Clinical Epidemiology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Assessing the Validity of Claims-Based Diagnostic Codes for Psychotic and Affective Disorders and the Influence of the Coding Transition from the ICD-9 to the ICD-10 in Taiwan's National Health Insurance Research Database

Yen-Wen Wang[1], Chen-Chung Liu[2], Hsi-Chung Chen[2], Chi-Shin Wu[3,4], Jen-Hui Chan[5], Cheng-Che Chen[6], Wei-Lieh Huang[4], Shih-Cheng Liao[2,5], Tzung–Jeng Hwang[2], Wei J Chen[1,2,7,8]

[1]Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan; [2]Department of Psychiatry, College of Medicine and National Taiwan University Hospital, National Taiwan University, Taipei, Taiwan; [3]National Center for Geriatrics and Welfare Research, National Health Research Institutes, Miaoli, Taiwan; [4]Department of Psychiatry, National Taiwan University Hospital Yunlin Branch, Yunlin, Taiwan; [5]NTU Hsin-Chu Hospital, National Taiwan University Hsin-Chu Branch, Hsinchu, Taiwan; [6]NTU Biomedical Park Hospital, National Taiwan University Hsin-Chu Branch, Hsinchu, Taiwan; [7]Department of Public Health, College of Public Health, National Taiwan University, Taipei, Taiwan; [8]Center for Neuropsychiatric Research, National Health Research Institutes, Miaoli, Taiwan

Correspondence: Wei J Chen, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan, Email wjchen@ntu.edu.tw

Purpose: No studies have validated psychiatric diseases diagnoses in Taiwan's National Health Insurance Research Database (NHIRD). We aimed to assess the interrater reliability of chart-review among psychiatrists, examine the validity of the diagnostic codes for psychotic disorders and affective diseases in the NHIRD against review-based diagnoses, and examine whether the change in the coding system from the ICD-9-CM to the ICD-10-CM affected the validity of the diagnostic codes.

Patients and Methods: The study participants were psychiatric inpatients aged 18 to 65 years who were admitted in 2015 and 2017, respectively, to the main and three branch hospitals of National Taiwan University Hospital. A chart review was conducted among 48 purposively selected inpatients with discharge diagnoses in five core categories to assess interrater reliability. This chart-review procedure was then used to generate diagnostic codes for a stratified sampling of 727 inpatients with discharge diagnoses in 12 diagnostic categories of psychotic disorders and affective disorders to examine the validity of the diagnostic codes.

Results: The intraclass correlation coefficient reliability of schizophrenia and three broad categories of diagnoses indicated good interrater reliability. The positive predictive value and sensitivity of common diagnoses in the narrow category (eg, schizophrenia) or the broad category (eg, psychotic disorders, bipolar disorders, and major depressive disorders) were high-performing (≥ 0.70), whereas those of the diagnoses of low prevalence were modest. The validity indices of claims-based diagnoses using the ICD-10-CM tended to be better than those using the ICD-9-CM.

Conclusion: This first-ever study validating psychiatric diagnoses in Taiwan's NHIRD using a structured chart review suggests that the diagnostic codes of narrow categories of schizophrenia or other broad categories are recommended for high-performing validity indices. Intensive training for the coding plus the specific details requested by the ICD-10 may increase the validity of the claims-based databases for psychotic and affective disorders.

Plain Language Summary: Mental disorders such as psychotic disorders and affective disorders have contributed substantially to the global burden of disease. Since Taiwan launched the National Health Insurance (NHI) in 1995, its claims data have been compiled into the National Health Insurance Research Database (NHIRD), which has a coverage rate over 99% of Taiwan's population and is thus widely used by researchers. However, the validity of the diagnostic codes for psychotic disorders and affective disorders and the impact of the coding system transition in the NHIRD have not been examined. In this study, we developed a standardized process for reviewing the medical records of psychiatric inpatients at one medical center and its branches in suburban or rural areas, which

Received: 7 April 2025
Accepted: 1 July 2025
Published: 10 July 2025

provided an opportunity to examine the validity of psychiatric diagnoses in hospitals across urbanicity levels in the NHIRD. Among 48 inpatients in five core categories, including schizophrenia, manic/mixed episode with psychotic features, depressive episode with psychotic features, bipolar disorder without psychotic features, and major depressive disorder without psychotic features, good interrater reliability was achieved. In another 727 inpatients with psychotic and affective disorders, the positive predictive value and sensitivity of common diagnoses in the narrow category (eg, SZ) or broad category (eg, psychotic disorders, bipolar disorders, and major depressive disorders) were high-performing (≥ 0.70), whereas those of the diagnoses of low prevalence were modest. Intriguingly, the validity indices of claims-based diagnoses using the International Classification of Diseases, Tenth Revision (ICD-10) tended to be better than those using the ICD-9.

