PCN Psychiatry and Clinical Neurosciences

REGULAR ARTICLE

Diagnosing psychiatric disorders from history of present illness using a large-scale linguistic model

Norio Otsuka, MD,[1] Yuu Kawanishi, MD,[1] Fumimaro Doi, MD,[1] Tsutomu Takeda, MD,[1] Kazuki Okumura, MD,[1] Takahira Yamauchi, MD, PhD,[1] Shuntaro Yada, PhD,[2] Shoko Wakamiya, PhD,[2] Eiji Aramaki, PhD[2]* and Manabu Makinodan, MD, PhD[1]*

Aim: Recent advances in natural language processing models are expected to provide diagnostic assistance in psychiatry from the history of present illness (HPI). However, existing studies have been limited, with the target diseases including only major diseases, small sample sizes, or no comparison with diagnoses made by psychiatrists to ensure accuracy. Therefore, we formulated an accurate diagnostic model that covers all psychiatric disorders.

Methods: HPIs and diagnoses were extracted from discharge summaries of 2,642 cases at the Nara Medical University Hospital, Japan, from 21 May 2007, to 31 May 31 2021. The diagnoses were classified into 11 classes according to the code from ICD-10 Chapter V. Using UTH-BERT pre-trained on the electronic medical records of the University of Tokyo Hospital, Japan, we predicted the main diagnoses at discharge based on HPIs and compared the concordance rate with the results of psychiatrists. The psychiatrists were divided into two groups: semi-Designated with 3–4 years of experience and Residents with only 2 months of experience.

Results: The model's match rate was 74.3%, compared to 71.5% for the semi-Designated psychiatrists and 69.4% for the Residents. If the cases were limited to those correctly answered by the semi-Designated group, the model and the Residents performed at 84.9% and 83.3%, respectively.

Conclusion: We demonstrated that the model matched the diagnosis predicted from the HPI with a high probability to the principal diagnosis at discharge. Hence, the model can provide diagnostic suggestions in actual clinical practice.

Keywords: *BERT-based prediction, diagnostic prediction, history of present illness, natural language processing.*

http://onlinelibrary.wiley.com/doi/10.1111/pcn.13580/full

With the increasing predictive accuracy of machine learning methods, there is a growing trend in medicine to use artificial intelligence (AI) for diagnosis, course prediction, and treatment. Among these, diagnostic support using AI methods has attracted considerable attention because of their potential for screening and reducing misdiagnoses.[1,2] Natural language processing (NLP) is particularly promising in psychiatry, where verbal descriptions are central to information gathering and treatment. Studies have predicted suicide risk and suicide attempts from social networking service posts,[3,4] differentiated disease from online questionnaires and biomarkers,[5] and identified cognitive decline from free conversation.[6] Discharge summaries predict mental health crises[7] and early readmission.[8] However, only a few studies have compared them to psychiatrists. Currently, the application of AI to mental illness is promising but is in the nascent stage.[9]

As per the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10),[10] and the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5),[11] mental disorders are diagnosed based on diverse criteria. The diagnosis is represented using a code, and the diagnosis process becomes a disease-coding task from the computer science perspective.

When a patient is admitted to hospital, a physician using natural language summarizes the relevant events or signs and symptoms leading up to the admission of the patient in the history of present illness (HPI) (Figs S1–S4). The HPI of a patient is among the most important factors in diagnosis. However, deciphering the HPI and listing multiple candidate diagnoses can be laborious and error-prone. Thus, AI assistance would be of great benefit. Using 500 cases, Dai et al. demonstrated that pre-training improved diagnostic accuracy when predicting five major mental disorders.[12] We used a larger number of cases and adopted UTH-BERT,[13] a BERT[12]-based NLP model[14] pre-trained in precise Japanese.

The research hypothesis is that AI may achieve accuracy comparable to or even better than clinical psychiatrists, and the purpose of the study is to provide useful suggestions for researchers how the AI model can perform compared to psychiatrists at the present and how it can be developed in the future.

[1] Department of Psychiatry, Nara Medical University, Kashihara, Japan
[2] Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Japan
* Correspondence: Email: aramaki@is.naist.jp and mmm@naramed-u.ac.jp
Eiji Aramaki and Manabu Makinodan contributed equally to this work.
1st General topics in psychiatry and related fields.
2nd Social psychiatry and epidemiology.

PCN Psychiatry and Clinical Neurosciences

Methods and Materials

Corpus

As shown in Figure 1, out of 4,840 cases we used 2,642 that met the inclusion criteria. We excluded cases where the diagnosis of psychiatric illness was insignificant or unclear, that is, cases immediately transferred to another hospital, admitted for addiction, accidental ingestion, or physical treatment, were poorly documented, or hospitalized with an unspecified diagnosis. Furthermore, only the last discharge summary was considered to prevent double counting, leaving 2,642 cases for our corpus. The discharge dates ranged from 21 May 2007, to 31 May 2021. The HPI is unstructured text data, written by a physician in natural language, describing the situation up to the time of admission. The diagnoses were grouped into 11 classes by considering only the first alphabet and the digit in the 10's column of the ICD-10 code from Chapter V (Mental and behavioral disorders; F0–F9) and others.

