

Original Paper

Comprehensive Symptom Prediction in Inpatients With Acute Psychiatric Disorders Using Wearable-Based Deep Learning Models: Development and Validation Study

Minseok Hong^{1,2}, MD; Ri-Ra Kang³, BS; Jeong Hun Yang^{2,4}, MD, PhD; Sang Jin Rhee¹, MD, PhD; Hyunju Lee¹, MD, PhD; Yong-gyom Kim³, BS; KangYoon Lee^{3,5}, PhD; HongGi Kim⁶, BS; Yu Sang Lee⁷, MD, PhD; Tak Youn^{8,9}, MD, PhD; Se Hyun Kim^{1,2}, MD, PhD; Yong Min Ahn^{1,2,10}, MD, PhD

¹Department of Neuropsychiatry, Seoul National University Hospital, Seoul, Republic of Korea

²Department of Psychiatry, Seoul National University College of Medicine, Seoul, Republic of Korea

³Department of IT Convergence Engineering, Gachon University, Seongnam-si, Republic of Korea

⁴Department of Psychiatry, Chungnam National University Sejong Hospital, Sejong, Republic of Korea

⁵Department of Computer Engineering, Gachon University, Seongnam-si, Republic of Korea

⁶Healthconnect Co. Ltd., Seoul, Republic of Korea

⁷Department of Psychiatry, Yong-In Mental Hospital, Yongin-si, Republic of Korea

⁸Department of Psychiatry and Electroconvulsive Therapy Center, Dongguk University International Hospital, Goyang-si, Republic of Korea

⁹Institute of Buddhism and Medicine, Dongguk University, Seoul, Republic of Korea

¹⁰Institute of Human Behavioral Medicine, Seoul National University Medical Research Center, Seoul, Republic of Korea

Corresponding Author:

Yong Min Ahn, MD, PhD

Department of Neuropsychiatry

Seoul National University Hospital

101 Daehak-ro

Jongno-Gu

Seoul, 03080

Republic of Korea

Phone: 82 2 2072 2450

Fax: 82 2 766 2450

Email: aym@snu.ac.kr

Abstract

Background: Assessing the complex and multifaceted symptoms of patients with acute psychiatric disorders proves to be significantly challenging for clinicians. Moreover, the staff in acute psychiatric wards face high work intensity and risk of burnout, yet research on the introduction of digital technologies in this field remains limited. The combination of continuous and objective wearable sensor data acquired from patients with deep learning techniques holds the potential to overcome the limitations of traditional psychiatric assessments and support clinical decision-making.

Objective: This study aimed to develop and validate wearable-based deep learning models to comprehensively predict patient symptoms across various acute psychiatric wards in South Korea.

Methods: Participants diagnosed with schizophrenia and mood disorders were recruited from 4 wards across 3 hospitals and prospectively observed using wrist-worn wearable devices during their admission period. Trained raters conducted periodic clinical assessments using the Brief Psychiatric Rating Scale, Hamilton Anxiety Rating Scale, Montgomery-Asberg Depression Rating Scale, and Young Mania Rating Scale. Wearable devices collected patients' heart rate, accelerometer, and location data. Deep learning models were developed to predict psychiatric symptoms using 2 distinct approaches: single symptoms individually (Single) and multiple symptoms simultaneously via multitask learning (Multi). These models further addressed 2 problems: within-subject relative changes (Deterioration) and between-subject absolute severity (Score). Four configurations were consequently developed for each scale: Single-Deterioration, Single-Score, Multi-Deterioration, and Multi-Score. Data of participants recruited before May 1, 2024, underwent cross-validation, and the resulting fine-tuned models were then externally validated using data from the remaining participants.

Results: Of the 244 enrolled participants, 191 (78.3%; 3954 person-days) were included in the final analysis after applying the exclusion criteria. The demographic and clinical characteristics of participants, as well as the distribution of sensor data, showed considerable variations across wards and hospitals. Data of 139 participants were used for cross-validation, while data of 52 participants were used for external validation. The Single-Deterioration and Multi-Deterioration models achieved similar overall accuracy values of 0.75 in cross-validation and 0.73 in external validation. The Single-Score and Multi-Score models attained overall R^2 values of 0.78 and 0.83 in cross-validation and 0.66 and 0.74 in external validation, respectively, with the Multi-Score model demonstrating superior performance.

Conclusions: Deep learning models based on wearable sensor data effectively classified symptom deterioration and predicted symptom severity in participants in acute psychiatric wards. Despite lower computational costs, Multi models demonstrated equivalent or superior performance than Single models, suggesting that multitask learning is a promising approach for comprehensive symptom prediction. However, significant variations were observed across wards, which presents a key challenge for developing clinical decision support systems in acute psychiatric wards. Future studies may benefit from recurring local validation or federated learning to address generalizability issues.

(*J Med Internet Res* 2024;26:e65994) doi: [10.2196/65994](https://doi.org/10.2196/65994)

KEYWORDS

digital phenotype; mental health monitoring; smart hospital; clinical decision support system; multitask learning; wearable sensor; local validation; mental health facility; deep learning

Introduction

Assessing psychiatric symptoms of patients with acute psychiatric disorders remains challenging [1-3]. Patients with psychosis, mania, and severe depression who require hospitalization often struggle to accurately report their symptoms. Poor insight, a short attention span, impaired cognition, paranoia, and severe avolition can result in underreporting, minimization, and impairment of the verbalization process [3,4]. Additionally, the inherent subjectivity in assessing psychopathology presents reliability concerns and requires experienced professionals [5,6]. Even well-staffed and equipped wards cannot provide continuous patient observation by trained medical personnel.

Recent advancements in wearable sensor technology and artificial intelligence (AI) have enabled the collection and analysis of vast amounts of data [7-9]. These data encompass characteristics related to human behavior, cognition, and mood associated with mental disorders [10] and are often referred to as digital phenotyping [11-13]. The continuous and objective measurement capabilities of sensor data can overcome the limitations of traditional psychiatric assessments [14-16]. Numerous studies have reported that sensor data from wearables and mobile phones can significantly predict health outcomes, such as disease risk and mortality, by unobtrusively monitoring sleep [17-20], circadian rhythms [21-23], and physical activity [24,25].

Using sensor data in acute psychiatric wards is expected to produce significant advantages. Acute psychiatric wards must always be prepared for the risks of aggression and self-harm, resulting in high staffing demands [26,27]. Health care workers in these settings are repeatedly exposed to such patient behaviors, introducing a significant risk of burnout [28,29]. Additionally, human rights standards and the legislation required for the care of psychiatric inpatients are progressing globally [30]. Consequently, digital technology adoption remains a pressing issue in addressing these challenges in psychiatric wards [31]. Furthermore, psychiatric wards provide a highly

controlled environment, enabling accurate and continuous collection of wearable sensor data and detailed clinical information. This creates suitable conditions for digital phenotyping research [32].

Despite these expected advantages, only a few studies have presented AI models to support medical personnel in the context of acute psychiatric wards. To the best of our knowledge, the task of symptom prediction has not yet been addressed. Psychiatric wards exhibit a wide range of architectural, staffing, and organizational variations [33-35], likely contributing to limited research in these contexts. Furthermore, psychiatric inpatients have diverse diagnoses and a wide symptom range [36,37]. Predicting comprehensive symptoms across multi-institutional pandiagnostic groups is necessary to provide a robust foundation for AI-assisted clinical decision support systems (AI-CDSSs) [38].

Therefore, we aimed to develop and validate a comprehensive symptom prediction model for patients in acute psychiatric wards using sensor data obtained using wearable devices and deep learning-based prediction models. This prospective, longitudinal, observational study was conducted across multiple institutions, representing various regions and hospital types. We constructed and evaluated models for predicting single symptoms individually and multiple symptoms simultaneously, as well as for predicting relative changes within subjects and absolute severity between subjects. Moreover, we explored the future directions and challenges of AI-CDSSs in psychiatric ward settings.

Methods

Overview

The study was presented according to the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [39], as well as the TRIPOD+AI guideline (version February 7, 2024) [40].

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki [41]. The study protocol was reviewed and approved by the Institutional Review Board of the Seoul National University Hospital (2210-073-1368). All participants provided informed consent. Participants received compensation of KRW 100,000 (approximately US \$70) for each week of participation. Research data were stored on servers within each hospital, separate from personally identifiable information except for study identification numbers. Access to and analysis of the data were restricted to preapproved researchers only. No identifiable images or personal information of participants are included in this manuscript or any supplementary materials.

Participants

The Sensor Application for Early Response in Closed Wards (SAFER) project aims to build a generalizable wearable-based AI-CDSS for acute psychiatric wards. This project is ongoing until 2024, and interim data were used in this study.

Four wards from 3 hospitals representing diverse regions and hospital types participated in the SAFER study: Seoul National University Hospital, a tertiary general hospital in the capital city with a single participating ward; Yongin Mental Hospital, a suburban psychiatric hospital with 2 wards, one for each sex; and Dongguk University Ilsan Hospital, a general hospital in a newly developed city with one participating ward.

Inpatients aged 13 years or older, diagnosed with a mood disorder (major depressive or bipolar disorder) or schizophrenia spectrum disorder, were recruited from each ward. For vulnerable participants, including minors and individuals whose symptoms might impair their ability to understand and consent, additional consent was obtained from their legal guardians. Exclusion criteria included dementia, intellectual disability, organic brain disorders, and any physical condition causing difficulty in data collection (eg, wrists too thin owing to low body weight of participants with anorexia nervosa). However, other coexisting diagnoses were included to ensure a pandiagnostic study population. As one of the analysis targets was a comparison with previous scores of the same individual, the analysis included only participants who had undergone at least 2 assessments, including the baseline assessment.