Keywords: psychiatry, psychotic disorders, health insurance database, positive predictive value, sensitivity, interrater reliability, intraclass correlation coefficient reliability, electronic health records

Introduction

Psychotic disorders, including schizophrenia, schizoaffective disorder, schizophreniform disorder, brief psychotic disorder, and affective disorder with psychotic features, are estimated to have a lifetime prevalence of approximately 3% worldwide.[1,2] Together with the nonpsychotic forms of depressive disorder and bipolar disorder, these mental disorders have contributed substantially to the global burden of disease. The proportion of global disability-adjusted life-years (DALYs) attributed to mental disorders increased from 3.1% in 1990 to 4.9% in 2019.[3] When disorders were ranked from highest to lowest in terms of DALYs for all ages, depressive disorders, schizophrenia, and bipolar disorder were among the top 5 mental disorders, with a global age-standardized prevalence in 2019 of 3.44% for depressive disorders, 0.29% for schizophrenia, and 0.49% for bipolar disorder.[3] The complexity of psychotic symptoms and varying tools for case finding pose challenges for epidemiological surveys.[1,4,5] A growing body of research on mental disorders has used administrative claims databases because they provide longitudinal real-world data on hospitalizations, major procedures, and medication use in large populations.[6–12]

Since Taiwan launched the National Health Insurance (NHI) in 1995, its claims data have been compiled into the National Health Insurance Research Database (NHIRD), which has a coverage rate over 99% of Taiwan's population and is thus widely used by researchers.[13–15] The majority of individuals who were first admitted for psychotic disorders from 1998–2007, according to the NHIRD, were affected by schizophrenia (72.5%), followed by other nonorganic psychoses (18.0%) and then affective psychoses (13.6%).[16] Nonetheless, the submission of claims data was mainly for financial and reimbursement purposes instead of research purposes, and the results of routine auditing of claims data were not available to researchers. Hence, it is important to validate the diagnostic categories in administrative claims databases to avoid or correct for misclassification biases in outcomes.[17,18] To examine the validity of the diagnostic codes in claims data, a commonly used approach is to have psychiatrists review patients' case notes or clinical records.[19–21] Moreover, the consistency of psychiatrists' review of medical records, measured as interrater reliability, needs to be assessed before the implementation of a validation study.[22,23]

In recent years, several research teams in Taiwan have conducted validation studies for the diagnostic codes for the diseases of their interest in the NHIRD. For example, studies have evaluated the validity of the diagnostic codes for ischemic stroke,[24,25] acute myocardial infarction,[26,27] psoriasis,[28] chronic obstructive pulmonary disease,[29] cancer,[30] and glaucoma.[31] The results revealed high positive predictive value and sensitivity for the diseases investigated in the NHIRD. In addition, the diagnostic coding system was changed from the International Classification of Diseases (ICD), Ninth Revision, Clinical Modification (ICD-9-CM) to the ICD, Tenth Revision, Clinical Modification (ICD-10-CM), in 2016. However, the validity of the diagnostic codes for psychotic disorders and affective disorders and the impact of the coding system transition in the NHIRD have not been examined. To fill this gap in the research, we developed a standardized process for reviewing the medical records of psychiatric inpatients at one medical center and its branches in suburban or rural areas, which provided an opportunity to examine the validity of psychiatric diagnoses in hospitals across urbanicity levels in the NHIRD. This study aimed to (1) assess the interrater reliability of chart review for psychiatric inpatients among psychiatrists; (2) examine the validity of the diagnostic codes for psychotic disorders and

affective diseases in the claims data submitted to the NHI against the chart review-based diagnoses; and (3) examine whether the change in the coding system from the ICD-9-CM to the ICD-10-CM affected the validity of the diagnostic codes in the claims data.

Materials and Methods

Study Participants

The study participants were psychiatric inpatients aged 18 to 65 years who were admitted in 2015 (using the ICD-9) and 2017 (using the ICD-10), respectively, to National Taiwan University Hospital (NTUH)-Taipei Main Hospital (in a metropolitan area, 68 acute beds) or its branches, including the Hsin-Chu (in an urban area, 36 acute beds), Biomedical Park (in a suburban area, 50 acute beds), and Yun-Lin (in a rural area, 50 acute beds) branches. The selection of psychiatric inpatients was based on the NTUH Integrated Medical Database, in which the discharge notes and the corresponding diagnostic categories in the claims data using the ICD coding system were stored. We included patients admitted in 2015 using the ICD-9-CM and those admitted in 2017 using the ICD-10-CM. We then selected 12 diagnostic categories of psychiatric diseases, including schizophrenia-spectrum disorder (hereafter referred to as schizophrenia (SZ)), schizoaffective disorder, manic/mixed episode with psychotic features (MEP), depressive episode with psychotic features (DEP), substance-induced psychotic disorder, delusional disorder, major depressive disorder without psychotic features (MDDNP), bipolar disorder without psychotic features (BDNP), depression not otherwise specified, cyclothymic disorder, dysthymic disorder, and other psychotic disorders.

The number of inpatients who fulfilled these 12 diagnostic categories was 1596 in 2015 and 1481 in 2017. Considering feasibility, we set the number of patients in each year to 400. For each year, to ensure sufficient representation for evaluating reliability and validity, we set the number to 100 for SZ and 50 for each of the other 4 categories (MEP, DEP, BDNP, and MDDNP). For the remaining 7 categories, if a category's number was 21 or less, all the patients were selected; for the remaining categories, the selection was proportional to their sizes. The total number and selected number of inpatients in different diagnostic categories and the corresponding codes in both the ICD-9 and the ICD-10 are provided in Table S1.

When an inpatient was discharged, the attending psychiatrist wrote a discharge note, in which compatible diagnoses in terms of the disease name were provided. At the end of each month, a group of health specialists in the hospital administrative center helped prepare the claims data, including the assignment of diagnostic codes in the ICD-9-CM or ICD-10-CM, on the basis of the information written in the discharge note. For this study, both the information of each patient's medical records and the corresponding NHRI claims data were extracted into separate datasets that could be reviewed by participating psychiatrists or the study researchers via the software Research Electronic Data Capture (REDCap).[32,33] This study was approved by the Research Ethics Committee of the National Taiwan University Hospital (NTUH-REC no. 201909027RINA).

