Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

Taiwan's National Health Insurance Research Database: past and future

This article was published in the following Dove Press journal:
Clinical Epidemiology

Cheng-Yang Hsieh[1,2,*]
Chien-Chou Su[1,*]
Shih-Chieh Shao[1,3]
Sheng-Feng Sung[4,5]
Swu-Jane Lin[6]
Yea-Huei Kao Yang[1]
Edward Chia-Cheng Lai[1,7]

[1]School of Pharmacy, Institute of Clinical Pharmacy and Pharmaceutical Sciences, College of Medicine, National Cheng Kung University, Tainan, Taiwan; [2]Department of Neurology, Tainan Sin Lau Hospital, Tainan, Taiwan; [3]Department of Pharmacy, Chang Gung Memorial Hospital, Keelung, Taiwan; [4]Division of Neurology, Department of Internal Medicine, Ditmanson Medical Foundation Chiayi Christian Hospital, Chiayi City, Taiwan; [5]Department of Information Management and Institute of Healthcare Information Management, National Chung Cheng University, Chiayi County, Taiwan; [6]Department of Pharmacy Systems, Outcomes & Policy, College of Pharmacy, University of Illinois at Chicago, Chicago, IL, USA; [7]Department of Pharmacy, National Cheng Kung University Hospital, Tainan, Taiwan

*These authors contributed equally to this work

Abstract: Taiwan's National Health Insurance Research Database (NHIRD) exemplifies a population-level data source for generating real-world evidence to support clinical decisions and health care policy-making. Like with all claims databases, there have been some validity concerns of studies using the NHIRD, such as the accuracy of diagnosis codes and issues around unmeasured confounders. Endeavors to validate diagnosed codes or to develop methodologic approaches to address unmeasured confounders have largely increased the reliability of NHIRD studies. Recently, Taiwan's Ministry of Health and Welfare (MOHW) established a Health and Welfare Data Center (HWDC), a data repository site that centralizes the NHIRD and about 70 other health-related databases for data management and analyses. To strengthen the protection of data privacy, investigators are required to conduct on-site analysis at an HWDC through remote connection to MOHW servers. Although the tight regulation of this on-site analysis has led to inconvenience for analysts and has increased time and costs required for research, the HWDC has created opportunities for enriched dimensions of study by linking across the NHIRD and other databases. In the near future, researchers will have greater opportunity to distill knowledge from the NHIRD linked to hospital-based electronic medical records databases containing unstructured patient-level information by using artificial intelligence techniques, including machine learning and natural language processes. We believe that NHIRD with multiple data sources could represent a powerful research engine with enriched dimensions and could serve as a guiding light for real-world evidence-based medicine in Taiwan.

Keywords: Health and Welfare Data Center of Taiwan, real-world data, big data analysis, validation, database cross-linkage

Introduction

Aging populations with increasing medical expenditures will challenge health care systems worldwide, with Taiwan being no exception.[1] However, some opportunities are present to address such a challenge. For example, Taiwan's National Health Insurance Research Database (NHIRD) exemplifies a population-level data source for health care research, generating evidence to support clinical decisions[2–5] and health care policy-making.[6] With the fast advent and uptake of data science technology, Taiwan's NHIRD can serve as a foundation for big data analysis and the procurement of real-world evidence. By adopting cost-effective evidence-based interventions, it is possible to improve patients' outcomes without increasing financial pressure on the health care systems.[7]

However, retrospective studies using the NHIRD may face a number of challenges, including unverified diagnosis coding or mortality outcomes, lack of

Correspondence: Edward Chia-Cheng Lai
School of Pharmacy, Institute of Clinical Pharmacy and Pharmaceutical Sciences, College of Medicine, National Cheng Kung University, No.1, University Road, 701 Tainan, Taiwan
Tel +886 6 2353535 ext 6209
Email edward_lai@mail.ncku.edu.tw

submit your manuscript | www.dovepress.com
DovePress
http://doi.org/10.2147/CLEP.S196293

information on disease severity, and unmeasured confounders.[8] One way to address those limitations is by linking with other nationwide databases, such as cause of death registry, disease/medication surveillance data, so as to obtain additional external data not available in the primary database.[9] However, privacy issues regarding disclosure of patients' identification while linking different databases raise additional concerns.

Recently, Taiwan's Ministry of Health and Welfare (MOHW) established a Health and Welfare Data Center (HWDC), a data repository site that centralizes the NHIRD and other health-related databases (eg, National Health Interview Survey) for data management and analyses. Several analytical approaches and policies have been developed to ensure the privacy and safety of data. Thus, the HWDC may become a solid foundation for studies utilizing NHIRD and other databases.

In this review, we will introduce the NHIRD through its history, and describe how it will be integrated with other databases in the HWDC. We address the strengths and limitations of the previous NHIRD studies and propose possible methodological solutions. To conclude, we offer our visions about the HWDC for facilitating high-quality big data analysis and biomedical research.

