Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Validation of Stroke Risk Factors in Patients with Acute Ischemic Stroke, Transient Ischemic Attack, or Intracerebral Hemorrhage on Taiwan's National Health Insurance Claims Data

Meng-Tsang Hsieh[1–3], Cheng-Yang Hsieh[4,5], Tzu-Tung Tsai[1], Sheng-Feng Sung[6,7]

[1]Stroke Center and Department of Neurology, E-Da Hospital, Kaohsiung, Taiwan; [2]School of Medicine, College of Medicine, I-Shou University, Kaohsiung, Taiwan; [3]Institute of Clinical Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan; [4]Department of Neurology, Tainan Sin Lau Hospital, Tainan, Taiwan; [5]School of Pharmacy, Institute of Clinical Pharmacy and Pharmaceutical Sciences, College of Medicine, National Cheng Kung University, Tainan, Taiwan; [6]Division of Neurology, Department of Internal Medicine, Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi City, Taiwan; [7]Department of Nursing, Min-Hwei Junior College of Health Care Management, Tainan, Taiwan

Correspondence: Sheng-Feng Sung, Division of Neurology, Department of Internal Medicine, Ditmanson Medical Foundation Chia-Yi Christian Hospital, 539 Zhongxiao Road, East District, Chiayi City, 60002, Taiwan, Tel +886 5 276 5041 ext 7283, Fax +886 5 278 4257, Email richard.sfsung@gmail.com

Purpose: Taiwan has changed the coding system to the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) coding since 2016. This study aimed to determine the optimal algorithms for identifying stroke risk factors in Taiwan's National Health Insurance (NHI) claims data.

Patients and Methods: We retrospectively enrolled 4538 patients hospitalized for acute ischemic stroke (AIS), transient ischemic attack (TIA), or intracerebral hemorrhage (ICH) from two hospitals' stroke registries, which were linked to NHI claims data. We developed several algorithms based on ICD-10-CM diagnosis codes and prescription claims data to identify hypertension, diabetes, hyperlipidemia, atrial fibrillation (AF), and ischemic heart disease (IHD) using registry data as the reference standard. The agreement of risk factor status between claims and registry data was quantified by calculating the kappa statistic.

Results: According to the registry data, the prevalence of hypertension, diabetes, hyperlipidemia, AF, and IHD among all patients was 77.5%, 41.5%, 47.9%, 12.1%, and 7.1%, respectively. In general, including diagnosis codes from prior inpatient or outpatient claims to those from the stroke hospitalization claims improved the agreement. Incorporating prescription data could improve the agreement for hypertension, diabetes, hyperlipidemia, and AF, but not for IHD. The kappa values of the optimal algorithms were 0.552 (95% confidence interval 0.524–0.580) for hypertension, 0.802 (0.784–0.820) for diabetes, 0.514 (0.490–0.539) for hyperlipidemia, 0.765 (0.734–0.795) for AF, and 0.518 (0.473–0.564) for IHD.

Conclusion: Algorithms using diagnosis codes alone are sufficient to identify hypertension, AF, and IHD whereas algorithms combining both diagnosis codes and prescription data are more suitable for identifying diabetes and hyperlipidemia. The study results may provide a reference for future studies using Taiwan's NHI claims data.

Keywords: administrative claims data, diagnosis, ICD-10-CM, stroke, risk factors

Introduction

Over the past three decades, stroke has been among the leading causes of death and disability globally, while the absolute numbers of incident and prevalent strokes have increased by 70% and 85%, respectively.[1] The 2021 stroke prevention guideline recommends adequate diagnostic evaluation to identify modifiable risk factors so as to target therapy to the identified risk factors to prevent stroke recurrence.[2] Approximately 80% of first-time strokes have at least one modifiable stroke risk factor,[3] while these modifiable risk factors account for about 90% of the population attributable risk of all strokes worldwide.[4] Despite numerous existing prevention guidelines and initiatives, the prevalence of stroke risk factors

Received: 10 December 2021
Accepted: 8 March 2022
Published: 17 March 2022

is still increasing.[5] Continuous epidemiological surveillance of stroke and its risk factors is thus important for policy makers intending to relieve the global burden of stroke.