F0 is 'Organic, including symptomatic, mental disorders' such as dementia, F1 is 'Mental and behavioral disorders due to psychoactive substance use' such as alcohol dependence, F2 is 'Schizophrenia, schizotypal and delusional disorders,' F3 is 'Mood [affective] disorders' such as depression, F4 is 'Neurotic, stress-related and somatoform disorders,' F5 is 'Behavioral syndromes associated with physiological disturbances and physical factors' such as eating disorders, F6 is 'Disorders of adult personality and behavior,' F7 is 'Mental retardation,' F8 is 'Disorders of psychological development' such as Autism, F9 is 'Behavioral and emotional disorders with onset usually occurring in childhood and adolescence' such as ADHD.

Although multiple diagnoses were occasionally listed together, the first author, with 5 years of experience as a psychiatrist, identified the primary diagnoses. The data was divided while maintaining the diagnostic proportions: 60% as training and 20% each as validation and test data. Similarly, we then divided the test data into three parts for psychiatrists. Table 1 shows the basic characteristics of each dataset.

Model

We used the full version of UTH-BERT, pre-trained on the electronic medical records at the University of Tokyo Hospital, to solve the text classification problem. The model and execution environment are presented in Table 2. The model was created based on the information in the page (https://github.com/KunikataJun/uth-bert-keras-colab). The sentences were tokenized using MeCab[15] as the morphological analyzer, J-Medic MANBYO_201907,[16] and mecab-ipadic-NEologd[17] as external dictionaries. A token is a sequence of characters that represents one unit of meaning in a sentence or text. We limited the token size to the first 400 elements and the batch size to eight, considering the limited capacity of the GPU. The distribution of the number of tokens in the HPI is shown in Figure 2. Because the token size was limited to 400, 536 HPI statements were partially truncated. In contrast, the participating psychiatrists were presented with all the sentences in the HPI, regardless of length.

A summary overview of the workflow is shown in Figure 3. The HPIs were extracted from summaries and tokenized. The diagnoses were converted to diagnostic labels. Following multiple training sessions with different hyperparameters, we adopted the model with the smallest loss on the validation set. The class was predicted from the distributed representation of the first token that went through the BERT process. The prediction results were evaluated by match rate (accuracy), precision, recall, and F1-score. Furthermore, an analysis of variance (ANOVA) was performed for the comparison in the match rates between the proposed model and clinical psychiatrists and in word counts in the HPI for each disease group. T-test was performed to compare the match rates to guaranteed cases between the model and the Residents.

When handling information on the electronic medical record, we used an 'opt-out' approach that guaranteed an opportunity to refuse, and only subjects who did not refuse were included in the study. Data handling and management methods followed those approved by the Ethical Review Committee of Nara Medical University, which conformed to the provisions of the Declaration of Helsinki.

Psychiatrists

Six psychiatrists from the Department of Psychiatry at Nara Medical University participated. Three had 2 years of initial training and only 2 months of experience as psychiatrists (Residents) and three had 3–4 years of experience and had applied or were about to apply for Designated Physicians of Mental Health certification as a full-fledged psychiatrist (semi-Designated), a Japanese national legal requirement. We divided the test data into three parts. Each psychiatrist predicted 176 or 177 diagnoses. The percentages of all the data by class were