Measurements

Each participant was provided with a wrist-worn device called URBAN HR (Partron Co, Ltd). Participants were instructed to wear the device continuously, except when showering or leaving the ward, although they could remove it freely if desired. The wearable devices collected data on heart rate, 3-axis acceleration (which measures acceleration along 3 perpendicular axes in space), and location. Using in-device algorithms, these metrics were used to calculate calories burned, steps walked, distance moved, and sleep index. Among these, the sleep index is used to analyze participants' sleep and is calculated based on heart rate and acceleration data using algorithms provided by the manufacturer. This sleep index quantifies the cumulative daily sleep duration, converting it into a score ranging from 30 (2.25 h or less) to 100 (7.5 h or more). The generated data were

transmitted in real time to the server at each hospital via Bluetooth gateways installed in each patient's room and common areas

The participants were assessed weekly for symptoms by raters using clinician-rated scales. Additional assessments were conducted following injection, seclusion, and restraint interventions: the Brief Psychiatric Rating Scale (BPRS) was used to evaluate psychotic and general psychiatric symptoms, the Hamilton Anxiety Scale (HAM-A) was used to assess anxiety, the Montgomery-Asberg Depression Rating Scale (MADRS) was used to evaluate depressive symptoms, and the Young Mania Rating Scale (YMRS) was used to measure manic symptoms. The raters were registered nurses with a minimum of 51 months and an average of 11 years of clinical experience. All raters completed 40 hours of specialized training for the study, and their ratings were periodically reviewed at research conferences with psychiatrists.

Feature Engineering

A 1-hour sliding window was used to synthesize time-series values from various sensors. The 3-axis accelerations were converted to a total acceleration, which is the magnitude of the summed acceleration. The first value, last value, mean, median, maximum, minimum, SD, and number of unique values within the 1-hour sliding window were calculated for total acceleration and heart rate. Calories burned, sleep index, number of steps walked, and distance traveled were calculated as cumulative values over this window. Regarding location data, the variance of the coordinate values and most frequent semantic location (room, hallway, other, and no signal) were computed over the sliding window.

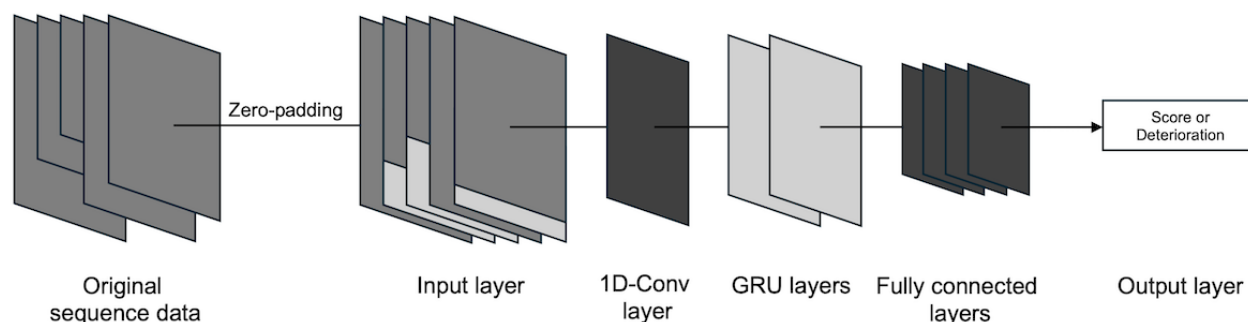
Location entropy quantifies the diversity and unpredictability of a participant's location patterns, which have demonstrated predictive value for mental health correlates [42-44]. We computed entropy values over 24-hour and 8-hour periods, delineated by midnight, 8 AM, and 4 PM. This approach was chosen as meaningful location entropy assessment requires longer periods than the 1-hour sliding window.

Missing values in the sensor data were imputed using the *softImpute* method [45]. The participant was excluded if missing values exceeded 50% of their total timespan. Sensor data from the 4 weeks preceding each symptom assessment were used in the model. In cases where available data preceding the assessment were less than 4 weeks, zero-padding was applied to extend the data to the required length [46]. The size of the input layer was 672 hours with 31 features (672×31) including 2 nonsensor features (age and sex). t-distributed Stochastic Neighbor Embedding (t-SNE) was used to visualize the distribution of these 31 features by ward.

Prediction Model and Performance Evaluation

A model designed to predict symptoms encompassing the BPRS, HAM-A, MADRS, and YMRS was constructed using a deep learning architecture that incorporates 1D convolutional layers and gated recurrent units (Figure 1). All models shared an identical structure, except for the terminal fully connected layer and output layer.

Figure 1. The architecture of deep learning models for predicting comprehensive symptoms in inpatients with acute psychiatric disorders. Original sequence data were adjusted to match the required length by removing data older than 4 weeks; shorter data were zero-padded to form the input layer. 1D-Conv, GRU, and fully connected layers were used sequentially. The model's performance was compared based on various output layer configurations. 1D-Conv: 1D convolutional; GRU: gated recurrent unit.



These models were differentiated based on the configuration of the output layer according to 2 criteria. First, the models were classified based on whether they predicted a single scale at a time (Single) or predicted all scales simultaneously (Multi). Multi models used a multitask learning (MTL) approach [47]. Second, the models were distinguished based on whether they classified instances where the symptom scale score increased from the previous assessment within an individual (Deterioration) or predicted the absolute severity score of the symptoms (Score). Deterioration models corresponded to a within-subject design, while Score models aligned with a between-subject design [48]. Consequently, 4 types of models were developed for each scale by intersecting these 2 criteria: Single-Deterioration, Single-Score, Multi-Deterioration, and Multi-Score.

Participants recruited prior to May 1, 2024, were allocated to the internal cross-validation set, while those recruited on or after May 1, 2024, were assigned to the external validation set. The internal cross-validation set underwent a 5-fold cross-validation. Hyperparameters were optimized using the random search method [49]. The final model was evaluated on the external validation set. The performance of Deterioration models was evaluated using accuracy, area under the curve (AUC), and receiver operating characteristic (ROC) curve. In the cross-validation, the ROC curve was generated by vertically pooling the curves from each fold. The performance of Score models was assessed using R^2 and normalized root mean squared error (NRMSE), with NRMSE chosen to account for varying score ranges across scales. Permutation feature importance was measured for features in the external validation set. The deep

learning model was built using *PyTorch* (version 2.2) [50], and model performance evaluation and t-SNE were performed using *scikit-learn* (version 1.3) [51].

Results

Overall, 244 participants were enrolled in the SAFER study between May 26, 2023, and August 5, 2024. Sixteen participants were excluded based on the exclusion criteria or consent withdrawal after initially agreeing to participate. Additionally, 37 participants were excluded from the analysis because they lacked at least 1 assessment or had > 50% missing data from wearable sensors. Ultimately, data from 191 participants, encompassing 3954 person-days of observation, were analyzed. Participants were observed for a mean of 20.7 (SD 17.5) days, with 4.3 (SD 2.9) assessments per participant. In the final sample, 57.6% (110/191) were female, and the most frequent diagnosis was a mood disorder with a current depressive episode (93/191, 48.7%). For validation purposes, 139 participants were assigned to the cross-validation set and 52 participants were assigned to the external validation set.

Significant differences were observed in sex, age, and household income across hospitals (all $P < .001$). The first psychiatric admission ($P = .17$) did not show any significant differences nor did BPRS ($P = .45$) or HAM-A ($P = .27$) scores, suggesting general clinical severity was similar. However, significant differences were observed in the number of assessments ($P = .02$), days of observation ($P = .007$), diagnostic groups ($P = .005$), and MADRS ($P = .002$) and YMRS ($P < .001$) scores across the hospitals (Table 1).

Table 1. Basic characteristics and statistical comparison of participants by hospital. The Kruskal-Wallis test was performed for continuous variables, and the chi-square test was performed for categorical variables.