Administrative claims data, generated in the billing process of healthcare services, have been widely used in stroke research, including epidemiology and outcome studies.[6–8] They are commonly used for monitoring disease burden and occurrence owing to their relatively low cost and wide coverage of population.[9] Because of the single-payer universal health insurance in Taiwan, virtually all residents' encounters with the healthcare system are included in Taiwan's National Health Insurance (NHI) claims data. The NHI Administration has been regularly releasing the claims data as the National Health Insurance Research Database (NHIRD). Consequently, this data source is ideal for answering population-level research questions.[10,11] However, as claims data are not primarily collected for research purposes, their validity for research needs to be established before using them for clinical epidemiological studies.

Claims-based studies generally ascertain diseases and medical conditions based on the International Classification of Diseases (ICD) diagnosis codes.[9] The validity of ICD-9 and ICD-10 codes for identifying common stroke risk factors has been well examined in prior studies.[12–14] Nevertheless, generalization of those findings to other areas should not be taken for granted because different healthcare systems may have different coding practices. Previously, algorithms using ICD-9 codes to identify stroke risk factors have been validated on the NHIRD.[15] However, Taiwan has changed the coding system to the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) coding since 2016.[10] Moreover, a direct use of ICD codes validated in other healthcare systems was found to cause significant changes in the prevalence of some common medical conditions during the transition from ICD-9 to ICD-10, suggesting that the use of ICD codes not validated locally may lead to incorrect assignment of disease status.[16] In order to facilitate stroke-related research using the NHIRD, a thorough validation of the ICD-10-CM algorithms for identifying stroke risk factors is required. Therefore, in this study, we aimed to determine the optimal algorithms for identifying common stroke risk factors in the NHIRD, including hypertension, diabetes, hyperlipidemia, atrial fibrillation (AF), and ischemic heart disease (IHD).

Patients and Methods

Data Source and Data Linkage

The study data were obtained from the stroke registries and administrative claims databases of two participating hospitals. The Ditmanson Medical Foundation Chia-Yi Christian Hospital is a regional teaching hospital with approximately 650 annual stroke admissions, and the E-Da Hospital is a would-be medical center with approximately 1000 stroke admissions per year. Following the protocol of the nationwide Taiwan Stroke Registry,[17] both hospitals prospectively registered all patients hospitalized for acute ischemic stroke (AIS), transient ischemic attack (TIA), or intracerebral hemorrhage (ICH) within 10 days of symptom onset. Data regarding demographics, medical history, risk factors, medications, interventions, and outcomes were collected by trained study nurses or stroke case managers. Stroke types and risk factors were defined according to the operation manual provided by the Taiwan Stroke Registry (Supplementary Table 1).

The administrative claims databases consisted of all claims data reported to the NHI, including claims for inpatient and outpatient care, physician services, procedures, and prescriptions. Each inpatient and outpatient record contains up to 20 and 5 ICD-10-CM diagnosis codes, respectively. All records for each patient can be linked between the stroke registry and the administrative claim database using a unique patient identifier and the date of admission.

The study protocol was independently approved by the Institutional Review Board of the Ditmanson Medical Foundation Chia-Yi Christian Hospital (IRB2021093) and the Institutional Review Board of the E-Da Hospital (EMRP-109-013). The requirement for informed consent was waived due to the retrospective design. The study data were kept under confidentiality to ensure the privacy of the study participants. This study was conducted in accordance with the Declaration of Helsinki.

Study Sample

All consecutive patients hospitalized between 2018 and 2020 were identified from the stroke registries. Only the first hospitalization was analyzed for patients with multiple hospitalizations. The data extracted from the stroke registries included age, sex, principal discharge diagnosis, and the status of five stroke risk factors, that is, hypertension, diabetes,

hyperlipidemia, AF, and IHD. The claim records for each patient were obtained from the administrative claims databases. Both the diagnosis and medication codes were extracted from the claim record for the index stroke hospitalization, as well as all inpatient and outpatient claim records within two years before the index stroke.

