BMJ Open

Analyses and identification of ICD codes for dementias in the research based on the NHIRD: a scoping review protocol

Ying-Jyun Shih ,[1,2] Jiun-Yi Wang,[1] Ya-Hui Wang,[2] Rong-Rong Shih,[2] Yung-Jen Yang [3]

To cite: Shih YJ, Wang JY, Wang YH, *et al*. Analyses and identification of ICD codes for dementias in the research based on the NHIRD: a scoping review protocol. *BMJ Open* 2022;**12**:e062654. doi:10.1136/bmjopen-2022-062654

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-062654).

Received 09 March 2022
Accepted 20 July 2022

[1]Department of Healthcare Administration, Asia University College of Medical and Health Science, Taichung City, Taiwan
[2]Department of Nursing, Tsaotun Psychiatric Center, Ministry of Health and Welfare, Tsaotun Township, Nan-Tou County, Taiwan
[3]Social Science Research Unit (SSRU), Institue of Education, University College London, London, UK

Correspondence to
Dr Yung-Jen Yang;
yung-jen.yang.12@ucl.ac.uk

ABSTRACT

Introduction Studies based on health claims data (HCD) have been increasingly adopted in medical research for their strengths in large sample size and abundant information, and the Taiwan National Health Insurance Research Database (NHIRD) has been widely used in medical research across disciplines, including dementia. How the diagnostic codes are applied to define the diseases/conditions of interest is pivotal in HCD-related research, but the consensus on the issue that diagnostic codes most appropriately define dementias in the NHIRD is lacking. The objectives of this scoping review are (1) to investigate the relevant characteristics in the published reports targeting dementias based on the NHIRD, and (2) to address the diversity by a case study.

Methods and analysis This scoping review protocol follows the methodological framework of the Joanna Briggs Institute Reviewer's Manual and the guidance of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping reviews. The review will be performed between 1 March and 31 December 2022 in five stages, including identifying the relevant studies, developing search strategies, individually screening and selecting evidence, collecting and extracting data, and summarising and reporting the results. The electronic databases of MEDLINE, EMBASE, CENTRAL, CINAHL, and PsycINFO, Airiti Library Academic Database, the National Health Insurance Administration's repository, and Taiwan Government Research Bulletin will be searched. We will perform narrative syntheses of the results to address research questions and will analyse the prevalence across the included individual studies as a case study.

Ethics and dissemination Our scoping review is a review of the published reports and ethical approval is not required. The results will provide a panorama of the dementia studies based on the NHIRD. We will disseminate our findings through peer-reviewed journals and conferences, and share with stakeholders by distributing the summaries in social media and emails.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ To the best of our knowledge, our study will be the first review exploring the characteristics and the utilisation of diagnostic codes in the dementia studies based on the National Health Insurance Research Database (NHIRD).
⇒ The scoping review methodology allows a broad perspective to depict the heterogeneity across the individual studies while maintaining transparency and accountability through systematic search and data extraction process.
⇒ The results of this scoping review are expected to lay the foundation for future dementia studies based on the NHIRD.
⇒ As the review is limited to literature published in English and Chinese, there are still potential publications in other languages that are not considered.
⇒ Due to the lack of consensus on the risk of bias tool for such type of study, a formal quality and risk of bias assessment of the included studies will not be performed, limiting the understanding of how the potential biases in the studies may influence the results.

INTRODUCTION

Health claims data (HCD) have been the major sources for the studies of epidemiological analyses, health service economics and outcomes, and disease-specific medical research.[1] Research based on large population-based databases enables researchers to explore medical conditions with low prevalence or interventions with small effect sizes.[2] HCD have been applied to many medical specialties, and many prestige organisations have advocated that researchers can take good advantages of big data, specifically the HCD, in dementia research.[3 4] In practice, HCD have been used in the research of dementias and their comorbidities, the disease progression trajectory, and the interaction between the biological and environmental factors.[3]