Current status of NHIRD use

Content/accessibility

On March 1, 1995, Taiwan launched a single-payer mandatory enrollment National Health Insurance (NHI) Program. Starting in 2002, Taiwan's National Health Research Institutes (NHRI) established and continue to maintain NHIRD for public research purposes. Up to 99.99% of Taiwan's population are enrolled under this program.[10] The NHIRD, derived from claims data of NHI beneficiaries, can thus illuminate the disease burden and health care process of the entire Taiwanese population.

The information contained in NHIRD is stored in different datasheets, including registry for beneficiaries, ambulatory care claims, inpatient claims, prescriptions dispensed at pharmacies, registry for medical facilities, and registry for board-certified specialists. These datasheets can be linked with individual personal identification numbers (PINs) to provide patient-level information on demographic characteristics for research.

Before 2016, the International Classification of Diseases, Ninth Edition, Clinical Modification (ICD-9-CM) was used for recording diagnosis in NHIRD, and the Tenth Edition (ICD-10) has been used since 2016. A specifically developed system, the NHI Drug Codes, was used to record medications dispensed. Because this drug coding system is not structured hierarchically, many researchers have mapped the codes to other international coding systems such as the WHO Anatomical Therapeutic Chemical Classification System (ATC code) while carrying out their studies.

In studies involving data derived from human subjects, privacy and data confidentiality are of the utmost concern.[11] To protect the privacy of patients, the NHRI has encrypted the names of patients, health care providers, and medical institutions with unique and anonymous identifiers. If necessary, investigators can use these identifiers for file linkage within the NHIRD data centers. The NHRI also has other restrictions on accessing the database. For example, every applicant seeking to use NHIRD must be a researcher or clinician from a university, research institute, or hospital, and the use of the data must be for research purposes only. All applications are to be reviewed by peer experts to ensure the rationality of the use. Researchers must follow the Computer-Processed Personal Data Protection Law and related regulations in Taiwan, and sign an agreement declaring that no attempt will be made to retrieve information potentially violating the privacy of patients or health care providers. Notably, a NHIRD study did not need an approval from the Institutional Review Board (IRB) prior to 2012. Because this might be considered a major problem in terms of ethical issue while using database for a clinical study, after 2012, applicants must receive an approval from the IRB before requesting access to the NHIRD for studies.

Beyond the protective strategies aforementioned, the NHRI has also restricted the amount of data requested by researchers to ≤10% of Taiwan's population. That means a researcher can only request data for up to about 2.3 million individuals. Researchers therefore should consider the disease prevalence and sample size before applying for a data-cut of NHIRD. For highly prevalent diseases (eg, diabetes mellitus), where the number of cases may be >2.3 million, investigators will need to consider a random sample selection from the entire diabetic cohort in the NHIRD. In addition to disease-specific databases, the NHRI also provides three Longitudinal Health Insurance Databases (LHID2000, LHID2005, and LHID2010) which randomly sampled 1 million beneficiaries from the original NHIRD in the years 2000, 2005, and 2010, respectively. The LHIDs contain the most updated claims data of sampled individuals since 1997. The representativeness of LHIDs has been validated by NHRI.[12]

Utilization

The representativeness and comprehensiveness of the NHIRD make it a good source for generating population-based evidences to support decisions.[13] In addition, the NHIRD provides opportunities for international collaborative study to compare differences between countries and races in health care, treatment, and outcomes. Several international multi-database studies that included NHIRD have been published.[14–17] The NHIRD has also become one of the core data resources of the Asian Pharmacoepidemiology Network (AsPEN) for conducting international comparative studies. AsPEN is an international initiative to support international study in Asia. Details of the AsPEN are described elsewhere.[18]

HWDC: opportunity for big data analysis

Although the NHIRD is well managed by governmental department, the use of the NHIRD for health care studies and the ownership of the data have recently aroused fervent discussions. Some civil groups have even filed lawsuits against the use of the NHIRD by the MOHW due to data privacy concerns. While the final verdict of the courts has not forbidden the use of the NHIRD, the debate on the legal aspects of the NHIRD is still ongoing.

To further strengthen the protection of health data, the MOHW of Taiwan has created large data repositories, the HWDC, and centralized health databases from different departments such as Taiwan's Center for Disease Control. The HWDC manages those databases, including the NHIRD, under tight supervision. Since November 2015, researchers are required to visit an HWDC to perform on-site analysis after accessing HWD through remote connection to MOHW servers. No electronic devices may be brought into a center, and individual-level data are forbidden to be taken out. All analysis results to be brought out from a center must be reviewed by data custodians to prevent any disclosure of patient identity.

Although the tight regulation of on-site analysis has led to inconvenience for analysts and has increased time and costs required for research, the HWDC has created opportunities for enriched dimensions of study by linking across the NHIRD and other databases. Currently, the HWDC contains more than 70 databases, which can be classified into two categories:

(A) **Linkable health care databases**: The databases can be linked to each other within HWDC by using PINs, such as NHIRD and National Health Interview Survey, etc.

(B) **Released health care databases**: The databases can be accessed outside HWDC, but these databases are de-identified and cannot be linked to other databases.