Development of Algorithms

We developed several algorithms using ICD-10-CM diagnosis codes and Anatomical Therapeutic Chemical (ATC) codes for medications to ascertain the status of the five stroke risk factors. Table 1 lists the diagnosis and medication codes related to each stroke risk factor. These algorithms (Figure 1) differed in the numbers of diagnosis codes extracted from the index hospitalization, whether to include inpatient and outpatient claims before the index stroke, the lookback period before the index stroke, and whether to combine prescription data and diagnosis codes with an "AND" or "OR" logic operand. Of note, although a maximum of 20 diagnosis codes can be recorded in each inpatient record, currently the NHIRD released by the NHI Administration contains only the first 5 diagnosis codes. The following illustration demonstrates how these algorithms (left side of Figure 1) were used to evaluate the presence or absence of a stroke risk factor. Take hyperlipidemia for example: algorithm 11 would identify a patient as having hyperlipidemia if both a hyperlipidemia-related ICD code (E78.x) and a hyperlipidemia-related ATC code (C10) were recorded in at least one inpatient or two outpatient claims within 2 years prior to the index hospitalization. On the other hand, algorithm 4 would identify a patient as having hyperlipidemia if a hyperlipidemia-related ICD code was included in the first 5 diagnosis codes of the index hospitalization.

Data Analyses

Continuous variables were summarized with the mean and standard deviation while categorical variables were summarized with counts and percentages. One-way analysis of variance with Bonferroni post-hoc test was used to compare continuous variables, and the chi-squared test was used to compare categorical variables. The presence of the five selected risk factor was determined by applying the developed algorithms on the claim records. The prevalence of each risk factor was estimated based on each algorithm and its 95% confidence interval (CI) was calculated using a binomial probability distribution. Using the status of risk factors recorded in the registry data as the reference standard, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) with 95% CI were calculated. Because registry data are not necessarily the gold standard, kappa statistics were calculated to assess the agreement of risk factor status between the claims and registry data. The degree of agreement was interpreted as follows: "slight" (0.00–0.20), "fair" (0.21–0.40), "moderate" (0.41–0.60), "substantial" (0.61–0.80), and "perfect" (0.81–1.00). The algorithm achieving the highest kappa value for each risk factor was considered the optimal algorithm. All statistical analyses were performed using Stata 15.1 (StataCorp, College Station, Texas). Two-tailed p values of <0.05 were considered statistically significant.

Table 1 Diagnosis and Medication Codes Used for Identifying Stroke Risk Factors

Risk Factor	ICD-10-CM Codes	ATC Codes
Hypertension	I10.x, I11.x, I12.x, I13.x, I14.x, I15.x	C02 (antihypertensives), C03 (diuretics), C07 (beta blocking agents), C08 (calcium channel blockers), C09 (agents acting on the renin-angiotensin system)
Diabetes	E10.x, E11.x, E12.x, E13.x, E14.x	A10 (drugs used in diabetes)
Hyperlipidemia	E78.x	C10 (lipid modifying agents)
AF	I48.x	B01AA (vitamin K antagonists), B01AE (direct thrombin inhibitors), B01AF (direct factor Xa inhibitors), C01BC03 (propafenone), C01BD01 (amiodarone)
IHD	I20.x, I21.x, I22.x, I25.x	C01 (cardiac therapy), C07 (beta blocking agents), C08 (calcium channel blockers), C09 (agents acting on the renin-angiotensin system), C10 (lipid modifying agents)

Abbreviations: AF, atrial fibrillation; ATC, Anatomic Therapeutic Chemical; ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification; IHD, ischemic heart disease.

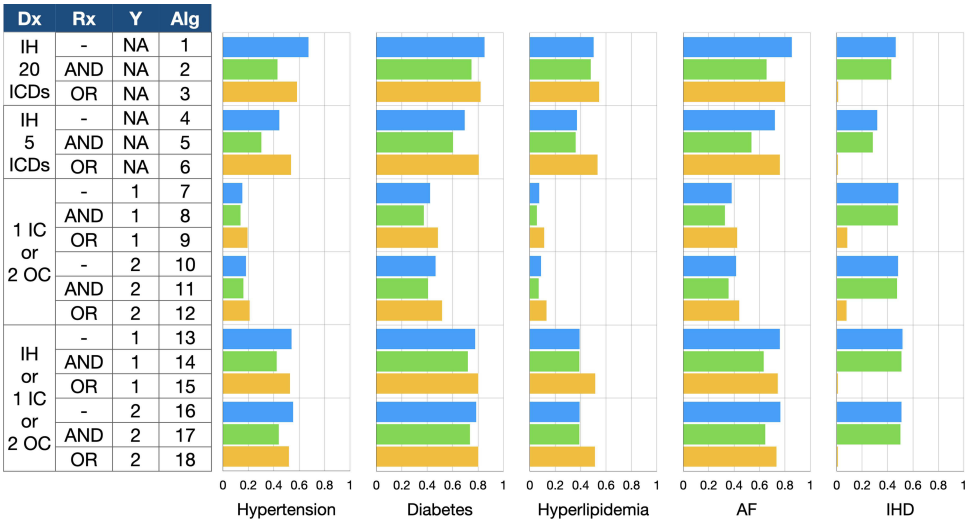

Figure 1 Agreement of diagnosis between National Health Insurance claims data and stroke registry data. X-axis denotes kappa statistics and y-axis represents various algorithms.