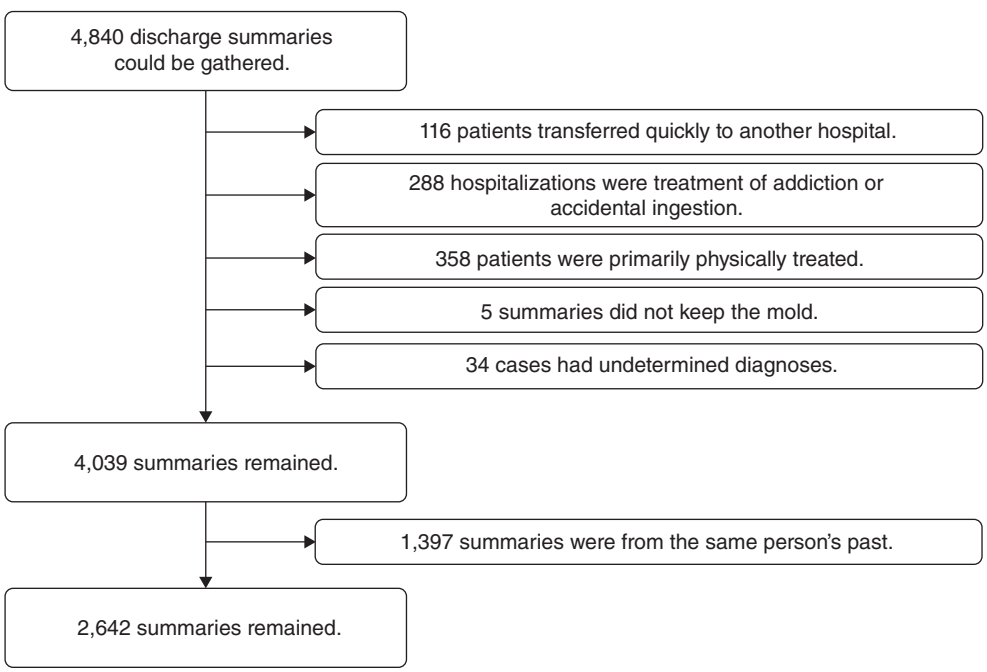

Fig. 1 Exclusion process: We excluded cases where the diagnosis of psychiatric illness was unimportant or unclear. Furthermore, only the last discharge summary was considered.

Table 1. Demographic features

			Test_set ($N = 529$)			
	Train set	Validataion set	Test_set 1	Test set 2	Test set3	Total
	($n = 1585$)	($n = 528$)	($n = 176$)	($n = 176$)	($n = 177$)	($n = 2642$)
Age at hospitalization (years, mean ± SD)	47.7 ± 20.4	47.1 ± 20.5	47.2 ± 20.3	47.8 ± 19.8	46.7 ± 19.8	47.5 ± 20.3
Female (%)	884 (55.8%)	294 (55.7%)	97 (55.1%)	102 (58.0%)	98 (55.4%)	1475 (55.8%)
Number of stays in hospital (days, 25% quantile/median/75% quantile)	47.0/81.0/93.0	48.0/80.0/91.0	51.5/84.0/92.2	45.0/76.0/96.2	55.0/84.0/99.0	48.0/81.0/92.8
F0, Organic disorders	211	70	23	23	24	351 (13.3%)
F1, Substance use disorders	94	31	10	11	10	156 (5.9%)
F2, Psychotic disorders	434	144	48	48	49	723 (27.4%)
F3, Mood disorders	413	138	46	46	46	689 (26.1%)
F4, Neurotic disorders	208	69	23	23	24	347 (13.1%)
F5, Eating disorders, etc.	100	33	11	11	12	167 (6.3%)
F6, Personality disorders	37	13	4	4	4	62 (2.3%)
F7, Mental retardation	23	8	3	3	2	39 (1.5%)
F8, Autism, etc.	41	13	5	5	4	68 (2.6%)
F9, Behavioral disorders	11	4	1	1	1	18 (0.7%)
Others	13	5	2	1	1	22 (0.8%)

indicated in advance. They referred to the ICD-10 and, without any time limit, aimed for a higher matching rate.

Results

The results are listed in Table 3. The match rate of the model was 74.3% (95% CI, 70.4%–78.2%), with 71.5% (95% CI, 62.0%–80.9%) for the semi-Designated and 69.4% (95% CI, 68.2%–70.5%) for the Residents. The scores were not statistically different between the model and the two groups of clinical psychiatrists (one-way ANOVA, F: 5.14, $P = 0.199$).

Table 4 shows the precision, recall, and F1-score for each diagnosis. In descending order of the F1-score, using the proposed model, the disorder groups are F5, F2, F3, F1, F0, F4 and F8, while F6, F7, F9 and others are not predictive. In the F9 group, neither the psychiatrist nor the model matched at all. The F1-scores were also extracted by the disorder group, as shown in Figure 4.

Table 5 shows the patterns and matching cases for the Residents and semi-Designated physicians, and the proposed model in eight patterns from (A) to (H). A matched answer was assigned a '✓' and a mismatched answer was assigned a '×.' In 53.9% of cases, all three

Table 2. Pre-trained model and execution environment

Pre-trained model	UTH-BERT-BASE-512-WWM (12-layer, 768-hidden, 12-heads)
Dataset	Electronic Medical Records at University of Tokyo Hospital
Max seq length	512
Max position embeddings	512
Vocabulary size	25,000
Morphological analyzer	MeCab
External dictionary	J-MeDic (MANBYO_201907), mecab-ipadic-neologd
Python	3.8.8
Tensorflow	2.3.0
Keras-BERT	0.86.0
GPU	NVIDIA GeForce RTX 3080(memory: 10GB × 1)
Max tokens	400
Batch size	8
Optimizer	AdamWarmup
Loss function	Categorical crossentropy
Learning rate	1e-4
Epochs	3