Procedures

To standardize the chart-review process, we designed a REDCap checklist covering all the diagnostic criteria for the 12 selected diagnoses according to both the ICD system (ICD-9-CM and ICD-10-CM) and the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), which is used in daily clinical teaching and practice in psychiatry nationwide in Taiwan. Psychiatrists who participated in this study were asked to fill out the REDCap checklist by reviewing the discharge notes. Following the privacy protection rule of the NTUH, the checklist of a patient cannot be linked directly to NHRI claims data using patient identifiers. As a result, we used a probabilistic record linkage method by simultaneously matching the medical record number and admission date to identify the records of the same individual across the medical record and claims datasets. This approach, widely used in administrative health research, has been shown to achieve acceptable accuracy when unique identifiers are not directly linkable.

Interrater Reliability

We started with a reliability assessment among 10 psychiatrists with a focus on five core diagnostic categories, including SZ, MEP, DEP, BDNP, and MDDNP. For this reliability assessment, 50 inpatients were selected from the list of 800 inpatients, with a balance in psychotic versus nonpsychotic disorders. Two inpatients were excluded from the interrater reliability assessment after being identified as having intellectual disabilities, which may involve different diagnostic considerations. The final sample included 48 inpatients. The allocation of these inpatients to psychiatrists (n = 10) at different sites is shown in Table S2.

Each medical record was individually evaluated by four psychiatrists, with a total of 192 rating records (Table S3). All the raters were blinded to the clinical diagnoses to avoid having their judgment influenced. After the interrater reliability was analyzed, for those patients whose chart review-based diagnoses differed, the participating psychiatrists discussed the case in detail to reach a consensus, and the format or procedures of the checklist were revised if indicated.

Validity Assessment

After the completion of the interrater reliability assessment, the remaining 750 medical records were used for validity assessments, of which 23 inpatients were excluded from subsequent analyses because they met the criteria for mental retardation or intellectual disability. A total of 16 psychiatrists reviewed the discharge notes of 727 patients. Each inpatient was reviewed by one psychiatrist, and the number of inpatients per psychiatrist ranged from 18 to 64 (more details are provided in Table S3).

Statistical Analyses

The interrater reliabilities in the diagnosis of major mental disorders were evaluated using the intraclass correlation coefficient reliability (ICCR), and their point estimates and 95% confidence intervals (CIs) were calculated using SAS statistical package version 9.4 (SAS Institute, Cary, NC, USA) via a one-way random effects (k = 4) model, since each subject was rated by a different set of 4 raters.[34,35] An ICCR of less than 0.40 was interpreted as poor, 0.40–0.59 as fair, 0.60–0.74 as good, and 0.75–1.00 as excellent.[36,37]

The validity of the psychiatric diagnostic codes recorded in the claims data was assessed against the psychiatrist's diagnosis after reviewing the discharge note, which was treated as the gold standard. Hence, the validity of the diagnostic codes was expressed as (1) the positive predictive value (PPV), the conditional probability of a diagnosis in the review-based diagnosis given that it appeared in the claims data, and (2) the sensitivity, the conditional probability of a diagnosis recorded in the NHIRD given that it appeared in the review-based diagnosis. Since few studies have investigated the acceptable cutoffs for the PPV or sensitivity, we referred to the results from previous studies on quantitative bias analysis, with a PPV of 70%–80% or greater indicating a high-performing algorithm.[18,38,39]

Results

Patient Characteristics

The sociodemographic and clinical characteristics of the 48 patients included in the interrater reliability study and the remaining 727 patients included in the validation study are presented in Table 1. Both samples were not different in terms of sex, age group, year of hospitalization, presence of psychotic features, age, age at onset, or length of index admission except at the hospital site, with the Biomedical Park branch contributing a lower proportion than the other hospitals did in the validation study.

Interrater Reliability Among the Psychiatrists

All of the diagnostic codes of the 48 inpatients, with each patient's codes assigned by 4 out of 10 psychiatrists, are displayed in Table S4. The ICCRs (95% CI) for the individual diagnostic categories are shown in Table 2. Among the five core categories of diagnoses, four had an ICCR > 0.6 (0.72 for SZ, 0.70 for MEP, 0.69 for BDNP, and 0.62 for MDDNP), indicating good reliability, whereas DEP had an ICCR of 0.47, likely due to the small number of cases. Moreover, the ICCRs of the three broad categories of diagnoses all indicated good reliability, ie, 0.64 for common psychotic disorders, 0.74 for bipolar disorders, and 0.60 for major depressive disorders.

Table 1 Sociodemographic and Clinical Characteristics of the Patients with Psychotic Disorders or Affective Disorders Selected for the Interrater Reliability and Validity Evaluations, Respectively, in This Study