Abbreviations: AF, atrial fibrillation; Alg, algorithm; Dx, diagnosis; IC, inpatient claims; ICD, International Classification of Diseases; IH, index hospitalization; IHD, ischemic heart disease; NA, not applicable; OC, outpatient claims; Rx, prescription; Y, years of lookback period.

Results

Characteristics of Study Sample

A total of 4538 patients were identified from the stroke registries, including 3058 with AIS, 575 with TIA, and 905 with ICH. Their mean age was 68.0 ± 13.7 years and 37.9% were female. According to the registry data, the prevalence of hypertension, diabetes, hyperlipidemia, AF, and IHD among all patients was 77.5%, 41.5%, 47.9%, 12.1%, and 7.1%, respectively. The demographics and prevalence of all the five stroke risk factors were significantly different among patient groups (Table 2). Patients with AIS were older than those with TIA ($p = 0.001$) and those with ICH ($p < 0.001$). Patients with TIA were also older than those with ICH ($p < 0.001$). Patients with AIS were more likely to have diabetes, hyperlipidemia, AF, and IHD whereas those with ICH were more likely to have hypertension.

Performance of Algorithms: Overview

Figure 1 depicts the agreement between the claims and registry data for each risk factor based on various algorithms. Overall, the agreement was the highest for diabetes, followed by AF, hypertension, hyperlipidemia, and the lowest for IHD. In general, the algorithm using 20 diagnosis codes from the index hospitalization (algorithms 1) attained higher

Table 2 Characteristics of the Study Sample

Characteristic	Total n = 4538	AIS n = 3058	TIA n = 575	ICH n = 905	*P*
Age, mean (SD)	68.0 (13.7)	69.5 (13.1)	67.4 (13.5)	63.2 (14.6)	<0.001
Female	1722 (37.9)	1152 (37.7)	259 (45.0)	311 (34.4)	<0.001
Hypertension	3516 (77.5)	2352 (76.9)	405 (70.4)	759 (83.9)	<0.001
Diabetes	1884 (41.5)	1379 (45.1)	228 (39.7)	277 (30.6)	<0.001
Hyperlipidemia	2172 (47.9)	1739 (56.9)	283 (49.2)	150 (16.6)	<0.001
AF	551 (12.1)	475 (15.5)	32 (5.6)	44 (4.9)	<0.001
IHD	320 (7.1)	243 (7.9)	36 (6.3)	41 (4.5)	0.001

Abbreviations: AF, atrial fibrillation; AIS, acute ischemic stroke; ICH, intracerebral hemorrhage; IHD, ischemic heart disease; SD, standard deviation; TIA, transient ischemic attack.

kappa values than that using only the first 5 diagnosis codes (algorithm 4). Including diagnosis codes from prior inpatient or outpatient records to the first 5 diagnosis codes from the index hospitalization (algorithms 13 and 16) improved the kappa values. Moreover, the length of lookback period (one versus two years) did not substantially change the kappa values. For hypertension, diabetes, hyperlipidemia, and AF, combining prescription data and diagnosis codes from the index hospitalization with an "OR" (algorithm 6) also improved the kappa values. For IHD, combining prescription data and diagnosis codes with an "OR" (algorithms 3, 6, 9, 12, 15, 18) caused a marked drop in the kappa values. The kappa values for algorithms that did not contain information from the index hospitalization (algorithms 7–12) were all below 0.60. In particular, all the kappa values for hyperlipidemia did not exceed 0.20, indicating merely slight agreement between the claims and registry data.

Algorithms Based on Index Hospitalization Claims Alone

When 20 diagnosis codes from the index hospitalization were used (algorithms 1–3 in Figure 1), the agreement was substantial to perfect for diabetes and AF, moderate to substantial for hypertension, moderate for hyperlipidemia, and slight to moderate for IHD. However, when only 5 diagnosis codes were available (algorithms 4–6), the agreement was still substantial to perfect for diabetes, but moderate to substantial for AF, fair to moderate for hypertension and hyperlipidemia, and slight to fair for IHD. Combining prescription data with an "AND" (algorithms 2 and 5) generally increased specificity at the expense of decreases in sensitivity and agreement (Supplementary Tables 2–6). By contrast, combining prescription data with an "OR" (algorithms 3 and 6) increased sensitivity but not necessarily improved agreement (Supplementary Tables 2–6).