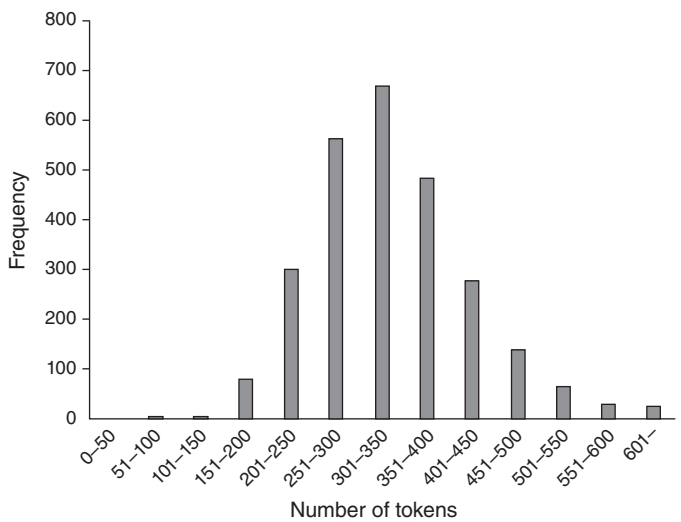

Fig. 2 Number of tokens in the history of present illness.

PCN Psychiatry and Clinical Neurosciences

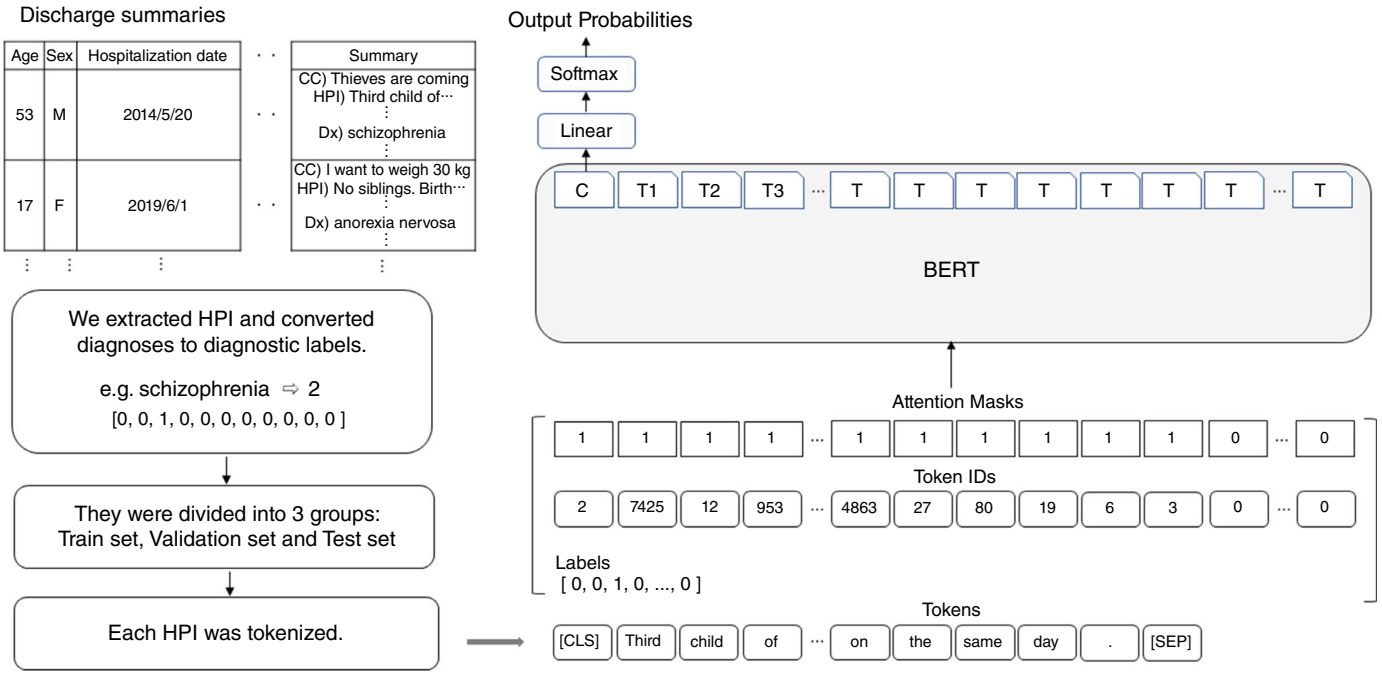

Fig. 3 Outline of the general workflow. The history of present illness (HPI) was extracted from the discharge summary, and the principal diagnosis at discharge was assigned to one of the 11 classes. Each HPI was tokenized and given an ID for each token; the IDs were converted to distributed representations acquired through pre-training within BERT. During the learning process, the parameters of both BERT and the classifier were updated (fine-tuning).

were consistent with the diagnosis at discharge (A). In contrast, in 12.9% of cases, none matched (B). Only the model differed in 5.7% of cases (C), only the Resident differed in 6.8% of the cases (D), and the semi-Designated physicians differed in 7.8% of the cases (H). The model alone matched 5.9% of the cases (E), while for the semi-Designated physicians, the percentage was 5.1% (G).

Considering that some of the HPIs may be inappropriate as questions, Table 6 compares only those questions whose quality was ensured by the match of the semi-Designated physician. The percentages of the match rate by the proposed model and the Residents were 84.9% (95% CI, 78.5%–92.5%) and 83.3% (95% CI, 81.1%–85.8%), respectively (t-test, $P = 0.545$).