However, due to the natures of administrative data or claims data, which are usually presumed for reimbursement rather than for research,[5 6] scholars have proposed important issues that may pose threats and limitations to such studies, including the quality and reliability of the data,[7 8] the lack of consensus and standardisation across the databases,

and the doubt of data accuracy as a result of erroneous linkage.[9 10] Furthermore, despite being powerful, studies based on HCD are still vulnerable to the under-recording of dementia, resulting in under-representativeness of the target group and threats to generalisability of the study results.[11]

Being the main components of HCD, diagnostic codes play important roles in HCD-based studies, and by following appropriate algorithms, researchers are able to aim at specific health outcomes of interest or identify populations with specific diagnoses within the HCD.[12 13] However, the quality of diagnostic codes fundamentally relies on the avoidance of misdiagnosis, miscoding and misclassification, which will otherwise limit or even flaw the results as mentioned by Stein *et al* in their systematic review.[14] The issues related to diagnostic codes in dementia database research become more complicated. For example, the diagnostic gap due to underidentification caused by miscoding of dementia,[15] misidentification of dementia,[16] misclassification of dementia in HCD[17] and the high heterogeneity in selecting the diagnostic codes of dementia[18] is not uncommon, and all the above issues probably bring about complexities in HCD-based dementia studies. As the differences in selecting diagnostic codes to define dementias in relevant research would result in misidentification of dementia, it will be helpful to develop a set of standardised diagnostic codes for dementia to minimise the potential problematic impact and improve the value of HCD in dementia research.

Scoping review is a relatively young methodology in the family of evidence synthesis. It is regarded as an appropriate way to explore, configure and aggregate the evidence, and can be used as a precursor to a systematic review. Scoping reviews are able to illustrate the ways how research has been executed in the specific fields or topics, and to identify the key concepts, rationales, types of evidence and the research gap. Rather than testing theories or hypotheses, they can serve to explore the contents, range, natures and heterogeneities across the individual studies, to summarise the results and to guide the researchers for the future research about the directions and methodologies.[19–21] The methodologies of scoping review have been first proposed by Arksey and O'Malley in 2005,[22] and later strengthened by Levac *et al* by proposing a practical five-step approach.[23] Despite their endeavours, inconsistencies in execution remain. In this way, the Joanna Briggs Institute (JBI) developed a comprehensive guideline to standardise the processes first in 2014 and updated revisions in both 2017 and 2020.[24] Tricco *et al* also proposed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping review (PRISMA-ScR) to aid the researchers in reporting their studies.[25] Considering the complex nature in selecting diagnostic codes to define dementias and the high heterogeneity in the database research, scoping review would be the best suitable methodology in response to our research questions.

Since 1995, Taiwan has launched the Taiwan National Health Insurance (TNHI) with the coverage of 99.9% of the whole 23 million population and established an HCD database which cumulates the health-related records of the users in the national health insurance system.[26 27] In practice, although increasing hospitals have begun to employ the clinical coders to help the task of coding in recent years, physicians across the levels in the health system are still the main persons responsible for coding and inputting the diagnostic codes as well as the interventional codes into the administrative systems for reimbursement in TNHI. From 1995 to 2016, the diagnostic codes were based on the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), which was replaced by ICD-10-CM after 2016.[27]

Thanks to its abundant information, the National Health Insurance Administration has built up the National Health Insurance Research Database (NHIRD) and released the purchasable datasets which included de-identifiable and encrypted sampling of the health records for the researchers in the academic organisations since 2000. At present time, there are three forms of the NHIRD datasets released at different chronological time, and they are general dataset from 2000, disease-specific dataset and the latest full population dataset, which has been released since 2016. These datasets consist both inpatient and outpatient claims data, and the sets include demographic profiles as well as clinical data of the codes for diagnoses, prescriptions and interventions. From 2016, the NHIRD data were authorised to link with other government databases at the Data Science Centre, including some national census data, disease registries, health surveys, social service data, death cause data and welfare registries. The linkage with other large national databases expands the applicability especially in research and health policymaking.[26 27]

In recent years, the NHIRD has provided the researchers with abundant resource for secondary database medical research, and hundreds of studies have been published, including the research for dementia. Many of them have been used as references for healthcare practice guidance and public policymaking. Despite its strengths, however, the NHIRD still bears the same inherent weakness of HCD and there have been inconsistencies in selecting the diagnostic codes in defining dementias in the research using the NHIRD. As a result, it is imperative to investigate the characteristics of dementia researches based on the NHIRD and how the diagnostic codes are selected as well as used in such studies. This will aid in identifying the potential research gap and reduce the research waste.