The detailed characteristics of the databases centralized in the HWDC are publicly accessible on its official website.[19] We summarized the information in Table 1. Because the NHIRD provides population-based samples, it can be employed as the core data source to link with other databases by using the unencrypted PINs. For example, we are able to use PINs to link the NHIRD with the National Health Interview Survey, to acquire more details such as the socioeconomic status and lifestyle behaviors of a patient. Additionally, we might be able to use the PINs of health care providers to link the NHIRD with the registry of certified specialists to acquire the characteristics of health-care providers, or use the identification numbers of hospitals to link between the NHIRD and facility databases to obtain additional data. Figure 1 is a schematic presentation of the cross-linkage between databases within the HWDC. By combining multiple databases, a researcher is able to access more variables than those available in the NHIRD, such as physical examination results, laboratory data, stage of cancer, level of disability, quality of life, body mass index, smoking, marital status, education, household income, etc. Such information can vastly improve the quality and capacity of research. For example, because the Taiwan Cancer Registry includes both clinical staging and pathologic staging developed by the American Joint Committee on Cancer, while we link the NHIRD to the Taiwan Cancer Registry, it is possible to acquire the stage information to determine cancer status and therapy for oncologic studies.

NHIRD is only available for on-site analysis after 2016 for the protection of personal information. The protection on personal data although inconvenient to the researchers, nonetheless is a valid one. The restriction meant to strike a balance between protecting privacy data and generating scientific evidence to benefit the society. Currently, only researchers who are Taiwanese nationals can have direct access to data. Foreign researchers, however, have always been welcomed to collaborate on NHIRD-based research, which is attested by several publications. Currently, we are also developing other approaches that would allow analyzing the data without compromising data privacy, such as using the distributed network approaches with common data model.[20] In the near future, more nationwide surveillance databases and hospital-based electronic health databases are

Table 1 Databases in the health and welfare databases center

A. Linkable health care databases

National Health Insurance Research Database (NHIRD)	Longitudinal Health Insurance Database (LHID)
Datasets • Ambulatory care expenditures by visits • Inpatient expenditures by admissions • Expenditures for prescriptions dispensed at contracted pharmacies • Details of ambulatory care orders • Details of inpatient orders • Details of prescriptions dispensed at contracted pharmacies • Health services utilization of medical facilities • Registry for contracted medical facilities • Registry for board-certified specialists • Accreditation profile of medical facilities • Registry for medical personnel • Registry for beneficiaries • Registry for catastrophic illness	• Longitudinal Health Insurance Database 2000 • Longitudinal Health Insurance Database 2005 • Longitudinal Health Insurance Database 2010 Note: the data of 200 million random beneficiaries from years 2000, 2005, and 2010, respectively from the original NHIRD. The LHIDs contain most updated claims data of selected individuals from 1997. Datasets in LHIDs are same as original NHIRD.

Birth, death, and maternal data	Disease and injury	Screening/cross-ministry	Welfare/society	Survey data
• Birth certificate • Cause of death data • Multiple cause of death data • Maternal and child health database	• Taiwan cancer registry • Taiwan cancer registry annual report • Artificial reproductive data • Rare disease data • 18 reporting and surveillance data • Traffic accident data	• Cancer screening data (breast cancer, cervical cancer, oral cancer, colorectal cancer) • Taiwanese aborigines (birth certificate, cause of death data, household registration)	• Disabled population profile • Low-income and middle-low-income household • Family violence data • Report data of protection of child and youths • Reported data of sexual assault	• National health interview survey • Hypertension, hyperglycemia and hyperlipidemia survey • Taiwan birth cohort study data • Knowledge, attitude, and practice of contraception • Report of the home care subsidy user condition survey • Report of the survey of requirement conditions of senior citizens welfare organizations in Taiwan • Other 12 released survey data

B. Released health care databases

Health behavior	Welfare/society	Health/lifestyle
• Global student health survey • Adult smoking behavior surveillance survey file • Global youth tobacco survey • Behavioral risk factor surveillance system	• The juvenile condition survey in Taiwan • Report of the senior citizens condition survey • Single parent family condition survey • Physically and mentally disabled citizens living and demand assessment survey • Report of the home care subsidy user condition survey • The low-income and middle-income family living condition survey • Women's living conditions survey	• Taiwan longitudinal study in aging

expected to be included in the HWD. With the continuing increase in the volume and quality of data and the quick advancement of analytic tools, stakeholders and researchers are likely to bear greater responsibility to maintain the ethical and legal requirements when accessing the data.

External criticism

Fishing expedition

Although a health care study might not always be driven by hypothesis, it is undeniable that the advances in computer techniques and the relatively lower cost of database maintenance have led to a considerable amount of "fishing expedition" studies, which might be of disservice to health care big data research. For example, Hampson et al indicated that some researchers may misuse NHIRD to "produce" papers en masse by applying templates, without due consideration of the essence of scientific research.[21] Such misconduct could damage the credibility of both the database and the research based on database.

This issue also reminds us again that, although every study should meet all the requirements of a scientific inquiry, database-driven studies should be held to an even higher standard because of the greater potential for bias in observational studies.[22,23] Health data researchers in Taiwan should work with the government to develop methods to remedy this problem and reestablish the credibility of NHIRD-based research.

