

Transforming Bipolar Disorder Diagnosis: Clinical Validation of AI-Enhanced Bipolar Interview (AEBI)

Enzhao Zhu¹, Huiqing Pan¹, Xiao Wang¹, Huan Wang¹, Zheyu Cai¹, Xiuake Bahuoja¹, Jiayi Wang¹, Xinyi Tang¹, Ruyi Shui¹, Weiwei Xu¹, Tongxing Ou¹, Zisheng Ai^{2,3,*}

¹School of Medicine, Tongji University, Shanghai, China

²Department of Medical Statistics, School of Medicine, Tongji University, Shanghai, China

³Putuo People's Hospital, School of Medicine, Tongji University, Shanghai, China

Email address:

zhuenzhao@outlook.com (Enzhao Zhu), 2211471@tongji.edu.cn (Huiqing Pan), naturalconan@163.com (Xiao Wang), whjoyce2021@163.com (Huan Wang), caizheyu@outlook.com (Zheyu Cai), shuakwj@163.com (Xiuake Bahuoja), jywang202207@126.com (Jiayi Wang), tangxinyi99@163.com (Xinyi Tang), 2252371@tongji.edu.cn (Ruyi Shui), 2231241@tongji.edu.cn (Weiwei Xu), outongxing2003@163.com (Tongxing Ou), azs1966@126.com (Zisheng Ai)

*Corresponding author

Abstract

Background: Bipolar disorder (BD) is prevalent in China, but limited resources and a shortage of trained professionals lead to frequent misdiagnosis and inappropriate treatment. It is thus imperative to develop an effective tool for accurate BD identification. **Objective:** The study develops the ICD-11-based AI-Enhanced Bipolar Interview (AEBI), an integrated system for BD assessment in outpatient settings. Within AEBI, the AEBI-Structured Interview (AEBI-SI), adapted from Flexible Interview for ICD-11, is a structured interview system, primarily based on patient self-reports, with additional support from nurses. AEBI-SI data are subsequently processed by a deep learning model to enable symptom identification and BD diagnosis. **Methods:** We developed AEBI-SI using the Delphi method. We trained the diagnostic model on AEBI-SI data from 4,661 first-visit outpatients across 14 hospitals in Shanghai (training/internal testing) and Sichuan (external testing), with the gold standard independently determined by chief psychiatrists after outpatient consultations. To enhance model generalization, we collected 56,688 Mental Status Examination (MSE) records retrospectively for domain adaptation. We compared 18 traditional and Transformer-based models for primary diagnosis and identification of six comorbid symptoms and conditions. Finally, a self-controlled trial was conducted from October 2024 to March 2025, involving 472 first-visit BD outpatients and six junior psychiatrists, to evaluate the impact of AEBI assistance on diagnostic accuracy and consultation efficiency compared to those without. **Results:** The final AEBI-SI retained 36 items following two iterative revisions: 22 BD-related and 14 comorbid symptom items. XLNet outperformed other models across three datasets, achieving a one-vs-rest macro-averaged F1 of 0.899 for primary diagnosis, an area under the receiver operating characteristic curve (AUROC) ranging from 0.971 to 0.983, and an area under the precision-recall curve (AUPRC) between 0.938 and 0.962. The model achieved an F1 of 0.901 for type I BD, 0.890 for type II BD, and 0.804 for BD NOS. The incorporation of MSE significantly enhanced model performance, increasing the F1 by 16.34%, AUROC by 8.80%, and AUPRC by 14.91% ($P < 0.001$) compared with those without MSE. In the self-controlled trial, AEBI significantly improved diagnostic performance, raising the F1 from 0.742 (95% CI, 0.722–0.764) to 0.829 (95% CI, 0.814–0.847) while enhancing consultation efficiency by 30.23% ($P < 0.001$).

Conclusion: AEBI is effective in diagnosing BD subtypes and identifying symptoms, enhancing diagnostic accuracy and efficiency for junior physicians in real-world settings. It facilitates the detection of key comorbid symptoms, potentially enabling more targeted and precise treatment strategies.

Keywords

Bipolar Disorders, Diagnostic Model, Deep Learning, Structured Clinical Interview, ICD-11, Transformers, Real-World Study