In the present study, by taking the advantages of scoping review methodology, the research team intends to identify the characteristics, address the heterogeneities and explore the diversities of diagnostic codes used to define dementias in the published studies. With the results of our study, we may lay the foundation for developing a set of standardised codes for defining dementias

for future dementia studies based on the NHIRD. In this manuscript, we present the protocol for informing the implementation of a scoping review.

Aim of the study and research question (study objective)

Based on empirical experience, there are heterogeneities in NHIRD studies, including the selections of diagnostic codes for defining diseases/conditions, the size of datasets, the time length of database used and the types of subdatasets (ie, inpatient or outpatient dataset). The main aims of our research are to investigate the relevant characteristics in the published reports targeting dementias based on the NHIRD, and to address the diversity by analysing the reported prevalence as a case study. We here define the following research questions for the scoping review.

1. Which diagnostic codes in ICD-9-CM and ICD-10-CM were used to define dementias in the studies based on the NHIRD?
2. To what extent the diagnostic codes varied across the studies when being used to define dementias, in relation with the research teams and the size and types of database?
3. What differences in terms of the additional approaches other than diagnostic codes used in the inclusion or exclusion criteria to identify the individuals with dementias in the databases and the time length of database were adopted across the studies?
4. How were other important publication characteristics of database studies reported across the studies?
5. As a case study, how does the prevalence differ across the studies based on the NHIRD in relation to the major variables above?

METHODS

The present scoping review protocol has been developed based on the methodological framework of the JBI Reviewer's Manual and has been constructed following the guidance of the PRISMA-ScR.[20 21 25] A PRISMA-ScR checklist can be found in the online supplemental file 1. This protocol has been registered with the Open Science Framework (OSF: osf.io/fc65g) on 26 February 2022, and the study will be implemented between 1 March and 31 December 2022 by following the five steps (subsections 2.1–2.5):

Identification of relevant studies

Inclusion criteria

Participants

The scoping review will aim at the published research reports on dementias using the NHIRD in the literature and will include all types of dementia, which are defined and identified in any way in the reports. There will be no restriction on the types of study designs, the age of the participants in the reports and the comorbidities so as to maximise the coverage of the types of dementia research. However, we will only include the reports written in English or Chinese for the ease of data collection and analyses.

Concept

The present scoping review intends to focus on dementia research using the NHIRD and explores as well as configures the elements of the studies. The elements include the ways of selecting diagnostic codes to define dementias, the usage of subdatasets of the NHIRD, the methodological spectra of dementia research based on the NHIRD, the heterogeneity and the potential impact on the outcomes.

Table 1 Search terms and concepts

Concept 1: dementia	"Dementia" "Major neurocognitive disorder" "Alzheimer's dementia" "Parkinson's disease dementia" "Dementia with Lewy body" "Frontotemporal dementia" "Vascular Dementia" "AIDS Dementia Complex" "Mixed dementia"
Concept 2: database research & national health insurance	"Claim data" "Insurance claim data" "Insurance data/database" "Administrative (or health) data/database" "Health data" "Electronic health record" "Government database" "National Health Insurance Research Database (NHIRD)" "National Health Programme/ Service" "National data (or nationwide or population-based) research" "Longitudinal (or follow-up) Study"
Concept 3: Taiwan	"Taiwan"

Context

The scoping review will specifically focus on the dementia studies based on the NHIRD that may be influenced by the contextual factors of the pragmatic medical practices and services in Taiwan as well as some of the specific regulations in TNHI. Rather than being a constraint factor, however, this is the unique contextual characteristic that this scoping review intends to address and can inform.