Validation

Like all electronic health databases, coding errors and purposely "upcoding" could be a problem of NHIRD. For example, to avoid refusal of reimbursement by the NHI, health care providers might resort to upcoding the diagnoses to more severe ones. Misclassification bias may thus become an issue if those diagnosis codes have not been properly validated. Although large data sets could potentially overcome this problem, the real impact of the incorrect coding awaits further elucidated. Same as other health care databases, researchers need to be mindful about the problem, and apply various strategies to improve data accuracy. For example, one might use stricter definition to increase the accuracy of diagnosis, such as only including patients who have had more than two records of the targeted diagnosis or those who have also been prescribed with related medications. Other efforts to ensure data validity including algorithms developed to identify targeted diagnosis and treatment by applying multiple criteria, and

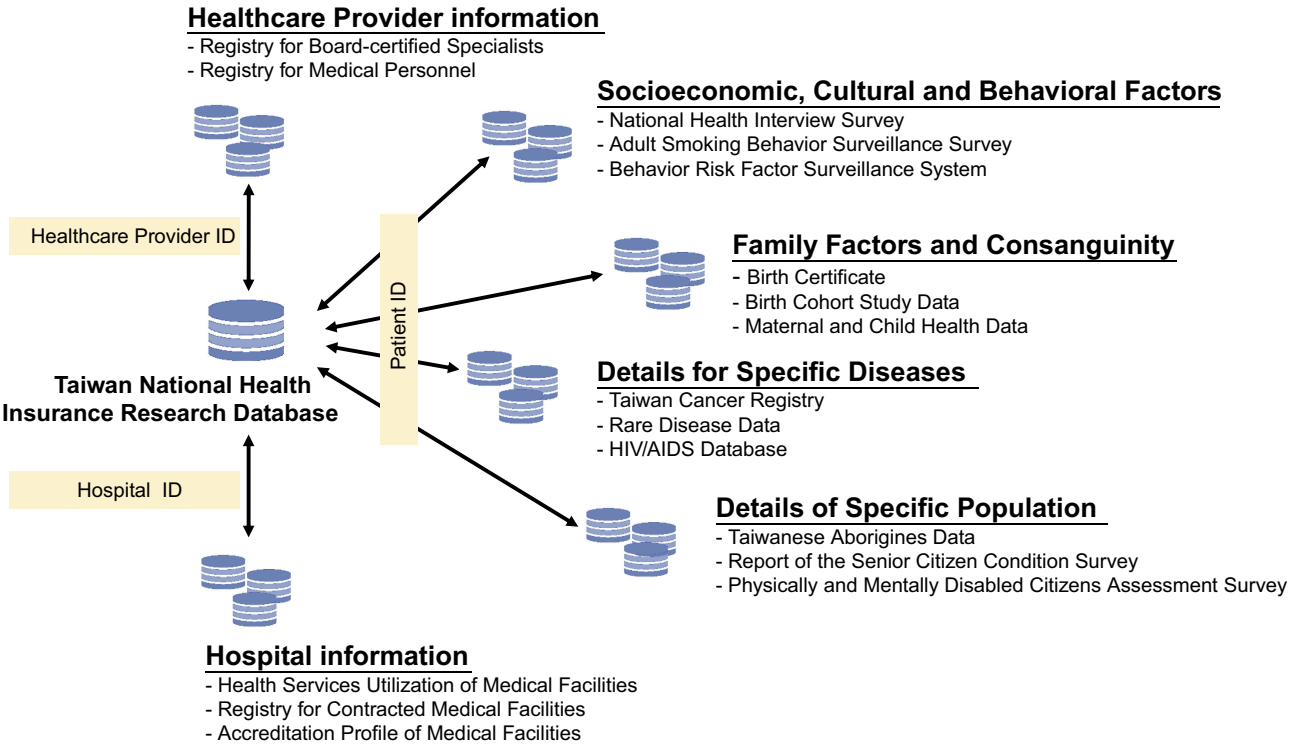

Figure 1 Conceptual presentation of database cross-linkage within HWDC.

validation studies. A continuous and concerted action by NHI and researchers is required to address the issue.

Several validation studies have been performed to evaluate the validity of diagnosis codes in the NHIRD (Table 2). Most of the validated diagnosis codes are for some common conditions or severe diseases, and with modest to high sensitivity and positive predictive values (eg, epilepsy,[24] ischemic stroke,[25] hypertension, diabetes, hyperlipidemia, fibrillation,[26] all cancer,[27] etc.). Notably, ICD-10 diagnosis codes adopted after the year 2015 in the NHIRD have not been validated yet. Furthermore, under the NHI program, patients with severe illnesses can apply for catastrophic illness certification, so as to be exempted from certain NHI payments and copayments for each health care encounter. All applications for catastrophic illness certification are reviewed by experts, and therefore the diagnosis can be considered highly accurate; hence, the catastrophic illness file has been used for case ascertainments in respective research.