	Total (N=191)	Hospital 1 ^a (n=73)	Hospital 2 ^b (n=88)	Hospital 3 ^c (n=30)	Statistics		P value
					Chi-square (<i>df</i>)	Kruskal-Wallis test (<i>H</i>)	
Sex, n (%)					19.6 (2)	— ^d	<.001
Male	81 (42.4)	17 (23.3)	51 (58.0)	13 (43.3)			
Female	110 (57.6)	56 (76.7)	37 (42.0)	17 (56.7)			
Age (years)					—	21.5	<.001
Mean (SD)	32.6 (13.8)	27.2 (10.7)	35.0 (13.5)	38.7 (17.1)			
Median (IQR)	27 (21-40)	24 (19-29)	30 (22-45)	32 (24-52)			
Diagnostic group, n (%)					14.8 (4)	—	.005
Schizophrenia spectrum	72 (37.7)	19 (26.0)	34 (38.6)	19 (63.3)			
Mood-depressive ^e	93 (48.7)	44 (60.3)	39 (44.3)	10 (33.3)			
Mood-manic ^f	26 (13.6)	10 (13.7)	15 (17.0)	1 (3.3)			
Household income (KRW 5 million=US \$3600), n (%)					24.1 (2)	—	<.001
>KRW 5 million	78 (40.8)	32 (43.8)	23 (26.1)	23 (76.7)			
≤KRW 5 million	113 (59.2)	41 (56.2)	65 (73.9)	7 (23.3)			
Education level, n (%)					8.5 (4)	—	.07
≥College	48 (25.1)	16 (21.9)	25 (28.4)	7 (23.3)			
High school	118 (61.8)	41 (56.2)	57 (64.8)	20 (66.7)			
≤Middle school	25 (13.1)	16 (21.9)	6 (6.8)	3 (10.0)			
Number of assessments					—	7.7	.02
Mean (SD)	4.3 (2.9)	4.1 (2.8)	4.6 (3.1)	3.5 (2.8)			
Observational days					—	9.9	.007
Mean (SD)	20.7 (17.5)	19.4 (14.5)	23.4 (19.3)	15.8 (17.5)			
Baseline symptom score, mean (SD)							
BPRS ^g	21.5 (9.1)	22.8 (10.9)	20.6 (8.1)	21.2 (6.7)	—	1.6	.45
HAM-A ^h	11.2 (6.4)	10.2 (4.6)	12.0 (8.2)	11.5 (3.3)	—	2.6	.27
MADRS ⁱ	18.8 (10.9)	22.0 (9.7)	16.8 (12.4)	16.7 (6.2)	—	12.2	.002
YMRS ^j	11.1 (8.3)	9.8 (8.5)	10.7 (8.4)	15.5 (6.1)	—	17.5	<.001
First psychiatric admission, n (%)					3.5 (2)	—	.17
Yes	66 (34.6)	26 (35.6)	34 (38.6)	6 (20.0)			
No	125 (65.4)	47 (64.4)	54 (61.4)	24 (80.0)			

^aHospital 1: Seoul National University Hospital.

^bHospital 2: Yongin Mental Hospital.

^cHospital 3: Dongguk University Ilsan Hospital.

^dNot applicable.

^eMood-depressive: Mood disorder with current depressive episode.

^fMood-manic: Mood disorder with current manic episode.

^gBPRS: Brief Psychiatric Rating Scale.

^hHAM-A: Hamilton Anxiety Rating Scale.

ⁱMADRS: Montgomery-Asberg Depression Rating Scale.

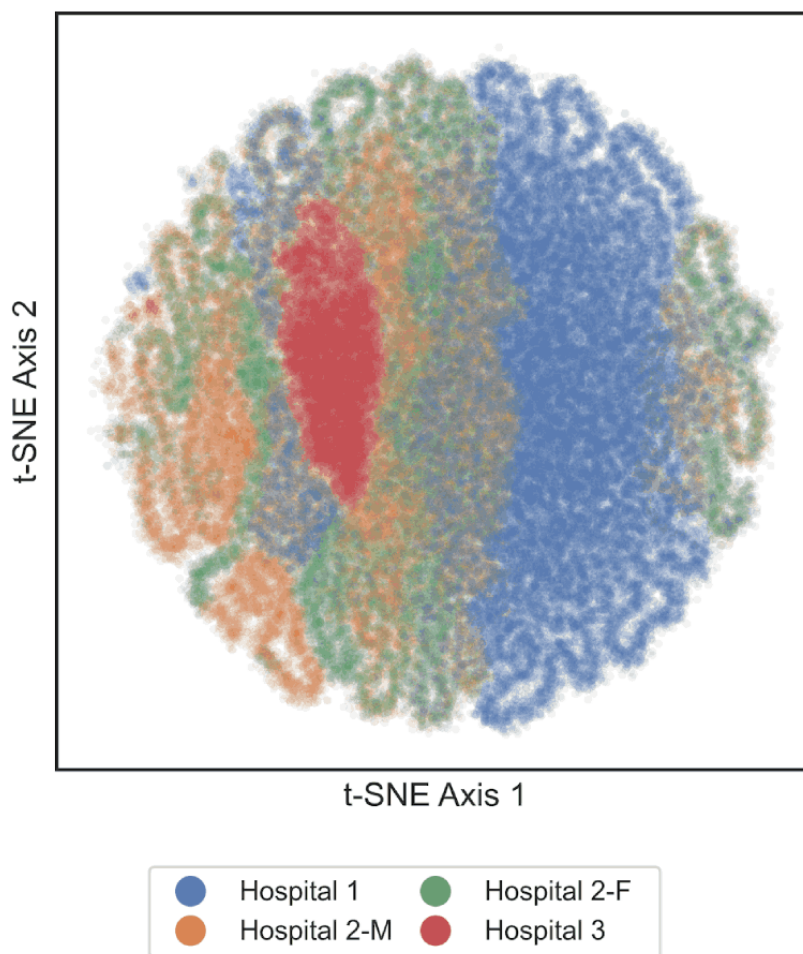
^jYMRS: Young Mania Rating Scale.

The group-level mean for the scale scores were 15.9 (SD 9.5) for the BPRS, 8.8 (SD 6.0) for the HAM-A, 14.1 (SD 10.5) for the MADRS, and 8.1 (SD 7.1) for the YMRS, with no significant differences between the cross-validation and external validation sets (all $P > .05$; Table S1 in [Multimedia Appendix 1](#)). The deterioration of symptom cases, where the assessment score was increased to that of the previous assessment for an individual, were 22.9% (143/624) for the BPRS, 26.1% (163/624) for the HAM-A, 28.8% (180/624) for the MADRS,

and 28.2% (176/624) for the YMRS; no significant differences were observed between the cross-validation and external validation sets (all $P > .05$; Table S2 in [Multimedia Appendix 1](#)).

The distribution of the multidimensional sensor data visualized using t-SNE revealed several clusters. Many clusters comprised data originating from 1 or 2 wards. Clusters with overlapping data from all wards were rare ([Figure 2](#)).

Figure 2. Visualization of the distribution of sensor data using t-SNE. Each point is color-coded to differentiate between hospital wards—hospital 1: Seoul National University Hospital; hospital 2: male (M) and female (F) wards in Yongin Mental Hospital; and hospital 3: Dongguk University Ilsan Hospital. t-SNE: t-distributed stochastic neighbor embedding.



In the Deterioration models, the Single-Deterioration and Multi-Deterioration models showed similar overall performance. Cross-validation revealed that both models achieved an accuracy of 0.75, with an AUC of 0.74 and 0.76, respectively. External validation revealed an accuracy of 0.73 for both models, with an AUC of 0.71 and 0.74, respectively. Among the scales, the BPRS demonstrated the highest accuracy across all models and sets, with accuracy ranging from 0.75 to 0.79 and AUC from 0.76 to 0.82. The MADRS performed the least accurately in cross-validation and the YMRS performed the least accurately

in external validation, regardless of model type ([Table 2](#)). The ROC curve for the BPRS in cross-validation showed no clear superiority between Single-Deterioration and Multi-Deterioration models within 1 SD. External validation indicated that the 2 models' curves intersected at both ends of the false positive rate range, with the Multi-Deterioration model generally outperforming in the central portion (0.1 to 0.6; [Figure 3](#)). Similar patterns were observed for other scales (Figures S1-3 in [Multimedia Appendix 1](#)).

Table 2. Performance of the Deterioration models measured by accuracy and area under the curve (AUC). The Deterioration models predict whether scale scores increased compared with the previous assessment.

	Cross-validation, mean (SD)		External validation	
	Single	Multi	Single	Multi
Accuracy				
Overall	0.75 (0.02)	0.75 (0.03)	0.73	0.73
BPRS ^a	0.79 (0.02)	0.79 (0.03)	0.79	0.75
HAM-A ^b	0.76 (0.05)	0.75 (0.05)	0.76	0.75
MADRS ^c	0.72 (0.03)	0.73 (0.05)	0.74	0.75
YMRS ^d	0.73 (0.04)	0.74 (0.02)	0.65	0.69
AUC				
Overall	0.74 (0.04)	0.76 (0.04)	0.71	0.74
BPRS	0.79 (0.07)	0.80 (0.06)	0.76	0.82
HAM-A	0.75 (0.06)	0.77 (0.05)	0.73	0.75
MADRS	0.69 (0.03)	0.72 (0.05)	0.69	0.73
YMRS	0.73 (0.05)	0.76 (0.04)	0.64	0.67

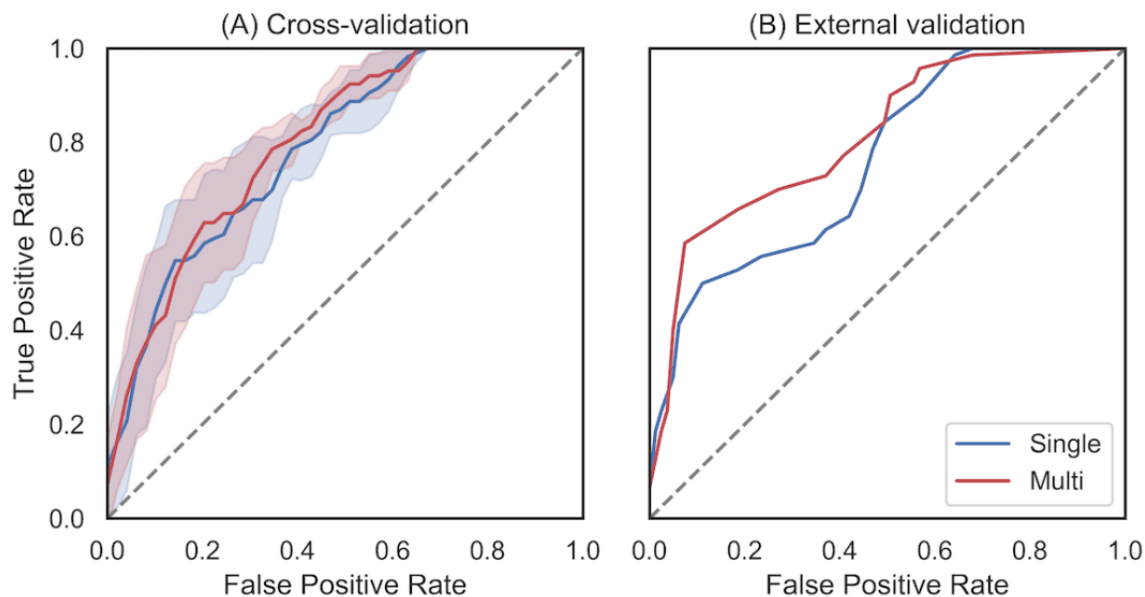
^aBPRS: Brief Psychiatric Rating Scale.