Effects of Incorporating Prescription Data

Combining prescription data and diagnosis codes had varied effects on the agreement for different risk factors. For algorithms using 20 diagnosis codes from the index hospitalization, incorporating prescription data did not improve the kappa values for all risk factors except for hyperlipidemia. However, for algorithms using the first 5 diagnosis codes from the index hospitalization, combining prescription data with an "OR" (algorithm 6) increased the kappa values for hypertension, diabetes, hyperlipidemia, and AF. Similar effects were observed in other algorithms that combined prescription data with an "OR" (algorithms 9, 12, 15, and 18), especially for diabetes and hyperlipidemia. By contrast, for IHD, combining prescription data only decreased the kappa values, particularly when an "OR" logical operand was used.

Optimal Algorithms

For studies using the NHIRD, algorithms using 20 diagnosis codes are not applicable because this database currently only includes 5 diagnosis codes for each inpatient record.[10] Consequently, algorithms 1–3, despite achieving higher agreement, were excluded from the list of optimal algorithms. Based on the kappa values, the optimal algorithms for identifying each risk factor are listed in Table 3. According to these algorithms, the prevalence of hypertension, diabetes, hyperlipidemia, AF, and IHD were estimated to be 72.2%, 42.8%, 59.8%, 10.4%, and 9.3%, respectively. The agreement between the claims and registry data was perfect for diabetes, substantial for AF, and moderate for hypertension, hyperlipidemia, as well as IHD.

Discussion

We assessed various algorithms for identifying stroke risk factors in Taiwan's NHI claims data. The algorithms using 20 diagnosis codes from the index hospitalization generally performed best in terms of the agreement between claims and registry data. Nevertheless, they are not applicable for the NHIRD since the NHIRD currently provides a maximum of 5 diagnosis codes for each inpatient record. The optimal algorithms for the NHIRD varied across risk factors. The algorithms that use 5 diagnosis codes from the index hospitalization plus diagnosis codes from prior inpatient and outpatient records usually had the highest agreement. Combining prescription data may further improve the agreement, particularly for diabetes and hyperlipidemia. The length of lookback period essentially did not alter algorithm performance.

Table 3 Optimal Algorithms for Identifying Stroke Risk Factors in Patients with Stroke Using the NHIRD

Risk Factor	Algorithm	Sen, % (95% CI)	Spe, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	Kappa (95% CI)	Estimated Prevalence, % (95% CI)
Hypertension	Alg 16: Diagnosis codes recorded in the index hospitalization, or in at least one inpatient, or in at least two outpatient claims in the past 2 years	85.7 (84.5–86.8)	74.1 (71.3–76.7)	91.9 (90.9–92.8)	60.1 (57.3–62.8)	0.552 (0.524–0.580)	72.2 (70.9–73.5)
Diabetes	Alg 15: Either diagnosis or medication codes recorded in the index hospitalization, or in at least one inpatient, or in at least two outpatient claims in the past 1 year	89.9 (88.4–91.2)	90.7 (89.5–91.8)	87.3 (85.7–88.7)	92.6 (91.6–93.6)	0.802 (0.784–0.820)	42.8 (41.3–44.2)
Hyperlipidemia	Alg 15: Either diagnosis or medication codes recorded in the index hospitalization, or in at least one inpatient, or in at least two outpatient claims in the past 1 year	86.9 (85.4–88.3)	65.1 (63.1–67.0)	69.6 (67.8–71.3)	84.4 (82.6–86.0)	0.514 (0.490–0.539)	59.8 (58.3–61.2)
AF	Alg 16: Diagnosis codes recorded in the index hospitalization, or in at least one inpatient, or in at least two outpatient claims in the past 2 years	73.5 (69.6–77.1)	98.3 (97.8–98.7)	85.6 (82.1–88.7)	96.4 (95.8–97.0)	0.765 (0.734–0.795)	10.4 (9.5–11.3)
IHD	Alg 13: Diagnosis codes recorded in the index hospitalization, or in at least one inpatient, or in at least two outpatient claims in the past 1 year	64.4 (58.9–69.6)	94.9 (94.2–95.6)	49.0 (44.2–53.9)	97.2 (96.7–97.7)	0.518 (0.473–0.564)	9.3 (8.4–10.1)

Abbreviations: Alg, algorithm; AF, atrial fibrillation; CI, confidence interval; IHD, ischemic heart disease; NHIRD, National Health Insurance Research Database; NPV, negative predictive value; PPV, positive predictive value; Sen, sensitivity, Spe, specificity.