Ways to deal with unmeasured variables

Important but unmeasured variables are another important issue for NHIRD studies. For example, indicators for disease severity are lacking in the NHIRD. This may lead to biased estimates of relationships, since disease severity at the baseline is among the most important confounding factors in healthcare research. The term "confounding by indication" describes the situation that the treatment indicated for a patient is not a random process, but rather is a consideration based on the baseline severity of the illness;[28] since baseline severity invariably also affects the outcome, it is difficult to isolate the real treatment effect without adjusting for the baseline severity. An example of confounding by indication is the comparison of gastrointestinal bleeding risk between cyclooxygenase-II selective nonsteroidal anti-inflammatory drugs (NSAIDs) and nonselective NSAIDs.[29] Because the selective NSAIDs may have fewer gastrointestinal side effects, clinicians are more likely to prescribe selective NSAIDs for patients who already have a higher baseline risk of gastrointestinal bleeding, resulting in a biased and unfavorable risk profile of the group receiving selective NSAIDs.

Although some unmeasured or unavailable variables of the NHIRD could probably be obtained through the linkage with external databases in the HWDC, at times, a linkage is not possible and/or the data are simply not stored in digital format. Methodological and statistical approaches have been developed to address this issue, such as the instrumental variable method.[30] Another method is to use self-controlled crossover design to control

Table 2 Summary of validation studies regarding the validity of diagnosis codes in the NHIRD

Diseases/ conditions	ICD-9-CM code	Sensitivity, %	Positive predictive value, PPV %	Note	References
Acute ischemic stroke	433.xx, 434.xx	94.5	97.9	Chart review by neurologic specialist as reference standard	Pharmacoepidemiol Drug Saf 2011[47]
Epilepsy	345.xx	81.4	76.8	Chart review by neurologic specialist as reference standard; specificity: 99.8%	Epilepsia 2012[24]
Pneumonia	480.xx–486.xx	92.3–94.7	Not available	Chart review by medical doctors as reference standard	CMAJ 2014[48]
Coronary artery bypass graft post-operative surgical site infection	996.03, 996.61, 996.72, 998.5 038.0–038.4, 038.8, 038.9, 682.6, 682.9, 780.6, 790.7, 875.0, 875.1, 891.0, 891.1, 996.03, 996.61, 996.72, 998.3, and 998.5.	35.3	19.4	Health care-associated infection surveillance data and manually reviewed medical charts as reference standard; specificity: 97.0%	BMC Med Inform Decis Mak 2014[49]
Acute myocardial infarction	410.xx	88.0	92.0	Chart review by neurologic specialist as reference standard	J Epidemiol 2014[50]
Renal dysfunction	250.4, 283.11, 403.x, 404.x, 580–589, 753.0, 753.1	38.4	76.0	Taiwan Stroke Registry as reference standard; only validated in stroke patients; specificity: 94.7%	Int J Stroke 2015[51]
Acute ischemic stroke	433.xx, 434.xx	97.3	88.4	Taiwan Stroke Registry as reference standard	J Formos Med Assoc 2015[25]
Tuberculosis contact	V01.1 with at least 1 chest radiographic examination or 795.5	98.3	Not available	Chart review by pulmonologists as reference standard	Medicine 2016[52]
Hypertension	401.x, 402.x, 403.x, 404.x, 405.x	92.4	88.5	Few conditions in patients with stroke. Taiwan Stroke Registry as reference standard	Int J Cardiol 2016[26]
Diabetes	250.x	90.9	92.0		
Hyperlipidemia	272.x	69.1	89.5		
Coronary artery disease	410.x, 411.x, 412.x, 413.x, 414.x	63.7	47.6		
Atrial fibrillation	427.31	72.8	71.1		
Tuberculosis	010–018 plus prescriptions of at least two anti-tuberculosis drugs	96.3	Not available	Chart review by pulmonologists as reference standard	Chest 2017[53]
Heart failure	428	Not available	97.6	Chart review by cardiologic specialist as reference standard	J Am Heart Assoc 2017[54]
Ischemic stroke	433–437	Not available	94.2		
All cancer	140–208	91.5	93.6	National Cancer Registry of Taiwan as reference standard	Pharmacoepidemiol Drug Saf 2018[27]
Varicose veins	454	Not available	98.0	Chart review as reference standard	JAMA 2018[55]

for time-constant confounders.[31] Active comparator design, instead of placebo-controlled, may help to eliminate the confounding effects if the unmeasured confounders are balanced between groups. Propensity score methods, including propensity score calibrations, two-stage calibration, and high-dimensional propensity score have also been used to reduce the potential confounding effects.[32–34] However, none of the methods is applicable or

appropriate for all situations. For example, it might be difficult to develop a high-quality instrumental variable or ideal active comparator for analysis, and propensity score methods are subject to the data availability and representativeness of databases.