^bHAM-A: Hamilton Anxiety Rating Scale.

^cMADRS: Montgomery-Asberg Depression Rating Scale.

^dYMRS: Young Mania Rating Scale.

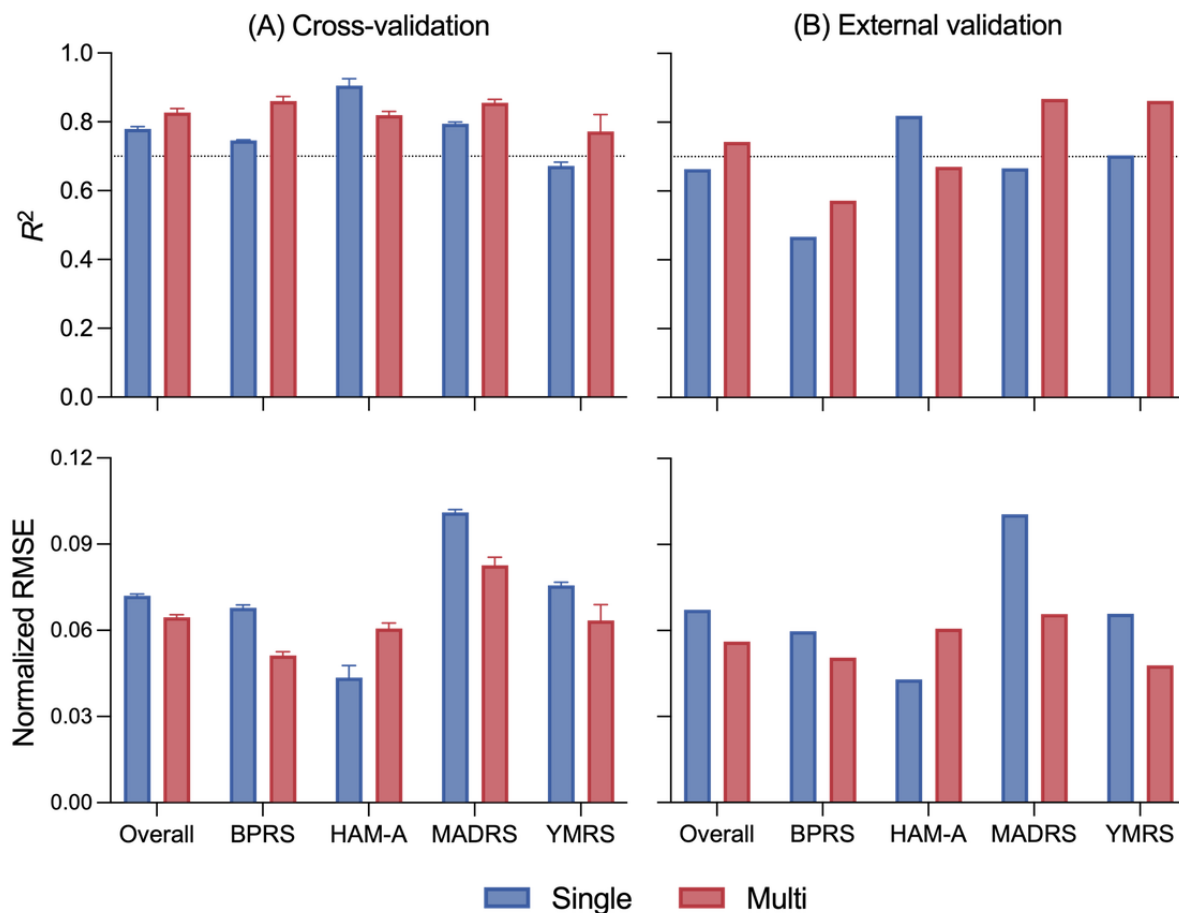
Figure 3. Receiver operating characteristic curve of the Deterioration models with respect to BPRS scores: (A) cross-validation and (B) external validation. The Deterioration models predict whether BPRS scores increased compared with the previous assessment. Colored areas in cross-validation represent the range of 1 SD. BPRS: Brief Psychiatric Rating Scale.



The Multi-Score model outperformed the Single-Score model in both R^2 and NRMSE. In the cross-validation set, the Single-Score and Multi-Score models achieved R^2 values of 0.78 and 0.83 and NRMSE values of 0.072 and 0.064, respectively. External validation showed that the Single-Score and Multi-Score models achieved R^2 values of 0.66 and 0.74

and NRMSE values of 0.067 and 0.056, respectively. Among the scales, the BPRS, which was predicted well in the Deterioration models, demonstrated the lowest R^2 values during external validation: 0.47 for Single-Score and 0.57 for Multi-Score. In contrast, the MADRS and YMRS exhibited R^2 values exceeding 0.85 in the external validation of the Multi-Score model (Figure 4).

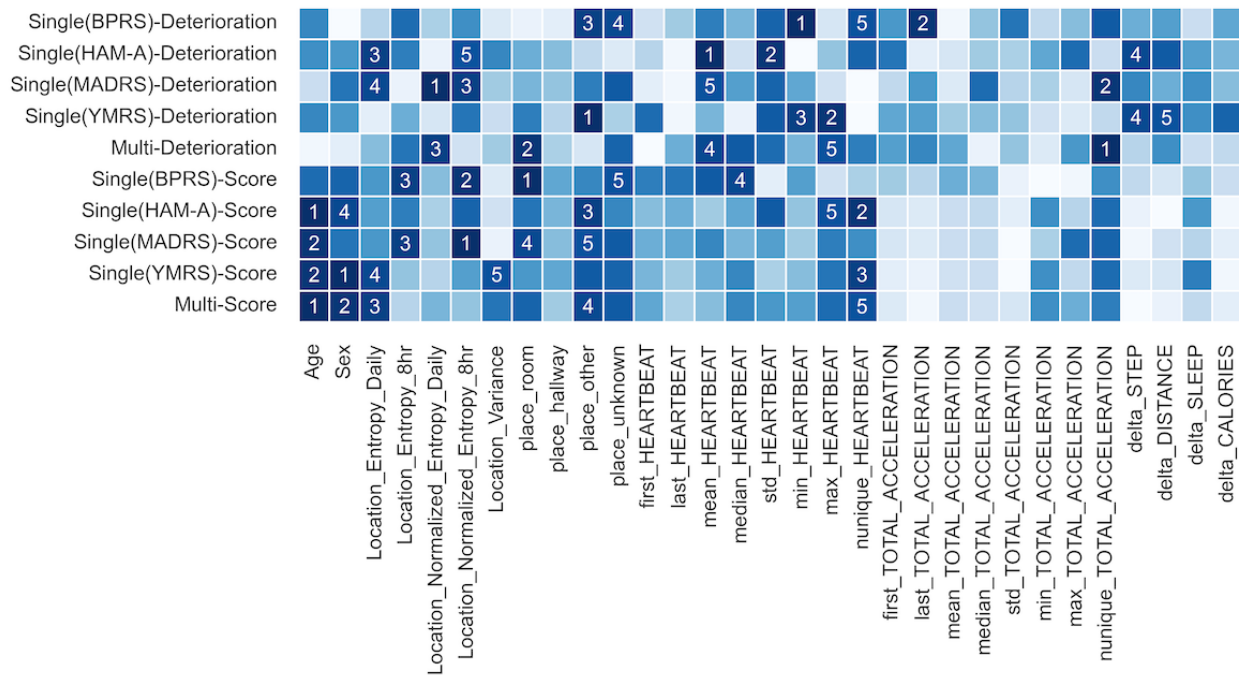
Figure 4. Performance of the Score model measured by R2 and NRMSE: (A) cross-validation and (B) external validation. The Score models predict scale scores; the error bar represents an interval of 1 SD. The dotted line indicates an R2 value of 0.7. BPRS: Brief Psychiatric Rating Scale; HAM-A: Hamilton Anxiety Rating Scale; MADRS: Montgomery-Asberg Depression Rating Scale; Multi: model predicting multiple symptoms simultaneously; NRMSE: normalized root mean squared error; Single: model predicting single symptoms individually; YMRS: Young Mania Rating Scale.



The permutation feature importance analysis indicated that time spent in places other than rooms or hallways (*place_other*, 5 models), the number of unique heart rate values (*unique_HEARTBEAT*, 4 models), and 2 location entropy indicators (*Location_Entropy_Daily* and *Location_Normalized_Entropy_8hr*, 4 models each) were among

the top 5 important features in 4 or more models. Notably, no single feature consistently ranked in the top 5 across more than half of the models. Nonsensor features such as age and sex did not rank in the top 5 important features for the Deterioration models but were frequently included in the Score models (Figure 5).