Variable	Sample for Reliability (N = 48)	Sample for Validity (N = 727)	Group Comparison[a]
	n (%)	*n (%)*	*P Value*
Sex			0.60
Male	21 (43.8)	290 (39.9)	
Female	27 (56.2)	437 (60.1)	
Age group			0.39
18–30	7 (14.6)	110 (15.1)	
31–40	12 (25.0)	176 (24.2)	
41–50	9 (18.8)	210 (28.9)	
51–65	20 (41.7)	231 (31.8)	
Hospital site			
Taipei Main Hospital	16 (33.3)	263 (36.2)	0.031
Hsin-Chu branch	11 (22.9)	181 (24.9)	
Biomedical Park branch	11 (22.9)	70 (9.64)	
Yun-Lin branch	10 (20.8)	213 (29.3)	
Year of hospitalization			0.72
2015	23 (47.9)	368 (50.6)	
2017	25 (52.1)	359 (49.4)	
Presence of psychotic features			0.72
Psychotic disorders	28 (58.3)	443 (60.9)	
Affective disorders without psychotic features	20 (41.7)	284 (39.1)	
	Mean (SD)	*Mean (SD)*	*P value*
Age (years)	45.5 (12.7)	43.5 (12.0)	0.28
Age of onset (years)	30.1 (13.9)	32.2 (13.3)	0.28
Length of the index admission (days)	76.6 (248.8)	100.4 (303.2)	0.60

Notes: [a]Chi-square test for categorical variables and *t* test for continuous variables.

Table 2 Interrater Reliability Evaluation of 48 Patients, with Each Patient Being Rated by Four Psychiatrists from a Pool of 10 Psychiatrists Affiliated with National Taiwan University Hospital[a]

Diagnosis	No. of Cases	No. of Noncases	ICCR (95% CI)
Narrow category of diagnosis			
1. Schizophrenia (SZ)	15	33	0.72 (0.63–0.79)
2. Manic/mixed episode with psychotic features (MEP)	8	40	0.70 (0.60–0.77)
3. Depressive episode with psychotic features (DEP)	5	43	0.47 (0.33–0.59)
4. Bipolar disorder without psychotic features (BDNP)	8	40	0.69 (0.59–0.76)
5. Major depressive disorder without psychotic features (MDDNP)	12	36	0.62 (0.50–0.71)
Broad category of diagnosis			
Psychotic disorders (1+2+3)	28	20	0.64 (0.54–0.73)
Bipolar disorders (2+4)	16	32	0.74 (0.66–0.80)
Major depressive disorders (3+5)	17	31	0.60 (0.48–0.69)

Notes: [a] The allocation of the medical records of the participating psychiatrists and the sources of the medical records are detailed in Table S3.
Abbreviations: ICCR, intraclass correlation coefficient reliability; CI, confidence interval.

Validity

Comparing the review-based diagnoses with the diagnostic codes in the NHIRD, Table 3 displays the number of joint events (present in both sources) and marginal events (present in one source), which can be used to derive the diagnoses-specific PPVs and sensitivities. Among the five core diagnoses, the PPVs of two diagnoses were > 0.70 (0.94 for SZ and

Table 3 The Validity of the NHIRD on the Basis of the Medical Records Reviewed by Psychiatrists in Narrow and Broad Disease Categories (n = 727)[a]

Categories	Medical Record (+) & NHIRD (+)	NHIRD (+)	Medical Record (+)	PPV[b] (95% CI)	Sensitivity[c] (95% CI)
Narrow category of diagnosis					
1. Schizophrenia (SZ)	169	180	202	0.94 (0.90–0.97)	0.84 (0.79–0.89)
2. Manic/mixed episode with psychotic features (MEP)	58	86	82	0.67 (0.58–0.77)	0.71 (0.61–0.81)
3. Depressive episode with psychotic features (DEP)	54	88	69	0.61 (0.51–0.72)	0.78 (0.69–0.88)
4. Bipolar disorder without psychotic features (BDNP)	52	89	82	0.58 (0.48–0.69)	0.63 (0.53–0.74)
5. Major depressive disorder without psychotic features (MDDNP)	69	89	114	0.78 (0.69–0.86)	0.61 (0.52–0.70)
6. Schizoaffective disorder	32	54	48	0.59 (0.46–0.72)	0.67 (0.53–0.80)
7. Substance-induced psychotic disorder	14	16	34	0.88 (0.71–1.00)	0.41 (0.25–0.58)
8. Delusional disorder	13	19	14	0.68 (0.48–0.89)	0.93 (0.79–1.00)
9. Depression not otherwise specified	12	47	27	0.26 (0.13–0.38)	0.44 (0.26–0.63)
10. Cyclothymic disorder	1	2	1	0.50 (0.00–1.00)	1.00 (1.00–1.00)
11. Dysthymic disorder	5	29	14	0.17 (0.03–0.31)	0.36 (0.11–0.61)
12. Other psychotic disorders	9	28	15	0.32 (0.15–0.49)	0.60 (0.35–0.85)
13. Others	0	0	25	n.a.	
Broad category of diagnosis					
Schizophrenia/schizoaffective disorder (1+6)	221	234	250	0.94 (0.92–0.97)	0.88 (0.84–0.92)
Psychotic disorders (1+2+3+6+7+8)	388	443	449	0.88 (0.85–0.91)	0.86 (0.83–0.90)
Bipolar disorders (2+4)	144	175	164	0.82 (0.77–0.88)	0.84 (0.79–0.89)
Major depressive disorders (3+5)	144	177	183	0.81 (0.76–0.87)	0.79 (0.73–0.85)

Notes: [a]Linked cases between the NHIRD and medical records; 727 cases underwent medical record review. [b]PPV: The proportion of records in the NHIRD that were also found in the medical records. [c]Sensitivity: Patients who received a diagnosis via medical records and were also present in the NHIRD.