Development of claims-basedseverity index: an example of stroke severity index

(SSI) Another way to overcome such a problem is to directly develop a claims-based index to measure disease severity. Taking stroke as an example, we can observe that stroke patients with claims of tube feeding may have either consciousness disturbance or swallowing difficulty, thus are likely to have suffered more severe stroke than those without. Using data mining techniques, we developed a novel 7-item NHIRD-based SSI to serve as a proxy for the National Institutes of Health Stroke Scale (NIHSS).[35] The SSI can be calculated easily using multiple linear regression equations as follows:

SSI=
+3.5083 (airway suctioning)
+1.3642 (bacterial sensitivity test)
+4.1770 (intensive care unit stay)
+4.5809 (nasogastric intubation)
+2.1448 (osmotherapy such as mannitol or glycerol)
+1.6569 (urinary catheterization)
−5.5761 (general ward stay)
+9.6804 (constant)

For example, the SSI for a patient who stayed in a general ward, and received bacterial sensitivity test and urinary catheterization during the stay would be 7.1254 (ie, 1.3642+1.6569–5.5761+9.6804). Subsequently, the SSI can be transformed into an "estimated NIHSS" (eNIHSS) by calibrating against the updated stroke registry data and the modified Rankin Scale for stroke. The eNIHSS has been validated in both acute ischemic stroke[36] and intracerebral hemorrhage,[37] and has been used to provide better case-mix adjustments for stroke outcome studies.[38–40] Transformation of SSI to eNIHSS can be done with the following equations:

$$\text{Ischemic stroke}: \text{eNIHSS} = 1.1722 \times \text{SSI} - 0.7533$$

$$\text{Intracerebral hemorrhage}: \text{eNIHSS} = 1.3894 \times \text{SSI} - 3.6788$$

To facilitate the use of eNIHSS for stroke study with NHIRD, we also created an easy to use online tool for researchers to easily evaluate the stroke severity when standard severity data are not available.[41] Users of such proxy measures do not need to cross-link multiple databases that may or may not offer a severity score; therefore, the method could reduce concerns of data privacy.

Visions and summary

We provide our visions for the improvement of big data analysis in health care research. First, there is a lag time of about 2 years in the availability of NHIRD data, which may diminish the possibility of timely assessments and efficiency of analysis derived from the NHIRD to support policy-making and clinical decisions. Regulators and researchers should figure out better ways to generate more real-time data. The HWDC in Taiwan should also proactively advocate efforts to obtain critical information not available in any of the current databases. We should also simplify all administrative procedures and employ efficient analytic methods to provide a better environment for research. Additionally, we should focus on education and training programs in the use of databases, incorporating the required epidemiological and statistical knowledge in order to cultivate more young researchers.

Despite the large volume of data in the HWDC, nearly all are structured data. However, data generated from routine clinical practice may exist in unstructured forms, such as clinical notes in electronic medical records (EMRs), waveforms from physiological monitors, and radiological images. The plethora of patient information in unstructured clinical notes may be a valuable data source for clinical research, even though this kind of data is less amenable to manipulation and utilization. Artificial intelligence techniques, including machine learning and natural language processing (NLP), may be used to process unstructured free text and draw conclusions to support clinical practice.[42] For example, algorithms using NLP techniques have been found to be advantageous over methods using disease codes in identifying cases with specific diagnoses from discharge summaries.[43] NLP techniques have been applied to extract meaningful information from EMRs to facilitate timely decision-making in intravenous thrombolytic therapy for acute ischemic stroke.[44]

In the era of artificial intelligence, our next step is to combine structured databases including the NHIRD with unstructured data sources. In Taiwan, with the development of the Taiwan Electronic Medical Record Template,[45] EMR systems have been widely used in almost all hospitals for routine clinical work. A large number of EMRs are generated

and accumulate during everyday patient care. Furthermore, a national EMR exchange system has been set up to provide sharing and exchange of EMRs between health care facilities.[46] Databases from this EMR exchange system can be incorporated into the data repository of the HWDC, offering researchers a very comprehensive database from which to distill more knowledge. Based on our successful experience with NHIRD, we hope that HWDC can someday transform from a data inventory warehouse into a powerful research engine, which can combine the data from different sources to serve as a guiding light for real-world evidence-based medicine in Taiwan.