Figure 5. Permutation feature importance in external validation. The horizontal axis represents 31 individual features, and the vertical axis represents the prediction models. To ensure comparability, the importance values are converted to ranks, with the top 5 annotated. BPRS: Brief Psychiatric Rating Scale; HAM-A: Hamilton Anxiety Rating Scale; MADRS: Montgomery-Asberg Depression Rating Scale; Multi: model predicting multiple symptoms simultaneously; Single: model predicting single symptoms individually; YMRS: Young Mania Rating Scale.



Discussion

Principal Findings

We conducted a longitudinal, prospective, multicenter study of inpatients using wearable devices to develop and validate deep learning models for predicting comprehensive symptoms. To our knowledge, this is the first study to predict clinical rating scale scores of pandiagnostic cases in acute psychiatric wards using wearable devices and deep learning models. While direct benchmarks are unavailable owing to our inpatient setting, we can compare our results to those of similar studies. A meta-analysis of depression prediction studies using wearable devices reported a pooled mean accuracy ranging from 0.70 to 0.89 [52]. Similarly, a meta-analysis of anxiety prediction studies using wearable devices reported a 95% CI for pooled mean accuracy of 0.71 to 0.89 [53]. Our Single-Deterioration and Multi-Deterioration models for MADRS, HAM-A, and overall performance fall within these ranges, demonstrating comparable performance. The Single-Score and Multi-Score models showed overall R^2 values of approximately 0.8 in cross-validation and 0.7 in external validation, indicating substantial explanatory power. In summary, our wearable-based deep learning models effectively predict comprehensive psychiatric symptoms.

Single Versus Multi Models

Our Multi models used an MTL framework to simultaneously predict 4 different scale scores. While Single models might be expected to outperform Multi models, given that they require approximately 4 times the computational resources and parameters to learn and solve the same tasks [47], our findings suggest otherwise. The Single-Deterioration and

Multi-Deterioration models, which predict changes from previous assessments for an individual, showed no material differences in both accuracy and AUC. Moreover, Multi-Deterioration models demonstrated a superior balance between sensitivity and specificity, as the AUC value and ROC curve indicated. Considering the class imbalance in symptom deterioration cases, this suggests that MTL-based models provide a more balanced predictive performance. Furthermore, the Single-Score and Multi-Score models, which predict absolute symptom severity, showed that the Multi-Score model demonstrated better explanatory power and lower errors in both cross-validation and external validation.

This study is not the first to attempt to predict mental health indicators using MTL-based models. Several studies have reported significant success using this approach with data from social media [54], electrocardiography [55], functional magnetic resonance imaging [56], and so on. Interestingly, while Harvey et al [57] argued that thousands of samples might be necessary for MTL to show benefits over single-task learning in predicting psychiatric diagnoses using functional magnetic resonance imaging, our study demonstrated these benefits with fewer samples. Although the comparability is limited due to the difference between cross-sectional brain imaging and longitudinal wearable sensor data, a notable distinction is that they predicted diagnoses, whereas we predicted symptoms.

Many contemporary psychopathologists agree that individual psychiatric symptoms form a complex network [58,59]. From this perspective, symptoms are not merely passive manifestations of specific diagnoses but active entities influencing other symptoms [60]. Thus, information learned in predicting one symptom can inform the prediction of others. In

contrast, psychiatric diagnoses are defined through operational criteria of symptom constellations [61]. Consequently, learning shared latent information for multiple symptom prediction may be more feasible and may require fewer samples than those required for predicting multiple diagnoses. A similar study predicting symptoms in individuals with schizophrenia using smartphone passive data found benefits from MTL with only 61 samples [19]. This was likely possible because the prediction targets were individual symptoms rather than diagnoses. Although there are not yet many studies directly comparable with the present research, our study findings suggest that MTL is an effective framework for psychiatric symptom prediction tasks.

Deterioration Versus Score Models

The Deterioration and Score models exhibit several notable differences. The purpose of the Deterioration models was to detect symptom changes within individuals compared with their previous state, whereas the Score models aimed to predict the absolute symptom severity between individuals. For feature importance, no shared patterns of important features were observed between the Deterioration and Score models predicting the same symptom scale. For instance, age and sex emerged as important in most Score models but not in Deterioration models. This is likely because predicting absolute symptom severity relies more on basic demographic information as crucial calibration data compared with predicting relative changes. Performance patterns during external validation also varied across scales. The BPRS demonstrated the best performance among all scales in both cross-validation and external validation for the Deterioration models, but the Score models showed a considerable R^2 reduction in external validation. Conversely, the YMRS exhibited the most significant performance decline in external validation for the Deterioration models but not for the Score models. This indicates that the Deterioration and Score models, even when predicting the same symptom scale, differ in their required information and generalization challenges. This serves as another example demonstrating the distinction between within-subject and between-subject designs [62]. Future studies on predicting symptoms in acute psychiatric inpatients should consider treating relative change and absolute severity as distinct analyses.

Challenges to Generalizability in AI-CDSSs for Acute Psychiatric Wards

The generalizability of models is a critical issue in developing AI models for predicting mental health indicators [63,64]. SAFER's ultimate goal was to develop a predictive model that can serve as the foundation for an AI-CDSS in acute psychiatric wards with broad generalizability. In this study, external validation based on time point maintained considerable performance compared with internal cross-validation, demonstrating the model's temporal robustness at least for the study period. However, recent studies have pointed out that external validation alone cannot fully guarantee reliability and generalizability in clinical prediction models [65]. In this study, challenges to generalizability remain in 2 aspects.

One challenge is interward variations. In this study, participant characteristics and sensor data distributions showed considerable

differences across institutions. This was partly inevitable owing to the study's aim to include diverse regions and hospital types. Nevertheless, the observed substantial interward variation suggests that applying an AI-CDSS based on this study to any other facility may encounter significant generalization difficulties. Studies on other clinical prediction models have noted that even models developed from large international multisite clinical trials often see their performance drop to chance levels when subjected to completely independent tests [66]. A potential solution to this issue is a method called recurring local validation or targeted validation [67,68]. This method involves validation and adjustment of prediction models to fit the specific settings, where they will be deployed. Future research should explore whether such ward-specific models could replace a single universal model.

The other challenge is the issue of selection bias [69]. In this study, we recruited pandiagnostic cases by minimizing the exclusion criteria based on clinical conditions and accepting various comorbidities. One reason for needing such an inclusive sample is that the end users of an AI-CDSS in acute psychiatric wards are primarily the ward's medical staff [31]; thus, its usability can only be ensured if it can be applied with minimal restrictions to patients in the ward. However, participants in this study were recruited based on informed consent. Even with minimized exclusion criteria, patients with acute psychiatric symptoms may be subject to selection bias owing to differences in their ability to consent caused by their symptoms, making it difficult to recruit participants in a completely inclusive manner, like those with stroke or Alzheimer disease [70,71]. Therefore, in the future deployment of this model, validation based on data collected close to usual care, similar to the concept of pragmatic trials, will be necessary [72]. Waivers of the requirement for informed consent could be helpful in this case, but they require various ethical and regulatory considerations [73]. One of these is that the collected sensor data contains sensitive personal information, which makes it difficult to analyze through data centralization in one place. A promising solution to this could be engineering techniques such as federated learning, which allows for learning and validation without exporting raw data outside the institution [74].

Limitations

First, although external validation was performed, it was conducted by dividing participants based on recruitment periods rather than by institutions. As confirmed in this study, acute psychiatric facilities varied significantly from one another, resulting in inherent generalizability issues. Thus, while the significance of the external validation is somewhat limited, we discussed various challenges and potential solutions to improve the model's generalizability. Second, this study was conducted in acute psychiatric wards within the context of South Korea's specific ethnic, sociocultural, legislative, and reimbursement systems. These contextual factors surrounding mental health facilities differ considerably between countries [30,75]. Therefore, developing a comprehensive symptom prediction model based on this study in other countries would require substantial consideration of these factors. Third, the average duration of the study was <3 weeks, which is relatively short. Therefore, patients with shorter treatment durations may have

been overrepresented, while relatively few participants provided data over a sufficiently long observation period. Fourth, the sample size was insufficient to fully cover the objectives of this study, which aimed at a diverse population across various settings. In particular, the number of participants older than 40 years or with low educational attainment was small, and the usefulness of our clinical prediction model may be reduced for participants with these demographic characteristics. Fifth, different raters were responsible for clinical assessments at each hospital. Efforts were made to improve measurement reliability through specialized training courses and periodic research conferences; however, interrater errors were not statistically measured or adjusted.

Conclusions

We developed and validated a model to predict multidimensional symptoms in various acute psychiatric wards using wearable sensor data and deep learning models. The constructed model proved effective in predicting a single symptom individually and multiple symptoms simultaneously. Notably, a model that predicted multiple symptoms simultaneously demonstrated more balanced within-subject classification and better between-subject symptom severity prediction. Substantial interward variations were also found in this study, suggesting that generalizability is a key issue. By discussing challenges and solutions for generalizability, we have contemplated the future direction of AI-CDSSs in acute psychiatric wards.

Acknowledgments

The authors appreciate the technical assistance provided by Dr Yunsun Kim at Hallym University Kangnam Sacred Heart Hospital in preparing the figures for this manuscript. This work was supported by the Mental Health Research and Development Project of the Korea Health Industry Development Institute (grant RS-2022-KH129879). The funding body had no role in the study design, data collection, analysis, interpretation, writing of the report, or decision to submit the article for publication. During the preparation of this work, the authors used ChatGPT 4o (OpenAI) [76] to enhance their English style. After using this tool/service, the authors reviewed and edited the content as needed. The authors assume full responsibility for the content of the published article.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to confidentiality assurances given during the informed consent process, as the study involves sensitive personal data from vulnerable participants. The analysis code is publicly [77].