Examples of attention visualizations are shown in Figures S1–S4. We created four fictitious HPI documents using Japanese and English translations. These are visualizations of the last attention weights from the head token [CLS] to the other tokens in the input summed over the number of heads.[14] The highlighted parts correspond to the input token effective for classification tasks. The match rate of the PyTorch model used for visualization was slightly lower than that of the original model (72.6% vs 74.3%), but a general trend can be observed. Tokens with higher attention weights are redder and accompanied by English translations. We also included English translations of words that were unintentionally divided into sub-words. The words described as unknown tokens [UNK] in this analysis are tabulated in Table S1. The word counts of HPI were significantly different between the diagnosis groups (F: 1.83, $P = 0.00119$) (Table S2).

Table 3. Match rate

	Residents	Semi-designated	Model
Test set1	69.9%	72.2%	76.1%
Test_set2	68.8%	66.5%	74.4%
Test_set3	69.5%	75.7%	72.3%
Total	69.4%	71.5%	74.3%

Discussion

Studies in practical settings requiring a comprehensive prediction of mental illness are limited. To apply AI in clinical practice, we assumed real-life situations that psychiatrists would encounter. The results suggest that the proposed model may perform as effectively as clinical psychiatrists. This result demonstrates that AI can successfully collect diagnostic information from HPI. The combined use of these linguistic models and biological research such as peripheral transcriptome[18] may increase the prediction rate even more.

Explicit diagnoses and preconceptions in the HPIs

In some cases, the diagnosis was included in the HPI. Even if it was not included, HPI contains other information based on the preconceived notions of the writer. However, this type of information is advantageous to AI and psychiatrists and does not diminish the importance of the comparison. For the semi-Designated physicians, the cases they have treated were included, contributing to the increased match rate.

Disease and performance

For diseases with a high number of cases, no large differences were observed between human predictions and the proposed model. However, minor differences persist. In the F4 area, the model outperformed psychiatrists by a relatively large margin regarding the F1-score. The F4 area includes anxiety disorder, adjustment disorder, and somatoform disorder, and words suggesting physical symptoms are likely to appear in the HPI. ‘Dizziness,’ ‘headache,’ ‘hyperventilation,’ and other medical terms that are relatively common in F4, thus improving the rate.

In contrast, psychiatrists generally outperformed the model in the F1 area, which refers to substance, such as alcohol and methamphetamine, abuse disorders. A possible reason is the difference in patient populations among the institutions; the University of Tokyo

Table 4. Precision, recall, and F1-score of participants and the model

	Residents			Semi-designated			Model			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
F0	0.796	0.614	0.694	0.789	0.643	0.709	0.697	0.657	0.676	70
F1	0.694	0.806	0.746	0.771	0.871	0.818	0.700	0.677	0.689	31
F2	0.803	0.814	0.808	0.833	0.862	0.847	0.833	0.897	0.864	145
F3	0.700	0.761	0.729	0.759	0.775	0.767	0.797	0.768	0.782	138
F4	0.642	0.486	0.553	0.675	0.386	0.491	0.563	0.700	0.624	70
F5	0.800	0.941	0.865	0.882	0.882	0.882	0.868	0.971	0.917	34
F6	0.059	0.083	0.069	0.158	0.250	0.194	0.000	0.000	0.000	12
F7	0.333	0.500	0.400	0.167	0.625	0.263	0.000	0.000	0.000	8
F8	0.294	0.357	0.323	0.467	0.500	0.483	0.500	0.571	0.533	14
F9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	3
Others	0.000	0.000	0.000	0.286	0.500	0.364	0.000	0.000	0.000	4
Accuracy	-	-	0.694	-	-	0.715	-	-	0.743	529
Macro avg	0.466	0.488	0.471	0.526	0.572	0.529	0.451	0.476	0.462	529
Weighted avg	0.699	0.694	0.692	0.743	0.715	0.720	0.713	0.743	0.726	529

Hospital, which generated UTH-BERT, provides no addiction treatment. Therefore, terms such as 'stimulant drug,' which is critical for F1 diagnosis, were treated as [UNK]. This probably reduced the model's matching rate. Thus, the frequency of occurrence and the recognition of words that suggest a diagnosis may contribute to diagnostic accuracy.

Matching patterns

As shown in Table 5, 53.9% of cases showed label concordance between the AI classification and medical evaluations (A). This is probably because directly naming the illness in HPI or including words strongly related to the diagnosis, such as 'mania' or 'low body weight,' make prediction easier. In our model, the accurate translation into psychiatric terminology of symptoms and course of events seems critical for diagnostic concordance. In contrast, 12.9% of cases are probably very difficult to predict or misleading (B). In certain cases, the diagnosis clearly changes during hospitalization, making prediction difficult. Consequent to a detailed evaluation or due to a change in condition, anxiety disorders may transition into a diagnosis of depression, and depression can subsequently transition into a diagnosis of dementia. Although full prediction is difficult, if there had been more episodes or terms in the HPIs that suggested the diagnosis, the concordance rate would have been slightly higher.