Acknowledgments

This work was supported by Ministry of Science and Technology of Taiwan (107-2320-B-006-070-MY3). The authors thank Health Data Science Center, National Cheng Kung University Hospital for providing administrative support. The funding institutions of this study had no further role in the collection, analysis, and interpretation of data, the writing of this paper, or the decision to submit it for publication.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Lin YY, Huang CS. Aging in Taiwan: building a society for active aging and aging in place. *Gerontologist*. 2016;56(2):176–183. doi:10.1093/geront/gnv107
2. Wu CY, Kuo KN, Wu MS, et al. Early Helicobacter pylori eradication decreases risk of gastric cancer in patients with peptic ulcer disease. *Gastroenterology*. 2009;137(5):1641–1648.e1641-1642. doi:10.1053/j.gastro.2009.07.060
3. Chien HC, Kao Yang YH, Bai JP. Trastuzumab-related cardiotoxic effects in Taiwanese women: a nationwide cohort study. *JAMA Oncol*. 2016;2(10):1317–1325. doi:10.1001/jamaoncol.2016.1269
4. Chang SH, Chou IJ, Yeh YH, et al. Association between use of non-vitamin K oral anticoagulants with and without concurrent medications and risk of major bleeding in nonvalvular atrial fibrillation. *JAMA*. 2017;318(13):1250–1259. doi:10.1001/jama.2017.13883
5. Wu CY, Chen YJ, Ho HJ, et al. Association between nucleoside analogues and risk of hepatitis B virus-related hepatocellular carcinoma recurrence following liver resection. *JAMA*. 2012;308(18):1906–1914.
6. Hsieh CY, Tsao WC, Lin RT, et al. Three years of the nationwide post-acute stroke care program in Taiwan. *J Chin Med Assoc*. 2018;81(1):87–88. doi:10.1016/j.jcma.2017.09.003
7. Schneeweiss S. Learning from big health care data. *N Engl J Med*. 2014;370(23):2161–2163. doi:10.1056/NEJMp1401111
8. Hsiao FY, Yang CL, Huang YT, et al. Using Taiwan's National Health Insurance Research Databases for pharmacoepidemiology research. *J Food Drug Anal*. 2007;15:99–108.
9. Lee JT, Huang N, Majeed A. The need for better linkage between administrative data and clinical datasets. *BMJ*. 2015;351:h5816. doi:10.1136/bmj.h6432
10. National Health Insurance Administration. NHI profile [cited July 9, 2018]. Available from: https://www.nhi.gov.tw/English/Content_List.aspx?n=8FC0974BBFEFA56D&topn=ED4A30E51A609E49. Accessed March 29, 2019.
11. Salerno J, Knoppers BM, Lee LM, Hlaing WM, Goodman KW. Ethics, big data and computing in epidemiology and public health. *Ann Epidemiol*. 2017;27(5):297–301. doi:10.1016/j.annepidem.2017.05.002
12. National Health Insurance Research Database. Data subsets [cited March 1, 2018]. Available from: https://nhird.nhri.org.tw/en/Data_Subsets.html. Accessed March 29, 2019.
13. Hsing AW, Ioannidis JP. Nationwide population science: lessons from the Taiwan National Health Insurance Research Database. *JAMA Intern Med*. 2015;175(9):1527–1529. doi:10.1001/jamainternmed.2015.3540
14. Pratt N, Andersen M, Bergman U, et al. Multi-country rapid adverse drug event assessment: the Asian Pharmacoepidemiology Network (AsPEN) antipsychotic and acute hyperglycaemia study. *Pharmacoepidemiol Drug Saf*. 2013;22(9):915–924. doi:10.1002/pds.3440
15. Roughead EE, Chan EW, Choi NK, et al. Proton pump inhibitors and risk of Clostridium difficile infection: a multi-country study using sequence symmetry analysis. *Expert Opin Drug Saf*. 2016;15(12):1589–1595. doi:10.1080/14740338.2016.1238071
16. Roughead EE, Chan EW, Choi NK, et al. Variation in association between thiazolidinediones and heart failure across ethnic groups: retrospective analysis of large healthcare claims databases in six countries. *Drug Saf*. 2015;38(9):823–831. doi:10.1007/s40264-015-0318-4
17. Pratt N, Chan EW, Choi NK, et al. Prescription sequence symmetry analysis: assessing risk, temporality, and consistency for adverse drug reactions across datasets in five countries. *Pharmacoepidemiol Drug Saf*. 2015;24(8):858–864. doi:10.1002/pds.3780
18. Andersen M, Bergman U, Choi NK, et al. The Asian Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. *Pharmacoepidemiol Drug Saf*. 2013;22(7):700–704. doi:10.1002/pds.3439
19. Ministry of Health and Welfare. Health and Welfare Data Science Center [cited February 5, 2018]. Availabe from: https://dep.mohw.gov.tw/DOS/np-2497-113.html. Accessed March 29, 2019.
20. Lai EC, Ryan P, Zhang Y, et al. Applying a common data model to Asian databases for multinational pharmacoepidemiologic studies: opportunities and challenges. *Clin Epidemiol*. 2018;10:875–885. doi:10.2147/CLEP.S149961
21. Hampson NB, Weaver LK. Carbon monoxide poisoning and risk for ischemic stroke. *Eur J Intern Med*. 2016;31:e7. doi:10.1016/j.ejim.2016.01.006
22. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. *J Clin Epidemiol*. 2012;65(2):126–131. doi:10.1016/j.jclinepi.2011.08.002
23. Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf*. 2012;21(1):1–10. doi:10.1002/pds.2229
24. Chen CC, Chen LS, Yen MF, et al. Geographic variation in the age- and gender-specific prevalence and incidence of epilepsy: analysis of Taiwanese National Health Insurance-based data. *Epilepsia*. 2012;53(2):283–290. doi:10.1111/j.1528-1167.2011.03332.x
25. Hsieh CY, Chen CH, Li CY, et al. 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. Sung SF, Hsieh CY, Lin HJ, et al. 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