In 2016, Taiwan's NHI started to use ICD-10-CM coding and increased the diagnoses columns from 5 to 20 for each inpatient claim. But up until now, the maximum number of diagnosis columns in the NHIRD has been restricted to 5. For patients with a primary diagnosis of stroke, only 4 diagnosis columns were available for coding comorbidities as well as complications during the hospital course. It is anticipated that algorithms based on such a limited number of diagnosis columns had lower sensitivity and agreement than those reported in the literature.[12] Consequently, before the restriction on the number of diagnosis columns in the NHIRD is lifted, researchers using the NHIRD are recommended to incorporate diagnosis codes from prior inpatient and outpatient claims, as well as prescription data when identifying stroke risk factors for hospitalized stroke patients. On the other hand, if researchers directly retrieve NHI claims data from study hospitals, the index hospitalization claims with 20 diagnosis columns may be enough to identify stroke risk factors, saving the extra time and effort required to process additional claims data from prior clinical encounters.

In line with our prior study that used ICD-9-CM diagnosis codes for identifying stroke risk factors,[15] algorithms incorporating both diagnosis codes and prescription data with an "OR" increased the agreement between claims and registry data for some stroke risk factors, particularly for hyperlipidemia. The improvement was most prominent for algorithms using 5 diagnosis codes from the index hospitalization. In practice hyperlipidemia is generally defined by the presence of elevated levels of lipid profile or a prior documentation of such condition in the medical history. Being considered a relatively minor medical condition, hyperlipidemia is commonly under-documented in the discharge summary,[18] let alone being coded within the first five diagnosis columns. For example, only 12.4% of patients hospitalized for AIS were found to have hyperlipidemia based on their discharge summaries[18] in contrast to a nearly 50% prevalence of hyperlipidemia according to the nationwide Taiwan Stroke Registry.[17] Under these circumstances, prescription data did help identify hyperlipidemia because lipid modifying agents are primarily indicated for hyperlipidemia. Multiple studies have similar findings showing that ICD codes alone were not sufficient to identify hyperlipidemia or hypercholesterolemia in patients with stroke or myocardial infarction,[12,19] whereas combining prescription data or even laboratory results did indeed increase the algorithm performance.[20]

On the other hand, prescription data was essentially useless for identifying IHD probably because the classes of medications for IHD have a broad range of indications other than IHD. For example, ATC codes of C07 (beta

blocking agents), C08 (calcium channel blockers), and C09 (agents acting on the renin-angiotensin system) are used to identify both hypertension and IHD. Since the prevalence of hypertension was 10 times higher than that of IHD, it would be tricky to accurately identify IHD in a patient with a diagnosis of hypertension who was prescribed with C07, C08, or C09 in the meantime. In this case, it might be possible to improve the algorithm performance by constructing a more sophisticated algorithm where medications used to identify IHD are adjusted according to whether a patient has hypertension. However, the more complex the algorithms, the more awkward it is to use them in research.

The variation in the identification of different stroke risk factors may be attributed to multiple factors, such as quality of charting by physicians, perceived importance of risk factors by coders, and insufficient time to "code everything".[12] In addition, the reimbursement policies of health insurance payers may play a role. For example, diabetes generally had the highest agreement among the five studied risk factors. Even algorithms using only data from prior inpatient and outpatient claims achieved moderate agreement between claims and registry data. Taiwan's NHI Administration has implemented a pay-for-performance program for diabetes care for more than two decades.[21] Consequently, healthcare providers might have stronger financial incentives to code diabetes accurately than other risk factors. By contrast, hyperlipidemia had the lowest agreement for algorithms using only data from prior inpatient and outpatient claims. In Taiwan, hyperlipidemia may go undiagnosed until the onset of the first stroke episode in 55% of patients.[22] Furthermore, Taiwan's NHI Administration has strict reimbursement criteria regarding lipid-lowering drugs for primary prevention of cardiovascular events.[23] Therefore, healthcare providers might find less need in coding hyperlipidemia.