The semi-Designated physician responsible for Test set 2, which had the lowest match rate, had the highest number of single matches and single discrepancies (G and H), suggesting that he may have focused on guessing relatively rare diseases. Without this bias, the matching rate from the semi-Designated group may have exceeded that of the model.

The difference between 5.9% (E) and 5.1% (G) is negligible, suggesting a minimally inappropriate increase in the accuracy of the model. The AI model could have used the individual writing style characteristics of the attending physicians and supervisors and bias toward the diseases for which they were responsible; however, this factor does not seem to matter. As shown in Table 4, even when limiting the cases that are consistent for the semi-Designated group, the model outperformed the Residents group, thus confirming its usefulness.

Limitations

Insufficient statistical comparisons

This study failed to show a statistically superiority or non-inferiority between the model and clinical psychiatrists; additional studies are needed to demonstrate the statistical superiority of the AI. As shown in Table S2, there was a significant difference in ANOVA regarding the number of words in the HPIs that may have influenced the results.

Disease classification and labeling errors

The study model only predicted the operative diagnosis at discharge in ICD-10, not the true disease. In addition, the ICD-10 codes used

Table 5. Match pattern

	A	B	C	D	E	F	G	H
Residents	✓	×	✓	×	×	✓	×	✓
Semi-designated	✓	×	✓	✓	×	×	✓	×
Model	✓	×	×	✓	✓	×	×	✓

	A	B	C	D	E	F	G	H	Total
Test set 1	99	23	7	13	9	4	8	13	176
Test set 2	90	23	6	10	11	5	11	20	176
Test set 3	96	22	17	13	11	2	8	8	177
Total	285 (53.9%)	68 (12.9%)	30 (5.7%)	36 (6.8%)	31 (5.9%)	11 (2.1%)	27 (5.1%)	41 (7.8%)	529 (100.0%)

PCN Psychiatry and Clinical Neurosciences

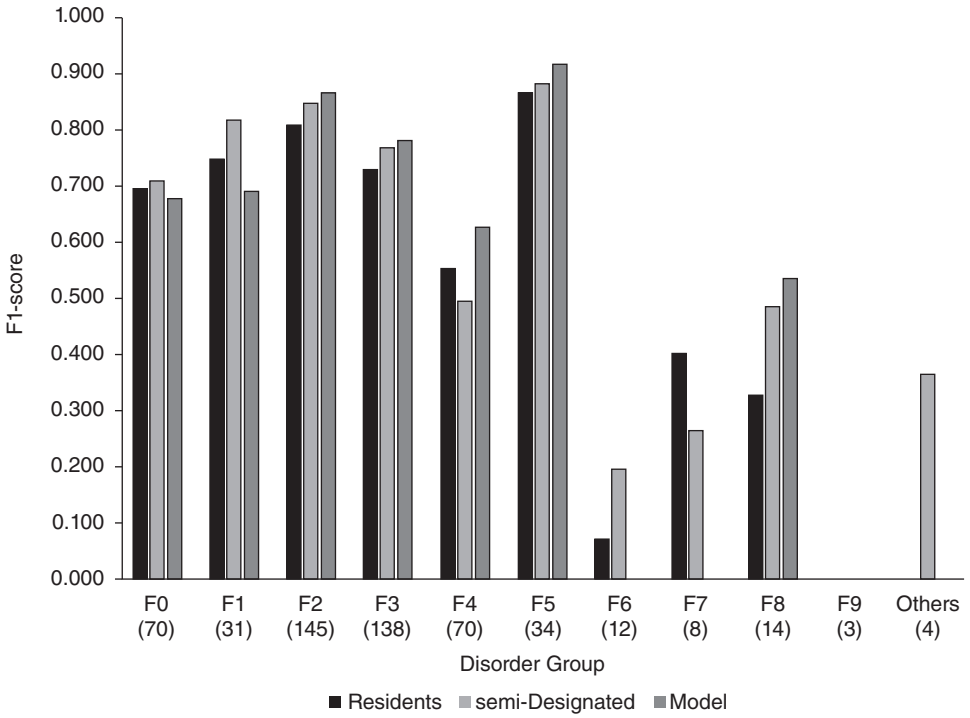

Fig. 4 F1-scores by disorder group. To see the approximate accuracy for each disorder group, only the F1-scores were extracted from Table 4. For reference, the total numbers of cases used for testing are appended under each disorder group.

for diagnosis prediction were limited to the digit in the 10's column to avoid fragmentation and facilitate AI training. Therefore, diverse pathologies were included within the same class. Multiple psychiatric diagnoses are also common. When two or more diagnoses are listed, it can be difficult to determine the primary diagnosis. The possibility cannot be excluded that such labeling errors unintentionally favored the model. Predicting mental illness is inherently a multi-label problem and requires further research.