0.78 for MDDNP), and those of the remaining 3 were moderate (0.67 for MEP, 0.61 for DEP, and 0.58 for BDNP). For the remaining narrow categories of diagnoses, the PPVs of three psychotic disorders were high (0.88 for substance-induced psychotic disorder) or moderate (0.58 for schizoaffective disorder and 0.68 for delusional disorder), whereas those of the other categories of low prevalence were ≤ 0.50. With respect to the sensitivities, three of the five core diagnoses with psychotic features had a sensitivity value > 0.70 (0.84 for SZ, 0.71 for MEP, and 0.78 for DEP), the other two without psychotic features had moderate sensitivity (0.63 for BDNP and 0.61 for MDDNP), and the remaining categories of low prevalence had modest estimates. Nevertheless, for the four broad categories of diagnoses, both the PPVs and sensitivities were > 0.70, ie, 0.94 and 0.88 for SZ/schizoaffective disorder, 0.88 and 0.86 for psychotic disorders, 0.82 and 0.84 for bipolar disorders, and 0.81 and 0.79 for major depressive disorders.

We then examined whether the claims data of the patients from the medical center might differ from those from the branch hospitals (Table S5). In general, the validity indices of the patients from both types of hospitals were similar, ie, the point estimate in one type was within the 95% CI of the counterpart in the other type, except those of bipolar disorder without psychotic features in patients from the branch hospitals, which were lower than their counterparts in the main hospital. Nevertheless, for the three broad categories of diagnoses, the validity indices of patients from both types of hospitals were all similarly high, with PPVs and sensitivities ≥ 0.79.

Validity Stratified by ICD Coding Versions

When the validity indices of five core categories of diagnoses were compared between patients admitted in 2015 (using the ICD-9) and those admitted in 2017 (using the ICD-10), the latter showed better validity indices in almost every diagnostic category (Table 4). For example, patients admitted in 2017 had greater sensitivity than patients admitted in 2015 for SZ (0.91 vs 0.78), MEP (0.74 vs 0.67), and BDNP (0.73 vs 0.55) and a greater PPV for MEP (0.78 vs 0.67) and BDNP (0.66 vs 0.51). However, all the validity indices of the three broad categories of diagnoses were similarly high, except for the lower sensitivity of major depressive disorders in the patients admitted in 2017 (0.74, 95% CI: 0.65–0.83) than in the patients admitted in 2015 (0.84, 95% CI: 0.76–0.92).

Table 4 The Validity of the NHIRD in Narrow and Broad Diagnostic Categories Among the Years of Admission in 2015 (Using the ICD-9) and 2017 (Using the ICD-10), Respectively, on the Basis of the Medical Records Reviewed by Psychiatrists[a]

Categories	Patients Admitted in 2015 (Using the ICD-9; N = 368)					Patients Admitted in 2017 (Using the ICD-10; N = 359)				
	MR (+) & NHIRD (+)	NHIRD (+)	MR (+)	PPV[b] (95% CI)	Sensitivity[c] (95% CI)	MR (+) & NHIRD (+)	NHIRD (+)	MR (+)	PPV[b] (95% CI)	Sensitivity[c] (95% CI)
Individual classification of diseases										
1. Schizophrenia (SZ)	83	89	107	0.93 (0.88–0.98)	0.78 (0.70–0.85)	86	91	95	0.95 (0.90–0.99)	0.91 (0.85–0.96)
2. Manic/mixed episode with psychotic features (MEP)	26	45	39	0.58 (0.43–0.72)	0.67 (0.52–0.81)	32	41	43	0.78 (0.65–0.91)	0.74 (0.61–0.87)
3. Depressive episode with psychotic features (DEP)	27	45	34	0.60 (0.46–0.74)	0.79 (0.66–0.93)	27	43	35	0.63 (0.48–0.77)	0.77 (0.63–0.91)
4. Bipolar disorder without psychotic features (BDNP)	23	45	42	0.51 (0.37–0.66)	0.55 (0.40–0.70)	29	44	40	0.66 (0.52–0.80)	0.73 (0.59–0.86)
5. Major depressive disorder without psychotic features (MDDNP)	34	45	53	0.76 (0.63–0.88)	0.64 (0.51–0.77)	35	44	61	0.80 (0.68–0.91)	0.57 (0.45–0.70)
Broad classification of diseases										
Psychotic disorders (1+2+3)	194	224	230	0.87 (0.82–0.91)	0.84 (0.80–0.89)	194	219	219	0.89 (0.84–0.93)	0.89 (0.84–0.93)
Bipolar disorders (2+4)	71	90	81	0.79 (0.70–0.87)	0.88 (0.80–0.95)	73	85	83	0.86 (0.78–0.93)	0.88 (0.81–0.95)
Major depressive disorders (3+5)	73	90	87	0.81 (0.73–0.89)	0.84 (0.76–0.92)	71	87	96	0.82 (0.73–0.90)	0.74 (0.65–0.83)

Notes: [a]Linked cases between the NHIRD and medical records; 727 cases underwent medical record review. [b]PPV: The proportion of records in the NHIRD that were also found in the medical records. [c]Sensitivity: Patients who received a diagnosis via medical records and were also present in the NHIRD.

Abbreviations: MR, medical record; NHIRD, National Health Insurance Research Database; PPV, positive predictive value; CI, confidence interval.

https://doi.org/10.2147/CLEP.S522618

Discussion

In this study, we established checklists for reviewing discharge notes to evaluate the interrater reliability among psychiatrists and then used review-based diagnoses to evaluate the validity of claims-based diagnostic codes. Among 48 inpatients in five core categories (SZ, MEP, DEP, BDNP, and MDDNP), good interrater reliability was achieved. In another 727 inpatients with psychotic and affective disorders, the PPV and sensitivity of common diagnoses in the narrow category (eg, SZ) or broad category (eg, psychotic disorders, bipolar disorders, and major depressive disorders) were high-performing (≥ 0.70), whereas those of the diagnoses of low prevalence were modest. Intriguingly, the validity indices of claims-based diagnoses using the ICD-10-CM tended to be better than those using the ICD-9-CM.