Limitations

This study has several limitations worth noting. First, this study included only hospitalized patients with AIS, TIA, or ICH. Whether the study findings could be generalized to those not hospitalized is undetermined. Patients with TIA and minor AIS are less likely to be hospitalized,[24] particularly under the influence of COVID-19.[25] Second, several stroke risk factors were not investigated, such as obesity, tobacco use, and excessive alcohol consumption. Further studies to determine the coding validity for these risk factors may be warranted. However, they are typically under-coded in administrative claims databases.[26–28] Third, this study did not use chart review or laboratory data to verify the status of stroke risk factors. Instead, we used registry data as the reference standard. Because stroke risk factors recorded in the stroke registry were collected by trained personnel following a standard protocol containing clear definitions,[17] we believe the stroke registry can serve as a reasonable reference standard.

Conclusion

We investigated various claims-based algorithms for identifying five important stroke risk factors in patients hospitalized for AIS, TIA, or ICH using ICD-10-CM diagnosis codes and prescription data. While algorithms using diagnosis codes alone are sufficient to identify hypertension, AF, and IHD, algorithms combining both diagnosis codes and prescription data are more suitable for identifying diabetes and hyperlipidemia. The study results may provide a reference for future studies using Taiwan's NHI claims data.

Acknowledgments

This research was funded in part by the Ditmanson Medical Foundation Chia-Yi Christian Hospital-National Chung Cheng University Joint Research Program (CYCH-CCU-2022-03). The funders of the research had no role in the design and conduct of the study, interpretation of the data, or decision to submit for publication. The authors would like to thank Ms. Li-Ying Sung for English language review.

Disclosure

The authors report no conflicts of interest in this work.