Quality and bias in HPI

The quality of the HPI varies depending on the experience and skill of the physician that wrote it, which may affect the predictive ratios. Because the HPI is often provided by a psychiatrist with knowledge of mental disorders, typical episodes and symptoms of the disorder are more likely to be included in the HPI if it is possible to predict the diagnosis in advance. It is unclear whether a similar predictive task is possible with an HPI written by a physician with little knowledge of mental disorders.

Unknown tokens and inappropriate sub-words

As shown in the Table S1, this model does not handle location and time series information, as many place names and time expressions become [UNK]. Other important words, such as 'abuse,' 'hanging up the neck,' and 'stimulant drug,' were also labeled as [UNK]. The inclusion of 'police officer,' 'arrest,' and 'involuntary hospitalization' indicates that Nara Medical University Hospital accepts 'involuntary hospitalization.' The visualized attention data (the Figs S1–S4) shows that the model paid close attention to seemingly important words such as 'loss of appetite' and 'auditory hallucination.' In contrast, words such as 'clozapine,' which identify the disease by itself in Japan, are not registered as drug names, and are divided into sub-words and not given proper attention. Therefore, there is room for further development when using the HPI information, which requires increasing the number of registered words and conducting preliminary studies involving multiple facilities.

Table 6. Match rate on guaranteed cases

	Residents	Model	Eligible cases
Test set 1	83.5%	88.2%	127
Test set 2	82.1%	85.5%	117
Test set 3	84.3%	81.3%	134
Total	83.3%	84.9%	378

Concordance rate in diagnosis

In comparing the concordance rate in diagnosis, we considered it desirable to inform psychiatrists of the percentages of the data per diagnosis. We had recognized differences in diagnostic thresholds, possibly reducing the concordance rate, particularly because inexperienced psychiatrists were probably unaware of diagnostic trends at the facility.

The tendency to ignore the true disease rates is termed base-rate neglect.[19] One reason for using AI is to correct biases, and this process reduced base-rate neglect by revealing disease proportionality. Nevertheless, for Test set 2 the semi-Designated group seemed more concerned with guessing the relatively unlikely disorder types. This could be interpreted as being in the nature of the physicians to make a more meaningful diagnosis. Therefore, the interest of the physicians in diagnosis should be considered when comparing diagnostic accuracy between physicians and AI. Alternatively, this could be interpreted as a limitation of the accuracy measure. The percentage of correct answers alone is insufficient to measure diagnostic ability. Clinically, indicators that consider the possibility of important diseases may be more appropriate.

Imbalanced data

Disorders with only a small amount of data were not well predicted. The training size problem can be solved by increasing the number of cases; for example, by collecting data from multiple facilities. However, other factors must also be considered. Li et al. stated that classification models tend to be biased toward majority classes and do not

provide adequate training for recognizing minority classes.[20] The imbalanced data problem was not well addressed in the present analysis. Krawczyk grouped the solutions to the imbalance problem into three categories: the data-level, algorithm-level, and hybrid methods.[21] The data-level methods modify the training distributions using techniques such as over-sampling or under-sampling. The algorithm-level methods involve changing the weight of each class. For example, Madabushi et al. incorporated cost-weighting into BERT.[22] The hybrid methods combine the two techniques. In this study, these methods were not implemented because of interpretation complexity. These modifications will be necessary, depending on the degree of imbalance or on the intended use of AI.

Explainability

Kundu contended that 'Black-box medicine without a clinical link is not good medicine.'[23] In this regard, attention visualization is a promising solution, as shown in the example in the Figures S1–S4. However, the meaning of high attention is unclear, with many criticisms against attention as a basis for AI decisions; as Jain and Wallace noted, 'standard attention modules do not provide meaningful explanations.'[24] Therefore, methods that improve explainability need to emerge.

Locality

The results of this study have not been validated with data from other institutions. The actual diagnostic threshold may differ between or even within facilities. Moreover, different regions and different eras will have different proportions of patient diagnoses. Therefore, the present model may not be generally applicable. Nonetheless, even using larger-scale linguistic models in the future, to predict diagnoses in clinical settings will require regional- or facility-specific data to achieve high accuracy. Moreover, combining information such as age, gender, head image, voice features, conversation content, facial expression data, and sequence data will further increase the accuracy.[25]

Conclusion

Predictions by the proposed AI model and psychiatrists of the main diagnosis of psychiatric disorders using histories of present illness were compared. In the future, we expect to conduct additional validation that considers diagnoses other than the primary diagnosis, improve coincident accuracy using data from various facilities to increase the vocabulary of the pre-trained models, and combine multiple data types to be tuned for use in actual clinical settings.

Disclosure statement

None.

Author contributions

Norio Otsuka processed and analyzed the data and wrote a draft of the paper. Yuu Kawanishi, Fumimaro Doi, Tsutomu Takeda, Kazuki Okumura, and Takahira Yamauchi provided specific suggestions on the concept, content, and wording of the paper from the early stages of the project. Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki provided guidance on analytical methods and content, mainly from the perspective of machine learning and natural language processing. Manabu Makinodan supervised the entire project and gave frequent feedback.