Authors' Contributions

MH contributed to the conceptualization, methodology, investigation, drafting of the original manuscript, and visualization. RRK was responsible for validation, formal analysis, and manuscript review and editing. JHY assisted with methodology and reviewed and edited the manuscript. SJR contributed to the validation, investigation, and review and editing of the manuscript. HL participated in the investigation and manuscript review and editing. YGK handled formal analysis and reviewed and edited the manuscript. KL provided resources and contributed to manuscript review and editing. HK was responsible for software, resources, data curation, and manuscript review and editing. YSL, TY, and SHK participated in the investigation and manuscript review and editing, with SHK also providing supervision. YMA was involved in conceptualization, supervision, project administration, funding acquisition, and manuscript review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary materials.

[DOCX File , 594 KB-Multimedia Appendix 1]

References

1. Liu F, Gao M, Wu Q, Yan M, Wu R, Shao P, et al. Diagnostic stability in psychiatric patients from hospital admission to discharge: a 10-year retrospective study. *Psychiatry Investig*. May 2023;20(5):461-470. [doi: [10.30773/pi.2022.0349](https://doi.org/10.30773/pi.2022.0349)] [Medline: [37253472](https://pubmed.ncbi.nlm.nih.gov/37253472/)]
2. Thrasher TW, Rolli M, Redwood RS, Peterson MJ, Schneider J, Maurer L, et al. 'Medical clearance' of patients with acute mental health needs in the emergency department: a literature review and practice recommendations. *WMJ*. Dec 2019;118(4):156-163. [Medline: [31978283](https://pubmed.ncbi.nlm.nih.gov/31978283/)]
3. Ho BC, Flaum M, Hubbard W, Arndt S, Andreasen NC. Validity of symptom assessment in psychotic disorders: information variance across different sources of history. *Schizophr Res*. Jun 01, 2004;68(2-3):299-307. [doi: [10.1016/j.schres.2003.07.006](https://doi.org/10.1016/j.schres.2003.07.006)] [Medline: [15099611](https://pubmed.ncbi.nlm.nih.gov/15099611/)]

4. Carpels A, de Smet L, Desplenter S, de Hert M. Falls among psychiatric inpatients: a systematic review of literature. *Alpha Psychiatry*. Sep 2022;23(5):217-222. [doi: [10.5152/alphapsychiatry.2022.21735](https://doi.org/10.5152/alphapsychiatry.2022.21735)] [Medline: [36426266](https://pubmed.ncbi.nlm.nih.gov/36426266/)]
5. Kang UG. *Psychopathology: A Conceptual Approach*. Seoul, South Korea. SNU Press; Dec 10, 2021.
6. Strauss J. Subjectivity and severe psychiatric disorders. *Schizophr Bull*. Oct 20, 2010;37(1):8-13. [doi: [10.1093/schbul/sbq116](https://doi.org/10.1093/schbul/sbq116)] [Medline: [20961994](https://pubmed.ncbi.nlm.nih.gov/20961994/)]
7. de Arriba-Pérez F, Caeiro-Rodríguez M, Santos-Gago J. Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios. *Sensors (Basel)*. Sep 21, 2016;16(9):1538. [doi: [10.3390/s16091538](https://doi.org/10.3390/s16091538)] [Medline: [27657081](https://pubmed.ncbi.nlm.nih.gov/27657081/)]
8. Chen L, Xia C, Zhao Z, Fu H, Chen Y. AI-driven sensing technology: review. *Sensors (Basel)*. May 07, 2024;24(10):2958. [doi: [10.3390/s24102958](https://doi.org/10.3390/s24102958)] [Medline: [38793814](https://pubmed.ncbi.nlm.nih.gov/38793814/)]
9. Kim D, Min J, Ko SH. Recent developments and future directions of wearable skin biosignal sensors. *Adv Sens Res*. Oct 24, 2023;3(2):2300118. [doi: [10.1002/adsr.202300118](https://doi.org/10.1002/adsr.202300118)]
10. Chia AZR, Zhang MWB. Digital phenotyping in psychiatry: a scoping review. *Technol Health Care*. 2022;30(6):1331-1342. [doi: [10.3233/thc-213648](https://doi.org/10.3233/thc-213648)] [Medline: [35661034](https://pubmed.ncbi.nlm.nih.gov/35661034/)]
11. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. *Nat Biotechnol*. 2015;33(5):462-463. [doi: [10.1038/nbt.3223](https://doi.org/10.1038/nbt.3223)] [Medline: [25965751](https://pubmed.ncbi.nlm.nih.gov/25965751/)]
12. Insel TR. Digital phenotyping: technology for a new science of behavior. *JAMA*. 2017;318(13):1215-1216. This renowned article by Insel suggests the need and potential for digital phenotyping in acute psychiatric wards. [doi: [10.1001/jama.2017.11295](https://doi.org/10.1001/jama.2017.11295)] [Medline: [28973224](https://pubmed.ncbi.nlm.nih.gov/28973224/)]
13. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. *J Med Internet Res*. Dec 13, 2023;25:e46778. [FREE Full text] [doi: [10.2196/46778](https://doi.org/10.2196/46778)] [Medline: [38090800](https://pubmed.ncbi.nlm.nih.gov/38090800/)]
14. Gurrin C, Smeaton AF, Doherty AR. *LifeLogging: Personal Big Data*. Norwell, MA. Now Publishers; 2014.
15. Oudin A, Maatoug R, Bourla A, Ferreri F, Bonnot O, Millet B, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. *J Med Internet Res*. 2023;25:e44502. [FREE Full text] [doi: [10.2196/44502](https://doi.org/10.2196/44502)] [Medline: [37792430](https://pubmed.ncbi.nlm.nih.gov/37792430/)]
16. Hsin H, Fromer M, Peterson B, Walter C, Fleck M, Campbell A, et al. Transforming psychiatry into data-driven medicine with digital measurement tools. *NPJ Digit Med*. 2018;1(1):37. [FREE Full text] [doi: [10.1038/s41746-018-0046-0](https://doi.org/10.1038/s41746-018-0046-0)] [Medline: [31304319](https://pubmed.ncbi.nlm.nih.gov/31304319/)]
17. Niemeijer K, Mestdagh M, Kuppens P. Tracking subjective sleep quality and mood with mobile sensing: multiverse study. *J Med Internet Res*. 2022;24(3):e25643. [FREE Full text] [doi: [10.2196/25643](https://doi.org/10.2196/25643)] [Medline: [35302502](https://pubmed.ncbi.nlm.nih.gov/35302502/)]
18. Sakal C, Li T, Li J, Yang C, Li X. Association between sleep efficiency variability and cognition among older adults: cross-sectional accelerometer study. *JMIR Aging*. 2024;7:e54353. [FREE Full text] [doi: [10.2196/54353](https://doi.org/10.2196/54353)] [Medline: [38596863](https://pubmed.ncbi.nlm.nih.gov/38596863/)]
19. Tseng VWS, Sano A, Ben-Zeev D, Brian R, Campbell AT, Hauser M, et al. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. *Sci Rep*. 2020;10(1):15100. [FREE Full text] [doi: [10.1038/s41598-020-71689-1](https://doi.org/10.1038/s41598-020-71689-1)] [Medline: [32934246](https://pubmed.ncbi.nlm.nih.gov/32934246/)]
20. Abd-Alrazaq A, Aslam H, AlSaad R, Alsahli M, Ahmed A, Damseh R, et al. Detection of sleep apnea using wearable AI: systematic review and meta-analysis. *J Med Internet Res*. 2024;26:e58187. [FREE Full text] [doi: [10.2196/58187](https://doi.org/10.2196/58187)] [Medline: [39255014](https://pubmed.ncbi.nlm.nih.gov/39255014/)]
21. Li X, Kane M, Zhang Y, Sun W, Song Y, Dong S, et al. Circadian rhythm analysis using wearable device data: novel penalized machine learning approach. *J Med Internet Res*. 2021;23(10):e18403. [FREE Full text] [doi: [10.2196/18403](https://doi.org/10.2196/18403)] [Medline: [34647895](https://pubmed.ncbi.nlm.nih.gov/34647895/)]
22. Gao C, Haghayegh S, Wagner M, Cai R, Hu K, Gao L, et al. Approaches for assessing circadian rest-activity patterns using actigraphy in cohort and population-based studies. *Curr Sleep Medicine Rep*. 2023;9(4):247-256. [doi: [10.1007/s40675-023-00267-4](https://doi.org/10.1007/s40675-023-00267-4)]
23. Haghayegh S, Gao C, Sugg E, Zheng X, Yang HW, Saxena R, et al. Association of rest-activity rhythm and risk of developing dementia or mild cognitive impairment in the middle-aged and older population: prospective cohort study. *JMIR Public Health Surveill*. 2024;10:e55211. [FREE Full text] [doi: [10.2196/55211](https://doi.org/10.2196/55211)] [Medline: [38713911](https://pubmed.ncbi.nlm.nih.gov/38713911/)]
24. Feng H, Yang L, Liang YY, Ai S, Liu Y, Liu Y, et al. Associations of timing of physical activity with all-cause and cause-specific mortality in a prospective cohort study. *Nat Commun*. 2023;14(1):930. [FREE Full text] [doi: [10.1038/s41467-023-36546-5](https://doi.org/10.1038/s41467-023-36546-5)] [Medline: [36805455](https://pubmed.ncbi.nlm.nih.gov/36805455/)]
25. Cao Z, Min J, Chen H, Hou Y, Yang H, Si K, et al. Accelerometer-derived physical activity and mortality in individuals with type 2 diabetes. *Nat Commun*. 2024;15(1):5164. [FREE Full text] [doi: [10.1038/s41467-024-49542-0](https://doi.org/10.1038/s41467-024-49542-0)] [Medline: [38886353](https://pubmed.ncbi.nlm.nih.gov/38886353/)]
26. Selenius H, Leppänen Östman S, Strand S. Self-harm as a risk factor for inpatient aggression among women admitted to forensic psychiatric care. *Nord J Psychiatry*. 2016;70(7):554-560. [doi: [10.1080/08039488.2016.1183707](https://doi.org/10.1080/08039488.2016.1183707)] [Medline: [27224513](https://pubmed.ncbi.nlm.nih.gov/27224513/)]
27. Brennan G, Flood C, Bowers L. Constraints and blocks to change and improvement on acute psychiatric wards - lessons from the city nurses project. *J Psychiatr Ment Health Nurs*. 2006;13(5):475-482. [doi: [10.1111/j.1365-2850.2006.00956.x](https://doi.org/10.1111/j.1365-2850.2006.00956.x)] [Medline: [16965464](https://pubmed.ncbi.nlm.nih.gov/16965464/)]