Despite the absence of "true negatives" in our sample, which limits the magnitude of the ICCR,[37] our results are similar to those of previous studies on SZ and other psychotic disorders in Danish patients (a kappa of 0.60 between clinical and algorithm-derived diagnoses)[40] and major depression (a kappa of 0.66 in a mixed sample of patients and nonpatients in the Netherlands).[41] Although not directly comparable, the majority of previous studies included in a systematic review of agreement between the source data and reference standards for most diagnostic categories were found to have a median kappa of approximately 0.5 for the diagnoses of schizophrenia, schizophrenia spectrum disorders, depressive disorder, and bipolar disorder.[12] Our use of REDCap-based checklists covering all selected diagnostic criteria might help decrease the variability in reaching a diagnosis among different psychiatrists, even from different hospitals.[42]

Since our selection of patients for validity assessment was based on random sampling with truncation for common diagnoses and enrichment for less common diagnoses, we were able to estimate both the PPV and sensitivity, which are important for future correction for biases incurring false-positive errors and false-negative errors.[18,38,39] Among the 12 narrow diagnostic categories, the PPVs of SZ, MDDNP, and substance-induced psychotic disorder met the high-performing criterion of 70%-80%. Moreover, the sensitivities of the diagnostic codes of SZ, MEP, DEP, and delusional disorder were also high. There were large discrepancies in performance for substance-induced psychosis (0.88 for PPV and 0.41 for sensitivity) and delusional disorder (0.68 for PPV and 0.93 for sensitivity), indicating that the coding by health specialists tended to incur false-negative errors for the former and false-positive errors for the latter. Nevertheless, when some narrow categories of diagnoses were combined into four broad categories (SZ/schizoaffective disorder, psychotic disorders, bipolar disorders, and major depressive disorders), all had high PPVs and sensitivities (≥ 79%). For comparison, previous validation studies on schizophrenia reported a high-performing PPVs, ranging from 91%[21,22] to 78%[43] and 58.3%,[20] and sensitivities, ranging from 71%[22] to 82.4%.[20] A previous validation study on bipolar disorder reported a high-performing PPV (72%) and sensitivity (84%).[22] However, previous validation studies on depressive disorder reported relatively low PPVs, ranging from 54.4%[44] to 70%,[22] and sensitivities, ranging from 52.6%[44] to 83%.[22] The relatively low validity indices for depressive disorder may reflect diagnostic complexity or the presence of psychiatric or medical comorbidities among inpatients, which can complicate accurate coding and clinical differentiation.

Furthermore, the validity indices for the diagnostic codes using the ICD-10-CM in 2017 tended to be slightly higher than those using the ICD-9-CM in 2015. A likely explanation is that the participating hospitals had requested that clinicians and supporting staff receive intensive training for the coding transition.[45] Another explanation is that the ICD-10-CM codes require more specific details than the ICD-9-CM codes do. Our findings are important in deciphering the inconsistent prevalences of psychotic and affective disorders derived from these two coding systems.[46]

This study has several limitations. First, since our patients were selected from inpatients in one medical center and its branches, our findings might not be generalizable to patients in other settings, eg, patients in other types of hospitals. Moreover, individuals with different socioeconomic backgrounds may have differential access to these hospitals, which could introduce selection bias. Second, the review-based diagnoses were based solely on the discharge notes in the existing electronic medical record databases, and not the entire medical charts or diagnostic interviews. Third, the number of patients with a psychiatric diagnosis of low prevalence was not sufficient to make precise estimates of the validity indices. Fourth, this study did not incorporate selection weights since our focus was on diagnostic validity rather than estimating population-level parameters. Our stratified sampling with purposive oversampling of key diagnostic categories, though appropriate for evaluating diagnostic validity, limits the generalizability of our results to the full inpatient population. Lastly, because patients were selected based on claims-based diagnoses, potential "true cases"

without recorded diagnoses may have been missed, possibly leading to an overestimation of sensitivity. Future studies should consider diagnosis-independent sampling and weighted analyses to improve generalizability and accuracy.

Conclusion

To our knowledge, this is the first study to validate psychiatric diagnoses in Taiwan's NHIRD using a structured chart review, providing an important foundation for future psychiatric epidemiological research in this population. We found that the validity of diagnostic codes for psychotic disorders and affective disorders varied with the breadth of the diagnostic categories in the NHIRD. For high-performing validity indices, the diagnostic codes of SZ or broad categories, such as psychotic disorders, bipolar disorders, and major depressive disorders, are recommended. Intensive training for the coding plus the specific details requested by the ICD-10 may increase the validity of the claims-based databases for psychotic and affective disorders.

Data Sharing Statement

The datasets analyzed for the current study are not publicly available due to the requirement of obtaining official permission to access the data but are available from the corresponding author upon reasonable request.

Ethical Statement

This study was approved by the Research Ethics Committee of the National Taiwan University Hospital (NTUH-REC no. 201909027RINA).

Author Contributions

All authors made a significant contribution to the reported work, whether in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas. All authors took part in drafting, revising or critically reviewing the article; they all gave approval for the final version to be published and all have agreed on the journal to which the article is to be submitted. Furthermore, all authors agree to be accountable for all aspects of the work.