Exclusion criteria

We will exclude the reports when the full texts are not readily available, and in addition, review articles, study protocols, grey literature and texts that are not peer-reviewed or fail to provide detailed information that is in line with our study are excluded. The grey literature here included letters, editorials or leading articles, commentaries, conference abstracts or presentations, and dissertations or theses.

Search strategy

The research team will develop search strategies that aim to include the study reports published between 2000 and 2022, ranging from the first year when the NHIRD was available to the academic researchers and the time to execute this review. Published studies in English or Chinese with all study designs will be included and the research team will search the major electronic databases of MEDLINE-OVID, EMBASE-OVID, Cochrane Central Register of Controlled Trials, CINAHL and PsycINFO to identify the reports on the topic. For the Chinese literature and grey literature, we will search the Airiti Library Academic Database, the National Health Insurance Administration's repository which collects the articles using Taiwan NHIRD and Taiwan Government Research Bulletin. The search will consist of three major steps; and the first step is to identify the search terms informing the three main concepts of the review, including dementia, database research and National Health Insurance Research Database, and Taiwan. The identified search terms of each concept will be first organised with the Boolean operator OR into each category, and then the three categories will be combined with the Boolean operator AND (table 1). The search strategies will be developed by a trained researcher (YJY) and then reviewed by a librarian in the university library. The second step is to use all the identified keywords and index terms to undertake search in the titles and abstracts of relevant articles in each electronic database. The final step is to export searched results into the electronic bibliographical software EndNote version X8 (Clarivate Analytics, Pennsylvania, USA). The second and third steps will be individually carried out by two different researchers (YJS and JYW). The pilot search strategies are shown in the online supplemental file 2. In addition to the electronic database search, we will conduct hand search for the potential studies in the reference lists of the articles initially identified when performing full-text reviewing.

Table 2 Data extraction table

Article information	Title Authors Affiliations with author(s) Main discipline of lead author Year of publication (based on the first identifiable date) Journals on which studies were published and the Journal Citation Reports (JCR) impact factor Language of publication Funding Type of the literature Other potential factors
Population	Sample size Inclusion and exclusion criteria of the individual study Age of the participants Sex of participants Diagnostic codes to define dementia Type of dementia reported Comorbidities or exposures investigated
Methods/results	Study designs reported Time length/follow-up time Subsets of database Whether and what methods used to validate the selection of diagnostic codes to define dementia Whether and what additional methods to increase the likelihood of dementia diagnosis used Statistical methods The number of people with dementia identified Prevalence, if any Main outcomes of the studies and the estimation such as incidence and mortality, if reported Other reported health-related outcomes

Selection process (evidence screening and selection)

Following the search, all identified citations will be collated and exported into the bibliographical software EndNote version X8, where duplicates will be removed. Two researchers (YJS and JYW) will independently screen potentially eligible studies according to the inclusion criteria in our review by screening titles and abstracts yielded by our comprehensive search. The selection results by the two individual researchers will then be compared and merged. Any discrepancy in the results will be solved through discussion by involving a third reviewer (YJY). Reasons for exclusion of full-text studies that do not meet the inclusion criteria and the results of the search in each step will be reported in detail in the final scoping review and presented in PRISMA flow chart diagram.

Data collection process

The research team will develop an electronic data extraction form. This Microsoft Excel data extraction form will be independently pilot tested by all team members for its applicability and the research team will achieve its final version based on the feedback. The targeted data to extract will include the characteristics of the reports such as the diagnostic codes to define dementia, the type of dementia reported, the study designs, the time length, the subsets of database, whether additional methods to increase the likelihood of dementia diagnosis were used, the statistical methods and the outcomes. Other general profiles of the reports include lead authors, affiliations, main discipline of the research team, year of publication, funding, the journals on which the studies were published and other potential factors (see table 2). When extracting the data for the review, a reviewer (YJS) will assess each eligible study and then input the assessed results in the prespecified data extraction form. Another reviewer (JYW) will then randomly examine 80% of the extracted results, and any inconsistency will be resolved through consensus with the involvement of the third reviewer (YJY).