28. Sófi G, Törő K, Csikós G, Fliegauf G. Perceiving violence against healthcare workers in a child and adolescent emergency psychiatric ward in Hungary: a qualitative pilot study. *Front Psychol.* 2023;14:1220183. [FREE Full text] [doi: [10.3389/fpsyg.2023.1220183](https://doi.org/10.3389/fpsyg.2023.1220183)] [Medline: [38023022](https://pubmed.ncbi.nlm.nih.gov/38023022/)]
29. Laker C, Cella M, Callard F, Wykes T. Why is change a challenge in acute mental health wards? a cross-sectional investigation of the relationships between burnout, occupational status and nurses' perceptions of barriers to change. *Int J Ment Health Nurs.* 2019;28(1):190-198. [FREE Full text] [doi: [10.1111/inm.12517](https://doi.org/10.1111/inm.12517)] [Medline: [29993168](https://pubmed.ncbi.nlm.nih.gov/29993168/)]
30. Mental Health, Human Rights and Legislation: Guidance and Practice. Geneva, Switzerland. World Health Organization and the United Nations (represented by the Office of the United Nations High Commissioner for Human Rights); 2023.
31. Westheimer JL, Moukaddam N, Lindsay JA, Sabharwal A, Najafi B, Iacobelli PA, et al. Technology implementation for mental health end users: a model to guide digital transformation for inpatient mental health professionals. *JMIR Ment Health.* 2023;10:e40429. [FREE Full text] [doi: [10.2196/40429](https://doi.org/10.2196/40429)] [Medline: [37023415](https://pubmed.ncbi.nlm.nih.gov/37023415/)]
32. Langholm C, Kowatsch T, Bucci S, Cipriani A, Torous J. Exploring the potential of apple sensorKit and digital phenotyping data as new digital biomarkers for mental health research. *Digit Biomark.* 2023;7(1):104-114. [FREE Full text] [doi: [10.1159/000530698](https://doi.org/10.1159/000530698)] [Medline: [37901364](https://pubmed.ncbi.nlm.nih.gov/37901364/)]
33. Chryssikou E. Architecture for Psychiatric Environments and Therapeutic Spaces. Amsterdam. IOS Press; 2014.
34. Staggs VS. National trends and variation in nurse staffing on inpatient psychiatric units. *Res Nurs Health.* 2019;42(5):410-415. [doi: [10.1002/nur.21979](https://doi.org/10.1002/nur.21979)] [Medline: [31429481](https://pubmed.ncbi.nlm.nih.gov/31429481/)]
35. Park S, Park S, Lee YJ, Park CS, Jung YC, Kim S. Nurse staffing and health outcomes of psychiatric inpatients: a secondary analysis of national health insurance claims data. *J Korean Acad Nurs.* 2020;50(3):333-348. [doi: [10.4040/jkan.19203](https://doi.org/10.4040/jkan.19203)] [Medline: [32632069](https://pubmed.ncbi.nlm.nih.gov/32632069/)]
36. Degli Esposti M, Ziauddeen H, Bowes L, Reeves A, Chekroud AM, Humphreys DK, et al. Trends in inpatient care for psychiatric disorders in NHS hospitals across England, 1998/99-2019/20: an observational time series analysis. *Soc Psychiatry Psychiatr Epidemiol.* 2022;57(5):993-1006. [FREE Full text] [doi: [10.1007/s00127-021-02215-5](https://doi.org/10.1007/s00127-021-02215-5)] [Medline: [34951652](https://pubmed.ncbi.nlm.nih.gov/34951652/)]
37. Sohn JH, Lee BC, Park S, Oh GH, Kim HR, Kim NY, et al. Health insurance issues and future prospects related to Korean mental health care: adult mental health and hospitalization services. *J Korean Neuropsychiatr Assoc.* 2024;63(2):83-95. [doi: [10.4306/jknpa.2024.63.2.83](https://doi.org/10.4306/jknpa.2024.63.2.83)]
38. Golden G, Popescu C, Israel S, Perlman K, Armstrong C, Fratila R, et al. Applying artificial intelligence to clinical decision support in mental health: what have we learned? *Health Policy Technol.* 2024;13(2):100844. [doi: [10.1016/j.hlpt.2024.100844](https://doi.org/10.1016/j.hlpt.2024.100844)]
39. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res.* 2016;18(12):e323. [FREE Full text] [doi: [10.2196/jmir.5870](https://doi.org/10.2196/jmir.5870)] [Medline: [27986644](https://pubmed.ncbi.nlm.nih.gov/27986644/)]
40. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. [FREE Full text] [doi: [10.1136/bmj-2023-078378](https://doi.org/10.1136/bmj-2023-078378)] [Medline: [38626948](https://pubmed.ncbi.nlm.nih.gov/38626948/)]
41. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA.* 2013;310(20):2191-2194. [doi: [10.1001/jama.2013.281053](https://doi.org/10.1001/jama.2013.281053)] [Medline: [24141714](https://pubmed.ncbi.nlm.nih.gov/24141714/)]
42. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res.* 2015;17(7):e175. [FREE Full text] [doi: [10.2196/jmir.4273](https://doi.org/10.2196/jmir.4273)] [Medline: [26180009](https://pubmed.ncbi.nlm.nih.gov/26180009/)]
43. Raugh IM, James SH, Gonzalez CM, Chapman HC, Cohen AS, Kirkpatrick B, et al. Geolocation as a digital phenotyping measure of negative symptoms and functional outcome. *Schizophr Bull.* 2020;46(6):1596-1607. [FREE Full text] [doi: [10.1093/schbul/sbaa121](https://doi.org/10.1093/schbul/sbaa121)] [Medline: [32851401](https://pubmed.ncbi.nlm.nih.gov/32851401/)]
44. Balliu B, Douglas C, Seok D, Shenhav L, Wu Y, Chatzopoulou D, et al. Personalized mood prediction from patterns of behavior collected with smartphones. *NPJ Digit Med.* 2024;7(1):49. [FREE Full text] [doi: [10.1038/s41746-024-01035-6](https://doi.org/10.1038/s41746-024-01035-6)] [Medline: [38418551](https://pubmed.ncbi.nlm.nih.gov/38418551/)]
45. Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. *J Mach Learn Res.* 2010;11:2287-2322. [FREE Full text] [Medline: [21552465](https://pubmed.ncbi.nlm.nih.gov/21552465/)]
46. Jlassi O, Dixon PC. The effect of time normalization and biomechanical signal processing techniques of ground reaction force curves on deep-learning model performance. *J Biomech.* 2024;168:112116. [FREE Full text] [doi: [10.1016/j.jbiomech.2024.112116](https://doi.org/10.1016/j.jbiomech.2024.112116)] [Medline: [38677026](https://pubmed.ncbi.nlm.nih.gov/38677026/)]
47. Zhang Y, Yang Q. An overview of multi-task learning. *Natl Sci Rev.* 2018;5(1):30-43. [doi: [10.1093/nsr/nwx105](https://doi.org/10.1093/nsr/nwx105)]
48. Edmonds WA, Kennedy TD. *An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods.* Thousand Oaks, CA. SAGE Publications, Inc; 2024:392.
49. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res.* 2012;13:281-305.
50. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G. PyTorch: an imperative style, high-performance deep learning library. 2019. Presented at: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); Dec 8-14, 2019:08-14; Vancouver, BC. URL: https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: Machine learning in python. *J Mach Learn Res.* 2011;12(85):2825-2830.
52. Abd-Alrazaq A, AlSaad R, Shuweihi F, Ahmed A, Aziz S, Sheikh J. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. *NPJ Digit Med.* 2023;6(1):84. [FREE Full text] [doi: [10.1038/s41746-023-00828-5](https://doi.org/10.1038/s41746-023-00828-5)] [Medline: [37147384](https://pubmed.ncbi.nlm.nih.gov/37147384/)]
53. Abd-Alrazaq A, AlSaad R, Harfouche M, Aziz S, Ahmed A, Damseh R, et al. Wearable artificial intelligence for detecting anxiety: systematic review and meta-analysis. *J Med Internet Res.* 2023;25:e48754. [FREE Full text] [doi: [10.2196/48754](https://doi.org/10.2196/48754)] [Medline: [37938883](https://pubmed.ncbi.nlm.nih.gov/37938883/)]
54. Benton A, Mitchell M, Hovy D. Multi-task learning for mental health using social media text. 2017. Presented at: Proceedings of the 15th Conference of the EACL; December 10, 2017:152-162; Valencia, Spain.
55. Zanna K, Sridhar K, Yu H, Sano A. Bias reducing multitask learning on mental health prediction. 2022. Presented at: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII); October 21, 2022:18-21; Nara, Japan. [doi: [10.1109/acii55700.2022.9953850](https://doi.org/10.1109/acii55700.2022.9953850)]
56. Huang ZA, Liu R, Zhu Z, Tan KC. Multitask learning for joint diagnosis of multiple mental disorders in resting-state fMRI. *IEEE Trans Neural Netw Learn Syst.* 2024;35(6):8161-8175. [doi: [10.1109/TNNLS.2022.3225179](https://doi.org/10.1109/TNNLS.2022.3225179)] [Medline: [36459608](https://pubmed.ncbi.nlm.nih.gov/36459608/)]
57. Harvey A, Moreau CA, Kumar K, Huguet G, Urchs SGW, Sharmarke H. Challenges in multi-task learning for fMRI-based diagnosis: benefits for psychiatric conditions and CNVs would likely require thousands of patients. *Imaging Neurosci.* 2024:1-20.
58. Hofmann SG, Curtiss J, McNally RJ. A complex network perspective on clinical science. *Perspect Psychol Sci.* 2016;11(5):597-605. [FREE Full text] [doi: [10.1177/1745691616639283](https://doi.org/10.1177/1745691616639283)] [Medline: [27694457](https://pubmed.ncbi.nlm.nih.gov/27694457/)]
59. McNally RJ. Network analysis of psychopathology: controversies and challenges. *Annu Rev Clin Psychol.* 2021;17(1):31-53. [FREE Full text] [doi: [10.1146/annurev-clinpsy-081219-092850](https://doi.org/10.1146/annurev-clinpsy-081219-092850)] [Medline: [33228401](https://pubmed.ncbi.nlm.nih.gov/33228401/)]
60. Robinaugh DJ, Hoekstra RHA, Toner ER, Borsboom D. The network approach to psychopathology: a review of the literature 2008-2018 and an agenda for future research. *Psychol Med.* 2020;50(3):353-366. [FREE Full text] [doi: [10.1017/S0033291719003404](https://doi.org/10.1017/S0033291719003404)] [Medline: [31875792](https://pubmed.ncbi.nlm.nih.gov/31875792/)]
61. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR). Washington, DC. American Psychiatric Association; 2022.
62. Hoffman L. Longitudinal Analysis: Modeling Within-Person Fluctuation and Change. New York, NY. Routledge; 2015.
63. Xu X, Liu X, Zhang H, Wang W, Nepal S, Sefidgar Y, et al. GLOBEM: cross-dataset generalization of longitudinal human behavior modeling. *Proc ACM Interact Mob Wearable Ubiquitous Technol.* 2023;6(4):1-34. [doi: [10.1145/3569485](https://doi.org/10.1145/3569485)]
64. Goetz L, Seedat N, Vandersluis R, van der Schaar M. Generalization-a key challenge for responsible AI in patient-facing clinical applications. *NPJ Digit Med.* 2024;7(1):126. [FREE Full text] [doi: [10.1038/s41746-024-01127-3](https://doi.org/10.1038/s41746-024-01127-3)] [Medline: [38773304](https://pubmed.ncbi.nlm.nih.gov/38773304/)]
65. van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med.* 2023;21(1):70. [FREE Full text] [doi: [10.1186/s12916-023-02779-w](https://doi.org/10.1186/s12916-023-02779-w)] [Medline: [36829188](https://pubmed.ncbi.nlm.nih.gov/36829188/)]
66. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. *Science.* 2024;383(6679):164-167. This paper provides empirical evidence suggesting the need for alternatives to external validation, particularly in the field of mental health AI. [doi: [10.1126/science.adg8538](https://doi.org/10.1126/science.adg8538)] [Medline: [38207039](https://pubmed.ncbi.nlm.nih.gov/38207039/)]
67. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. *Nat Med.* 2023;29(11):2686-2687. [doi: [10.1038/s41591-023-02540-z](https://doi.org/10.1038/s41591-023-02540-z)] [Medline: [37853136](https://pubmed.ncbi.nlm.nih.gov/37853136/)]
68. Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagn Progn Res.* 2022;6(1):24. [FREE Full text] [doi: [10.1186/s41512-022-00136-8](https://doi.org/10.1186/s41512-022-00136-8)] [Medline: [36550534](https://pubmed.ncbi.nlm.nih.gov/36550534/)]
69. Wan B, Caffo B, Vedula SS. A unified framework on generalizability of clinical prediction models. *Front Artif Intell.* 2022;5:872720. [FREE Full text] [doi: [10.3389/frai.2022.872720](https://doi.org/10.3389/frai.2022.872720)] [Medline: [35573904](https://pubmed.ncbi.nlm.nih.gov/35573904/)]
70. Hotter B, Ulm L, Hoffmann S, Katan M, Montaner J, Bustamante A, et al. Selection bias in clinical stroke trials depending on ability to consent. *BMC Neurol.* 2017;17(1):206. [FREE Full text] [doi: [10.1186/s12883-017-0989-9](https://doi.org/10.1186/s12883-017-0989-9)] [Medline: [29202730](https://pubmed.ncbi.nlm.nih.gov/29202730/)]
71. Wilkins JM, Forester BP. Informed consent, therapeutic misconception, and clinical trials for alzheimer's disease. *Int J Geriatr Psychiatry.* 2020;35(5):430-435. [FREE Full text] [doi: [10.1002/gps.5262](https://doi.org/10.1002/gps.5262)] [Medline: [31895474](https://pubmed.ncbi.nlm.nih.gov/31895474/)]
72. Ford I, Norrie J. Pragmatic trials. *N Engl J Med.* 2016;375(5):454-463. [FREE Full text] [doi: [10.1056/NEJMr1510059](https://doi.org/10.1056/NEJMr1510059)] [Medline: [27518663](https://pubmed.ncbi.nlm.nih.gov/27518663/)]
73. Largent EA, Joffe S, Dickert NW, Morain SR. The ethical value of consulting community members in non-emergency trials conducted with waivers of informed consent for research. *Clin Trials.* 2024:17407745241259360. [FREE Full text] [doi: [10.1177/17407745241259360](https://doi.org/10.1177/17407745241259360)] [Medline: [38916109](https://pubmed.ncbi.nlm.nih.gov/38916109/)]
74. Kim YG, Yang S, Lee K. Multicenter analysis of emergency patient severity through local model evaluation client selection: optimizing client selection based on local model evaluation. *Appl Sci.* 2024;14(16):6876. [doi: [10.3390/app14166876](https://doi.org/10.3390/app14166876)]
75. Joe KH, Jang M. Comparison of mental health act about involuntary admission among 4 East Asian countries. *J Korean Neuropsychiatr Assoc.* 2019;58(4):297-313. [doi: [10.4306/jknpa.2019.58.4.297](https://doi.org/10.4306/jknpa.2019.58.4.297)]