27. 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
28. Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. *Am J Epidemiol*. 1999;149(11):981–983.
29. Wolfe F, Flowers N, Burke TA, et al. Increase in lifetime adverse drug reactions, service utilization, and disease severity among patients who will start COX-2 specific inhibitors: quantitative assessment of channeling bias and confounding by indication in 6689 patients with rheumatoid arthritis and osteoarthritis. *J Rheumatol*. 2002;29(5):1015–1022.
30. Davies NM, Smith GD, Windmeijer F, et al. COX-2 selective non-steroidal anti-inflammatory drugs and risk of gastrointestinal tract complications and myocardial infarction: an instrumental variable analysis. *Epidemiology*. 2013;24(3):352–362. doi:10.1097/EDE.0b013e318289e024
31. Petersen I, Douglas I, Whitaker H. Self controlled case series methods: an alternative to standard epidemiological study designs. *BMJ*. 2016;354:i4515. doi:10.1136/bmj.i4515
32. Brookhart MA, Wyss R, Layton JB, et al. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes*. 2013;6(5):604–611. doi:10.1161/CIRCOUTCOMES.113.000359
33. Sturmer T, Schneeweiss S, Rothman KJ, et al. Performance of propensity score calibration – a simulation study. *Am J Epidemiol*. 2007;165(10):1110–1118. doi:10.1093/aje/kwm074
34. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512–522. doi:10.1097/EDE.0b013e3181a663cc
35. Sung SF, Hsieh CY, Kao Yang YH, et al. Developing a Stroke Severity Index based on administrative data was feasible using data mining techniques. *J Clin Epidemiol*. 2015;68(11):1292–1300. doi:10.1016/j.jclinepi.2015.01.009
36. Sung SF, Hsieh CY, Lin HJ, et al. Validity of a Stroke Severity Index for administrative claims data research: a retrospective cohort study. *BMC Health Serv Res*. 2016;16(1):509. doi:10.1186/s12913-016-1769-8
37. Hung LC, Sung SF, Hsieh CY, et al. Validation of a novel claims-based Stroke Severity Index in patients with intracerebral hemorrhage. *J Epidemiol*. 2017;27(1):24–29. doi:10.1016/j.je.2016.08.003
38. Hsieh CY, Lin HJ, Hu YH, et al. Stroke severity may predict causes of readmission within one year in patients with first ischemic stroke event. *J Neurol Sci*. 2017;372:21–27. doi:10.1016/j.jns.2016.11.026
39. 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
40. Hsieh CY, Lin HJ, Chen CH, et al. "Weekend effect" on stroke mortality revisited: application of a claims-based Stroke Severity Index in a population-based cohort study. *Medicine (Baltimore)*. 2016;95(25):e4046. doi:10.1097/MD.0000000000004864
41. Stroke Severity Index Calculator [cited February 18, 2019]. Availabe from: http://140.123.175.14:508/SSI/hdmlab/ssi2.jsp. Accessed March 29, 2019.
42. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2:e000101. doi:10.1136/svn-2017-000101
43. Li L, Chase HS, Patel CO, et al. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. *AMIA Annu Symp Proc*. 2008;2008:404–408.
44. Sung SF, Chen K, Wu DP, et al. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: a feasibility study. *Int J Med Inform*. 2018;112:149–157. doi:10.1016/j.ijmedinf.2018.02.005
45. Rau HH, Hsu CY, Lee YL, et al. Developing electronic health records in Taiwan. *IT Prof*. 2010;12:17–25. doi:10.1109/MITP.2010.53
46. Li YC, Yen JC, Chiu WT, et al. Building a national electronic medical record exchange system - experiences in Taiwan. *Comput Methods Programs Biomed*. 2015;121(1):14–20. doi:10.1016/j.cmpb.2015.04.013
47. Cheng CL, Kao YH, Lin SJ, Lee C-H, 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
48. Su VY, Liu CJ, Wang HK, et al. Sleep apnea and risk of pneumonia: a nationwide population-based study. *CMAJ*. 2014;186(6):415–421. doi:10.1503/cmaj.131547
49. Yu TH, Hou YC, Lin KC, Chung K-P. Is it possible to identify cases of coronary artery bypass graft postoperative surgical site infection accurately from claims data? *BMC Med Inform Decis Mak*. 2014;14:42. doi:10.1186/1472-6947-14-42
50. Cheng CL, Lee CH, Chen PS, et al. Validation of acute myocardial infarction cases in the National Health Insurance Research Database in Taiwan. *J Epidemiol*. 2014;24(6):500–507.
51. Hsieh CY, Cheng CL, Lai EC, et al. Identifying renal dysfunction in stroke patients using diagnostic codes in the Taiwan National Health Insurance Research Database. *Int J Stroke*. 2015;10(1):E5. doi:10.1111/ijs.12380
52. Su VY, Yen YF, Pan SW, et al. Latent tuberculosis infection and the risk of subsequent cancer. *Medicine (Baltimore)*. 2016;95(4):e2352. doi:10.1097/MD.0000000000004864
53. Su VY, Su WJ, Yen YF, et al. Statin use is associated with a lower risk of TB. *Chest*. 2017;152(3):598–606. doi:10.1016/j.chest.2017.04.170
54. Lin YS, Chen TH, Chi CC, et al. Different implications of heartfailure, ischemic stroke, and mortality between nonvalvular atrial fibrillation and atrial flutter-a view from a national cohort study. *J Am Heart Assoc*. 2017;6(7):e006406. doi:10.1161/JAHA.117.006406
55. Chang SL, Huang YL, Lee MC, et al. Association of varicose veins with incident venous thromboembolism and peripheral artery disease. *JAMA*. 2018;319(8):807–817. doi:10.1001/jama.2018.0246