* Development of prediction models for screening depression and anxiety using
616 smartphone and wearable-based digital phenotyping: protocol for the Smartphone and
617 Wearable Assessment for Real-Time Screening of Depression and Anxiety (SWARTS-DA)
618 observational study in Korea. *BMJ Open* **15**, e096773 (2025).
619 https://doi.org/10.1136/bmjopen-2024-096773

620 26 Ali, M. *et al.* Speech as a Multimodal Digital Phenotype for Multi-Task LLM-based Mental Health
621 Prediction. arXiv:2505.23822 (2025).
622 <https://ui.adsabs.harvard.edu/abs/2025arXiv250523822A>.

623 27 Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–
624 265 (2023).

625 28 Linardon, J., Anderson, C., Messer, M., Liu, C. & Torous, J. Transdiagnostic-focused apps for
626 depression and anxiety: a meta-analysis. *NPJ Digital Medicine* **8**, 443 (2025).

627 29 Jeon, S. *et al.* Machine-learning validated short form of the Korean version of the Sleep-Related
628 Behaviors Questionnaire-10 items: SRBQ-10. *Behavioral Sleep Medicine* **24**, 1–14 (2026).

629 30 Cawiding, O. R. *et al.* Cancer-related Dysfunctional Beliefs and Attitude about Sleep-6 (C-DBAS-
630 6): a practical and accurate shortened version using XGBoost and SymScore. *Sleep and Biological*
631 *Rhythms* **23**, 465–475 (2025).

632 31 Jo, H. *et al.* Dysfunctional Beliefs and Attitudes about Sleep-6 (DBAS-6): Data-driven shortened
633 version from a machine learning approach. *Sleep Medicine* **119**, 312–318 (2024).

634 32 Jo, H. *et al.* Data-driven shortened Insomnia Severity Index (ISI): a machine learning approach.
635 *Sleep and Breathing* **28**, 1819–1830 (2024).

636 33 Ha, S. *et al.* Predicting the risk of sleep disorders using a machine learning–based simple
637 questionnaire: Development and validation study. *Journal of medical Internet research* **25**,
638 e46520 (2023).

639 34 Weber, S. *et al.* Using a fine-tuned large language model for symptom-based depression
640 evaluation. *npj Digital Medicine* **8**, 598 (2025).

641 35 Kim, J. *et al.* Optimizing large language models for detecting symptoms of depression/anxiety in
642 chronic diseases patient communications. *NPJ Digital Medicine* **8**, 580 (2025).

643 36 Kim, J. *et al.* Large language models outperform mental and medical health care professionals in
644 identifying obsessive-compulsive disorder. *NPJ digital medicine* **7**, 193 (2024).

645 37 Stamatis, C. A. *et al.* Differential temporal utility of passively sensed smartphone features for
646 depression and anxiety symptom prediction: a longitudinal cohort study. *Npj Mental Health*
647 *Research* **3**, 1 (2024).

648 38 Balliu, B. *et al.* Personalized mood prediction from patterns of behavior collected with
649 smartphones. *NPJ Digital Medicine* **7**, 49 (2024).

650 39 Teferra, B. G. *et al.* Leveraging large language models for automated depression screening. *PLOS*
651 *Digital Health* **4**, e0000943 (2025).

40 Onnela, J.-P. & Rauch, S. L. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* **41**, 1691–1696 (2016).

41 Huckvale, K., Venkatesh, S. & Christensen, H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ digital medicine* **2**, 88 (2019).

42 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

43 Epskamp, S., Borsboom, D. & Fried, E. I. Estimating psychological networks and their accuracy: A tutorial paper. *Behavior research methods* **50**, 195–212 (2018).

44 Clark, L. A. & Watson, D. Distress and fear disorders: an alternative empirically basedtaxonomy of the 'mood'and 'anxiety'disorders. *The British Journal of Psychiatry* **189**, 481–483 (2006).

45 de Jonge, P. *et al.* The cross-national structure of mental disorders: results from the World Mental Health Surveys. *Psychological medicine* **48**, 2073–2084 (2018).

46 Krueger, R. F. The structure of common mental disorders. *Archives of general psychiatry* **56**, 921–926 (1999).

47 Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* **62**, 593–602 (2005).

48 Mitchell, A. J., Vaze, A. & Rao, S. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet* **374**, 609–619 (2009).

49 Daniore, P. *et al.* From wearable sensor data to digital biomarker development: ten lessons learned and a framework proposal. *NPJ digital medicine* **7**, 161 (2024).

50 Langener, A. M. *et al.* It's All About Timing: Exploring Different Temporal Resolutions for Analyzing Digital-Phenotyping Data. *Advances in Methods and Practices in Psychological Science* **7**, 25152459231202677 (2024). https://doi.org/10.1177/25152459231202677

51 Mohr, D. C., Zhang, M. & Schueller, S. M. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* **13**, 23–47 (2017).

52 Torous, J., Onnela, J. & Keshavan, M. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Translational psychiatry* **7**, e1053–e1053 (2017).

53 Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine* **16**, 606–613 (2001).

54 Park, S.-J., Choi, H.-R., Choi, J.-H., Kim, K.-W. & Hong, J.-P. Reliability and validity of the Korean version of the Patient Health Questionnaire-9 (PHQ-9). *Anxiety and mood* **6**, 119–124 (2010).