Funding

This work was supported by grants from the National Health Research Institutes, Taiwan (09A1-PP10) and the Ministry of Science and Technology, Taiwan (109-2314-B-002-172-MY3). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Disclosure

Dr Wei-Lieh Huang reports grants from National Health Research Institutes, National Science and Technology Council, and National Taiwan University Hospital Yunlin Branch; personal fees from Janssen, Servier, Boehringer Ingelheim, Pfizer/Viatris, Sumitomo, and Otsuka, outside the submitted work. The authors report no other conflicts of interest in this work.

References

1. Perala J, Suvisaari J, Saarni SI. et al. Lifetime prevalence of psychotic and bipolar I disorders in a general population. *Arch Gen Psychiatry*. 2007;64(1):19–28. doi:10.1001/archpsyc.64.1.19
2. Lieberman JA, First MB. Psychotic Disorders. *N Engl J Med*. 2018;379(3):270–280. doi:10.1056/NEJMra1801490
3. GBD. Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Psychiatry*. 2022;9(2):137–150. doi:10.1016/s2215-0366(21)00395-3.
4. Kessler RC, Birnbaum H, Demler O, et al. The prevalence and correlates of nonaffective psychosis in the National Comorbidity Survey Replication (NCS-R). *Biol Psychiatry*. 2005;58(8):668–676. doi:10.1016/j.biopsych.2005.04.034
5. Moreno-Küstner B, Martín C, Pastor L. Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. *PLoS One*. 2018;13(4):e0195687. doi:10.1371/journal.pone.0195687
6. Melfi CA, Croghan TW. Use of claims data for research on treatment and outcomes of depression care. *Med Care*. 1999;37(4 Suppl Lilly):As77–80. doi:10.1097/00005650-199904001-00010

7. Frayne SM, Miller DR, Sharkansky EJ, et al. Using administrative data to identify mental illness: what approach is best? *Am J Med Qual*. 2009;25(1):42–50. doi:10.1177/1062860609346347
8. Kisely S, Lin E, Lesage A, et al. Use of administrative data for the surveillance of mental disorders in 5 provinces. *Can J Psychiatry*. 2009;54(8):571–575. doi:10.1177/070674370905400810
9. Steele LS, Glazier RH, Lin E, Evans M. Using administrative data to measure ambulatory mental health service provision in primary care. *Med Care*. 2004;42(10):960–965. doi:10.1097/00005650-200410000-00004
10. Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. *J Epidemiol Community Health*. 2014;68(3):283–287. doi:10.1136/jech-2013-202744
11. Cadarette SM, Wong L. An introduction to health care administrative data. *Can J Hosp Pharm*. 2015;68(3):232–237. doi:10.4212/cjhp.v68i3.1457
12. Davis KAS, Sudlow CLM, Hotopf M. Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. *BMC Psychiatry*. 2016;16(1):263. doi:10.1186/s12888-016-0963-x
13. Lin LY, Warren-Gash C, Smeeth L, Chen PC. Data resource profile: the National Health Insurance Research Database (NHIRD). *Epidemiol Health*. 2018;40:e2018062. doi:10.4178/epih.e2018062
14. Hsieh CY, Su CC, Shao SC, et al. Taiwan's National Health Insurance Research Database: past and future. *Clin Epidemiol*. 2019;11:349–358. doi:10.2147/clep.S196293
15. Lin Y-H, Wu C-S, Liu -C-C, Kuo P-H, Chan H-Y, Chen WJ. Comparative effectiveness of antipsychotics in preventing readmission for first-admission schizophrenia patients in national cohorts from 2001 to 2017 in Taiwan. *Schizoph Bull*. 2022;48(4):785–794. doi:10.1093/schbul/sbac046
16. Chiang C-L, Chen P-C, Huang L-Y, et al. Time trends in first admission rates for schizophrenia and other psychotic disorders in Taiwan, 1998–2007: a 10-year population-based cohort study. *Soc Psychiatry Psychiatr Epidemiol*. 2017;52(2):163–173. doi:10.1007/s00127-016-1326-0
17. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol*. 2011;64(8):821–829. doi:10.1016/j.jclinepi.2010.10.006
18. Lanes S, Beachler DC. Validation to correct for outcome misclassification bias. *Pharmacoepidemiol Drug Saf*. 2023;32(6):700–703. doi:10.1002/pds.5601
19. Byrne N, Regan C, Howard L. Administrative registers in psychiatric research: a systematic review of validity studies. *Acta Psychiatr Scand*. 2005;112(6):409–414. doi:10.1111/j.1600-0447.2005.00663.x
20. Kurdyak P, Lin E, Green D, Vigod S. Validation of a population-based algorithm to detect chronic psychotic Illness. *Can J Psychiatry*. 2015;60(8):362–368. doi:10.1177/070674371506000805
21. Svensson E, Voldsgaard I, Haller LG, Baandrup L. Validation study of the population included in the Danish Schizophrenia Registry. *Dan Med J*. 2019;66(10):1.
22. Davis KAS, Bashford O, Jewell A, et al. Using data linkage to electronic patient records to assess the validity of selected mental health diagnoses in English Hospital Episode Statistics (HES). *PLoS One*. 2018;13(3):e0195002. doi:10.1371/journal.pone.0195002
23. Vernal DL, Stenstrøm AD, Staal N, et al. Validation study of the early onset schizophrenia diagnosis in the Danish Psychiatric Central Research Register. *Eur Child Adolesc Psychiatry*. 2018;27(8):965–975. doi:10.1007/s00787-017-1102-z
24. Cheng CL, Kao YH, Lin SJ, Lee CH, Lai ML. Validation of the National Health Insurance Research Database with ischemic stroke cases in Taiwan. *Pharmacoepidemiol Drug Saf*. 2011;20(3):236–242. doi:10.1002/pds.2087
25. Hsieh CY, Chen CH, Li CY, Lai ML. Validating the diagnosis of acute ischemic stroke in a National Health Insurance claims database. *J Formos Med Assoc*. 2015;114(3):254–259. doi:10.1016/j.jfma.2013.09.009
26. Cheng C-L, Lee C-H, Chen P-S, Y-H L, S-J L, Yang Y-HK. Validation of acute myocardial infarction cases in the National Health Insurance Research Database in Taiwan. *J Epidemiol*. 2014;24(6):500–507. doi:10.2188/jea.JE20140076
27. Cheng CL, Chien HC, Lee CH, Lin SJ, Yang YH. Validity of in-hospital mortality data among patients with acute myocardial infarction or stroke in National Health Insurance Research Database in Taiwan. *Int J Cardiol*. 2015;201:96–101. doi:10.1016/j.ijcard.2015.07.075
28. Lee MS, Yeh YC, Chang YT, Lai MS. All-cause and cause-specific mortality in patients with psoriasis in Taiwan: a nationwide population-based study. *J Invest Dermatol*. 2017;137(7):1468–1473. doi:10.1016/j.jid.2017.01.036
29. Ho TW, Ruan SY, Huang CT, Tsai YJ, Lai F, Yu CJ. Validity of ICD9-CM codes to diagnose chronic obstructive pulmonary disease from National Health Insurance claim data in Taiwan. *Int J Chron Obstruct Pulmon Dis*. 2018;13:3055–3063. doi:10.2147/copd.S174265
30. Kao WH, Hong JH, See LC, et al. Validity of cancer diagnosis in the National Health Insurance database compared with the linked National Cancer Registry in Taiwan. *Pharmacoepidemiol Drug Saf*. 2018;27(10):1060–1066. doi:10.1002/pds.4267
31. Lu PT, Tsai TH, Lai CC, Chuang LH, Shao SC. Validation of diagnostic codes to identify glaucoma in Taiwan's claims data: a multi-institutional study. *Clin Epidemiol*. 2024;16:227–234. doi:10.2147/clep.S443872
32. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377–381. doi:10.1016/j.jbi.2008.08.010
33. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/j.jbi.2019.103208
34. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemp Clin Trials*. 2012;33(5):869–880. doi:10.1016/j.cct.2012.05.004
35. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163. doi:10.1016/j.jcm.2016.02.012
36. Fleiss JL. *Statistical Methods for Rates and Proportions*. 2nd edn ed. New York: John Wiley & Sons; 1981.
37. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6(4):284–290. doi:10.1037/1040-3590.6.4.284
38. Newcomer SR, Xu S, Kulldorff M, Daley MF, Fireman B, Glanz JM. A primer on quantitative bias analysis with positive predictive values in research using electronic health data. *J Am Med Inform Assoc*. 2019;26(12):1664–1674. doi:10.1093/jamia/ocz094
39. Weinstein EJ, Ritchey ME, Lo Re III V. Core concepts in pharmacoepidemiology: validation of health outcomes of interest within real-world healthcare databases. *Pharmacoepidemiol Drug Saf*. 2023;32(1):1–8. doi:10.1002/pds.5537