76. ChatGPT. OpenAI. URL: <https://chatgpt.com/?model=gpt-4o> [accessed 2024-11-04]
77. Source code repository for "Comprehensive Symptom Prediction in Acute Psychiatric Inpatients Using Wearable-Based Deep Learning Models: Development and Validation Study". GitHub. URL: <https://github.com/gachon-CCLab/SAFER-Symptom-Prediction.git> [accessed 2024-10-30]

Abbreviations

AI: artificial intelligence
AI-CDSS: artificial intelligence–assisted clinical decision support system
AUC: area under the curve
BPRS: Brief Psychiatric Rating Scale
HAM-A: Hamilton Anxiety Scale
MADRS: Montgomery-Asberg Depression Rating Scale
MTL: multitask learning
NRMSE: normalized root mean squared error
ROC: receiver operating characteristic
SAFER: Sensor Application for Early Response in Closed Wards
t-SNE: t-distributed stochastic neighbor embedding
YMRS: Young Mania Rating Scale

Edited by G Eysenbach, A Coristine; submitted 31.08.24; peer-reviewed by C Sakal, L Yang; comments to author 30.09.24; revised version received 20.10.24; accepted 20.10.24; published 13.11.24

Please cite as:

Hong M, Kang R-R, Yang JH, Rhee SJ, Lee H, Kim Y-G, Lee K, Kim H, Lee YS, Youn T, Kim SH, Ahn YM

Comprehensive Symptom Prediction in Inpatients With Acute Psychiatric Disorders Using Wearable-Based Deep Learning Models: Development and Validation Study

J Med Internet Res 2024;26:e65994

URL: <https://www.jmir.org/2024/1/e65994>

doi: [10.2196/65994](https://doi.org/10.2196/65994)

PMID:

©Minseok Hong, Ri-Ra Kang, Jeong Hun Yang, Sang Jin Rhee, Hyunju Lee, Yong-gyom Kim, KangYoon Lee, HongGi Kim, Yu Sang Lee, Tak Youn, Se Hyun Kim, Yong Min Ahn. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 13.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.