References

1. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol*. 2021;20(10):795–820. doi:10.1016/s1474-4422(21)00252-0.
2. Kleindorfer DO, Towfighi A, Chaturvedi S, et al. 2021 Guideline for the Prevention of Stroke in Patients With Stroke and Transient Ischemic Attack: a Guideline From the American Heart Association/American Stroke Association. *Stroke*. 2021;52(7):e364–e467. doi:10.1161/str.0000000000000375
3. Gorelick PB. Stroke Prevention. *Arch Neurol*. 1995;52(4):347–355. doi:10.1001/archneur.1995.00540280029015
4. O'Donnell MJ, Rangarajan S, Xavier D, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. *Lancet*. 2016;388(10046):761–775. doi:10.1016/s0140-6736(16)30506-2
5. Otite FO, Liaw N, Khandelwal P, et al. Increasing prevalence of vascular risk factors in patients with stroke: a call to action. *Neurology*. 2017;89(19):1985–1994. doi:10.1212/wnl.0000000000004617
6. Hsieh CY, Wu DP, Sung SF. Trends in vascular risk factors, stroke performance measures, and outcomes in patients with first-ever ischemic stroke in Taiwan between 2000 and 2012. *J Neurol Sci*. 2017;378:80–84. doi:10.1016/j.jns.2017.05.002
7. Chang KW, Xian Y, Zhao X, et al. Antiplatelet patterns and outcomes in patients with atrial fibrillation not prescribed an anticoagulant after stroke. *Int J Cardiol*. 2020;321:88–94. doi:10.1016/j.ijcard.2020.08.011
8. Shim DH, Kim Y, Roh J, et al. Hospital Volume Threshold Associated with Higher Survival after Endovascular Recanalization Therapy for Acute Ischemic Stroke. *J Stroke*. 2020;22(1):141–149. doi:10.5853/jos.2019.00955
9. Yu AYX, Holodinsky JK, Zerna C, et al. Use and Utility of Administrative Health Data for Stroke Research and Surveillance. *Stroke*. 2016;47(7):1946–1952. doi:10.1161/strokeaha.116.012390
10. Hsieh CY, Su CC, Shao SC, et al. Taiwan's National Health Insurance Research Database: past and future. *Clin Epidemiology*. 2019;11:349–358. doi:10.2147/clep.s196293
11. Sung SF, Hsieh CY, Hu YH. Two Decades of Research Using Taiwan's National Health Insurance Claims Data: bibliometric and Text Mining Analysis on PubMed. *J Med Internet Res*. 2020;22(6):e18457. doi:10.2196/18457
12. Kokotailo RA, Hill MD. Coding of Stroke and Stroke Risk Factors Using International Classification of Diseases, Revisions 9 and 10. *Stroke*. 2005;36(8):1776–1781. doi:10.1161/01.str.0000174293.17959.a1
13. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43(11):1130–1139. doi:10.1097/01.mlr.0000182534.19832.83
14. Tonelli M, Wiebe N, Fortin M, et al. Methods for identifying 30 chronic conditions: application to administrative data. *BMC Med Inform Decis Mak*. 2015;15(1):31. doi:10.1186/s12911-015-0155-5
15. Sung SF, Hsieh CY, Lin HJ, Chen YW, Yang YHK, Li CY. Validation of algorithms to identify stroke risk factors in patients with acute ischemic stroke, transient ischemic attack, or intracerebral hemorrhage in an administrative claims database. *Int J Cardiol*. 2016;215:277–282. doi:10.1016/j.ijcard.2016.04.069
16. Hsu M, Wang C, Huang L, Lin C, Lin F, 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
17. Hsieh F-I, Lien L-M, Chen S-T, et al. Get With The Guidelines-Stroke Performance Indicators: surveillance of Stroke Care in the Taiwan Stroke Registry. *Circulation*. 2010;122(11):1116–1123. doi:10.1161/circulationaha.110.936526
18. Cheng CL, Kao YHY, 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
19. Youngson E, Welsh RC, Kaul P, McAlister F, Quan H, Bakal J. Defining and validating comorbidities and procedures in ICD-10 health data in ST-elevation myocardial infarction patients. *Medicine*. 2016;95(32):e4554. doi:10.1097/md.0000000000004554
20. Oake J, Aref-Eshghi E, Godwin M, et al. Using Electronic Medical Record to Identify Patients With Dyslipidemia in Primary Care Settings: international Classification of Disease Code Matters From One Region to a National Database. *Biomed Inform Insights*. 2017;2017(9). doi:10.4137/bii.s40801
21. Chen TT, Oldenburg B, Hsueh YS. Chronic care model in the diabetes pay-for-performance program in Taiwan: benefits, challenges and future directions. *World J Diabetes*. 2021;12(5):578–589. doi:10.4239/wjd.v12.i5.578
22. Sung SF, Lai ECC, Wu DP, Hsieh CY. Previously undiagnosed risk factors and medication nonadherence are prevalent in young adults with first-ever stroke. *Pharmacoepidemiol Drug Saf*. 2017;26(12):1458–1464. doi:10.1002/pds.4250
23. Hsu CY, Chen WJ, Chen HM, Tsai HY, Hsiao FY. Impact of changing reimbursement criteria on statin treatment patterns among patients with atherosclerotic cardiovascular disease or cardiovascular risk factors. *J Clin Pharm Ther*. 2021;46(2):415–423. doi:10.1111/jcpt.13299
24. Kapral MK, Hall R, Fang J, et al. Predictors of Hospitalization in Patients With Transient Ischemic Attack or Minor Ischemic Stroke. *Can J Neurol Sci*. 2016;43(4):523–528. doi:10.1017/cjn.2016.12
25. Butt JH, Fosbøl EL, Østergaard L, et al. Effect of COVID-19 on First-Time Acute Stroke and Transient Ischemic Attack Admission Rates and Prognosis in Denmark. *Circulation*. 2020;142(12):1227–1229. doi:10.1161/circulationaha.120.050173
26. Kim HM, Smith EG, Stano CM, et al. Validation of key behaviourally based mental health diagnoses in administrative data: suicide attempt, alcohol abuse, illicit drug abuse and tobacco use. *BMC Health Serv Res*. 2012;12(1):18. doi:10.1186/1472-6963-12-18
27. Kazzi ESA, Lau B, Li T, Schneider EB, Makary MA, Hutfless S. Differences in the Prevalence of Obesity, Smoking and Alcohol in the United States Nationwide Inpatient Sample and the Behavioral Risk Factor Surveillance System. *PLoS One*. 2015;10(11):e0140165. doi:10.1371/journal.pone.0140165
28. Suissa K, Schneeweiss S, Lin KJ, Brill G, Kim SC, Patorno E. Validation of obesity-related diagnosis codes in claims data. *Diabetes Obes Metab*. 2021;23(12):2623–2631. doi:10.1111/dom.14512