40. Jakobsen KD, Frederiksen JN, Hansen T, Jansson LB, Parnas J, Werge T. Reliability of clinical ICD-10 schizophrenia diagnoses. *Nordic J Psychiatry*. 2005;59(3):209–212. doi:10.1080/08039480510027698
41. Lobbestael J, Leurgans M, Arntz A. Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID I) and Axis II Disorders (SCID II). *Clin Psychol Psychother*. 2011;18(1):75–79. doi:10.1002/cpp.693
42. Berendsen S, van der Paardt JW, Van HL, et al. Staging and profiling for schizophrenia spectrum disorders: inter-rater reliability after a short training course. *Prog Neuropsychopharmacol Biol Psychiatry*. 2020;99:109856. doi:10.1016/j.pnpbp.2019.109856
43. Pihlajamaa J, Suvisaari J, Henriksson M, et al. The validity of schizophrenia diagnosis in the Finnish Hospital Discharge Register: findings from a 10-year birth cohort sample. *Nord J Psychiatry*. 2008;62(3):198–203. doi:10.1080/08039480801983596
44. Townsend L, Walkup JT, Crystal S, Olfson M. A systematic review of validated methods for identifying depression using administrative data. *Pharmacoepidemiol Drug Saf*. 2012;21 Suppl 1(S1):163–173. doi:10.1002/pds.2310
45. Stewart CC, Lu CY, Yoon TK, et al. Impact of ICD-10-CM Transition on Mental Health Diagnoses Recording. *EGEMS*. 2019;7(1):14. doi:10.5334/egems.281
46. Hsu MC, Wang CC, Huang LY, Cy L, Fj L, Toh S. Effect of ICD-9-CM to ICD-10-CM coding system transition on identification of common conditions: an interrupted time series analysis. *Pharmacoepidemiol Drug Saf*. 2021;30(12):1653–1674. doi:10.1002/pds.5330