55 Spitzer, R. L., Kroenke, K., Williams, J. B. & Löwe, B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine* **166**, 1092–1097 (2006).

56 Lee, S. H. *et al.* Validation of the Korean version of the generalized anxiety disorder 7 self-rating scale. *Asia-Pacific Psychiatry* **14**, e12421 (2022).

57 Weathers, F. *et al.* The PTSD checklist for DSM-5 (PCL-5) national center for PTSD. *Boston, MA, USA* (2013).

58 Oh, S.-I. *et al.* Reliability and validity of the Korean version of the PTSD checklist civilian version in North Korean defectors. *Journal of Korean Neuropsychiatric Association* **53**, 410–417 (2014).

59 Marx, B. P. *et al.* Reliable and clinically significant change in the clinician-administered PTSD Scale for DSM-5 and PTSD Checklist for DSM-5 among male veterans. *Psychological assessment* **34**, 197 (2022).

60 Saunders, J. B., Aasland, O. G., Babor, T. F., De la Fuente, J. R. & Grant, M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction* **88**, 791–804 (1993).

61 Lee, B. Development of Korean version of alcohol use disorders identification test (AUDIT-K): Its reliability and validity. *Journal of Korean Academy of Addiction Psychiatry* **4**, 83 (2000).

62 Lee, E. H., Kim, J. H. & Yu, B. H. Reliability and validity of the self-report version of the Panic Disorder Severity Scale in Korea. *Depression and anxiety* **26**, E120–E123 (2009).

63 Metalsky, G. I. & Joiner Jr, T. E. The hopelessness depression symptom questionnaire. *Cognitive Therapy and Research* **21**, 359–384 (1997).

64 Suh, S. *et al.* Validation of the Korean depressive symptom inventory-suicidality subscale. *Psychiatry investigation* **14**, 577 (2017).

65 Dorogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. *ArXiv* **abs/1810.11363** (2018).

Acknowledgments

This study was funded by Institute for Basic Science (IBS-R029-C3) (J.K.K.), the National Research Foundation of Korea (NRF) (RS- 2022- nR068758) (J.K.K.), the National Institute for Mathematical Sciences (NIMS) grant funded by the Korea government (MSIT) (No. B26810000) (D.J.), and the National IT Industry Promotion Agency (NIPA) of the Republic of Korea (H1002-25-1002) (S.S., J.B., and S.C.). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author Contribution

Concept and design by D.J., S.C., and J.K.K. Recruitment of participants and data collection by S.S., J.B., and S.C. Development of algorithms by M.L., D.J., and J.K.K. Analysis of data by M.L., S.S., J.B., D.J., S.C., and J.K.K. Drafting the initial manuscript by M.L., S.S., J.B., D.J., S.C., and J.K.K. Critical revision of the manuscript for important intellectual content by M.L., S.S., J.B., D.J., S.C., and J.K.K. All authors read and approved the manuscript.

Competing interests

All authors declare no financial or non-financial competing interests.

Tables

Table 1. Baseline demographic characteristics of the participants (N=1,600)

Variable	Survey I (N=1,000) Mean ± SD, N (%)	Survey II (N=600) Mean ± SD, N (%)
Male, n (%)	500 (50.0%)	300 (50.0%)
Age (years)	45.8 ± 13.4	49.1 ± 16.5
Measures		
Patient Health Questionnaire-9 (PHQ-9)	5.6 ± 5.9	4.08 ± 4.9
PHQ-9 ≥ 9	241 (24.1%)	100 (16.7%)
Generalized Anxiety Disorder-7 (GAD-7)	2.9 ± 4.1	3.2 ± 4.1
GAD-7 ≥ 5	241 (24.1%)	157 (26.2.0%)
PTSD Checklist for DSM-5 (PCL-5)	23.5 ± 18.3	22.8 ± 17.6
PCL-5 ≥ 28	392 (39.2%)	221 (36.8%)
Alcohol Use Disorder Identification Test (AUDIT)	6.5 ± 7.3	6.5 ± 7.2
AUDIT ≥ 14 in men / 10 in women	189 (18.9%)	122 (20.3%)
Panic Disorder Severity Scale (PDSS)	1.3 ± 3.5	1.7 ± 2.9
PDSS ≥ 7	105 (10.5%)	47 (7.8%)
Depression Symptom Inventory - Suicidality Subscale (DSI-SS)	1.0 ± 1.9	1.0 ± 2.0
DSI-SS ≥ 4	154 (15.4%)	110 (18.3%)
Number of thresholds met		
No clinical cut-offs met	436 (43.6%)	274 (45.7%)
1 cut-off met	232 (23.2%)	136 (22.7%)
2 cut-offs met	130 (13.0%)	71 (11.8%)
3 cut-offs met	75 (7.5%)	41 (6.8%)
4 cut-offs met	55 (5.5%)	46 (7.7%)
5 cut-offs met	47 (4.7%)	20 (3.3%)
6 cut-offs met	25 (2.5%)	12 (2.0%)

PTSD, Post-Traumatic Stress Disorder; DSM, Diagnostic and Statistical Manual of Mental Disorders