Funding Information

AMED-PRIME (grant number 21gm6310015h0002 to M.M.), AMED-CREST (grant number 22gm1510009h0001 to M.M.), AMED (grant number 21wm04250XXs0101 to M.M.), AMED (grant number 21uk1024002h0002 to M.M.).

Data Availability Statement

The data are not publicly available due to containing information that could compromise the privacy of research participants.

References

1. Squires M, Tao X, Elangovan S *et al*. Deep learning and machine learning in psychiatry: A survey of current progress in depression detection, diagnosis and treatment. *Brain Inform.* 2023; **10**: 10.
2. Shiba K, Daoud A, Kino S, Nishi D, Kondo K, Kawachi I. Uncovering heterogeneous associations of disaster-related traumatic experiences with subsequent mental health problems: A machine learning approach. *Psychiatry Clin. Neurosci.* 2022; **76**: 97–105.
3. Ophir Y, Tikochinski R, Asterhan CSC, Sisso I, Reichart R. Deep neural networks detect suicide risk from textual Facebook posts. *Sci. Rep.* 2020; **10**: 16685.
4. Wang N, Luo F, Shivtare Y *et al*. Learning models for suicide prediction from social media posts. **In:** *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology*, Association for Computational Linguistics, 2021; 87–92.
5. Tomasik J, Han SYS, Barton-Owen G *et al*. A machine learning algorithm to differentiate bipolar disorder from major depressive disorder using an online mental health questionnaire and blood biomarker data. *Transl. Psychiatry* 2021; **11**: 41.
6. Horigome T, Hino K, Toyoshiba H *et al*. Identifying neurocognitive disorder using vector representation of free conversation. *Sci. Rep.* 2022; **12**: 12461.
7. Garriga R, Mas J, Abraha S *et al*. Machine learning model to predict mental health crises from electronic health records. *Nat. Med.* 2022; **28**: 1240–1248.
8. Rumshisky A, Ghassemi M, Naumann T *et al*. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl. Psychiatry* 2016; **6**: e921.
9. Cearns M, Hahn T, Baune BT. Recommendations and future directions for supervised machine learning in psychiatry. *Transl. Psychiatry* 2019; **9**: 271.
10. The International Classification of Diseases. 10th revision. https://icd.who.int/browse10/2019/en#. Accessed 2023/3/10 2019. World Health Organization.
11. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. American Psychiatric Association, Arlington, 2013.
12. Dai HJ, Su CH, Lee YQ *et al*. Deep learning-based natural language processing for screening psychiatric patients. *Front. Psych.* 2020; **11**: 533949.
13. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed using a huge Japanese clinical text corpus. *PLoS One* 2021; **16**: e0259763.
14. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. **In:** *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. **1**, Association for Computational Linguistics, Minneapolis, Minnesota, 2019; 4171–4186.
15. MeCab TK. Yet another part-of-speech and morphological analyzer. https://github.com/taku910/mecab. Accessed March 21 2023.
16. Ito K, Nagai H, Okahisa T *et al*. J-medic: A Japanese disease name dictionary based on real clinical usage. **In:** *Proceedings of the 11th International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA), Miyazaki, Japan, 2018; 2365–2369 http://sociocom.jp/~data/2018-manbyo. Accessed March 21 2023.
17. Sato T, Hashimoto T, Okumura M. Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval: NLP2017-B6–1. **In:** *Proceedings of the Twenty-Three Annual Meeting of the Association for Natural Language Processing*. The Association for Natural Language Processing, Tsukuba, Japan, 2017; NLP2017–B6.
18. Weichen S, Lihua X, Tianhong Z *et al*. Peripheral transcriptome of clinical high-risk psychosis reflects symptom alteration and helps prognosis prediction. *Psychiatry Clin. Neurosci.* 2022; **76**: 268–270.
19. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad. Med.* 2003; **78**: 775–780.
20. Li J, Fong S, Mohammed S, Fiaidhi J. Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms. *J. Supercomput.* 2016; **72**: 3708–3728.
21. Krawczyk B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* 2016; **5**: 221–232.

22. Madabushi HT, Kochkina E, Castelle M. Cost-sensitive BERT for generalisable sentence classification with imbalanced data. **In:** *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. ACL, Hong Kong, Chinac, 2019; 125–134.
23. Kundu S. AI in medicine must be explainable. *Nat. Med.* 2021; **27**: 1328.
24. Jain S, Wallace BC. Attention is not explanation. **In:** *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. ACL, Minneapolis, Minnesota, 2019; 3543–3556.
25. Soenksen LR, Ma Y, Zeng C *et al*. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digit. Med.* 2022; **5**: 149.

Supporting Information

Additional supporting information can be found online in the Supporting Information section at the end of this article.