Collating, summarising and reporting the results

As the scoping review intends to map the characteristics of the published reports and to address the diversity of the specific outcome for case study, we will not perform critical appraisal of each individual study. The extracted data will be analysed by two researchers who will work at the same time by following a prespecified analytical plan. In responding to the research questions, the variables mentioned above will first be summarised quantitatively, and descriptive statistics including number counts, frequencies and rates will be used to summarise the results. Then the researchers will perform narrative discussions based on the results to address the mapping and research questions that cannot be answered quantitatively. Considering the vast differences in reporting prevalence (ie, non-reporting, different time frames or ways of calculation) in individual studies, we will not attempt to pool the prevalence with statistical methods on our own, and instead, the reporting conditions of prevalence in each individual study will be presented with potential reasons discussed. We will also perform a subgroup analysis based on the type of literature to compare whether there are differences between the formal and grey literature. Finally, we will abide by the PRISMA-ScR guideline[25] to ensure the reporting standard of the final manuscript, and submit to a peer-reviewed journal.

Patient and public involvement

Our scoping review will not plan to involve patients with dementia or the public, but we will also invite clinicians who have to type in the codes for dementia in their practices in the review process and when consulting, discussing and verifying the results.

Ethics and dissemination

Our scoping review is a review of the published reports which does not involve access to individual data, and ethical approval is not required. The results will be disseminated through peer-reviewed journals and conferences focusing on dementia and related topics. We will also share our results with stakeholders, including the non-governmental organisations for dementia and the policymakers in the field of health informatics, by distributing our results in brief and plain language in the social media and emails. If there is any amendment to the protocol, the revised information will be disclosed on the protocol registry space (OSF: osf.io/fc65g) and will be stated in any future publication based on the review.

DISCUSSION

The proposed scoping review aims to investigate and address the diversity in the published reports targeting dementias based on the NHIRD, and to discuss the potential influence through a case study of the reported prevalence. To the best of our knowledge, our study will be the first review of the studies based on the NHIRD on such topic. We believe that by the strengths of a scoping review methodology, a panorama view of the dementia studies based on the NHIRD will help to inform the researchers of the status quo of the research in its field and thereby improve the quality and value of database research by taking advantage of the NHIRD.

Acknowledgements We especially appreciated the help from the staff in the Department of Healthcare Administration, College of Medical and Health Science and the librarians in Asia University.

Contributors YJS and YJY conceptualised the research. YJS, JYW, YHW and RRS helped the development of ideas and drafted the initial protocol. YJY supervised and revised the manuscript. All authors read and agreed with the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Ying-Jyun Shih http://orcid.org/0000-0002-0714-9532
Yung-Jen Yang http://orcid.org/0000-0003-2852-3252

REFERENCES

1 Konrad R, Zhang W, Bjarndóttir M, *et al*. Key considerations when using health insurance claims data in advanced data analyses: an experience report. *Health Syst* 2019;9:317–25.
2 Thygesen LC, Ersbøll AK. When the entire population is the sample: strengths and limitations in register-based epidemiology. *Eur J Epidemiol* 2014;29:551–8.
3 Anderson G, Oderkirk J, OECD. *Dementia research and care: can big data help? big data approaches to dementia: opportunities and challenges*. Paris: OECD Publishing, 2015.
4 Pike KJ, Fazio S, Bynum JPW, *et al*. Resources, methods, and data infrastructure to promote research in dementia care, caregiving, and services. *J Am Geriatr Soc* 2021;69:1793–800.
5 Reinke C, Buchmann N, Fink A, *et al*. Diabetes duration and the risk of dementia: a cohort study based on German health claims data. *Age Ageing* 2022;51:afab231.
6 Shi L, Steenland K, Li H, *et al*. A national cohort study (2000-2018) of long-term air pollution exposure and incident dementia in older adults in the United States. *Nat Commun* 2021;12:6754.
7 Feder SL. Data quality in electronic health records research: quality domains and assessment methods. *West J Nurs Res* 2018;40:753–66.
8 Rabinstein AA. Administrative medical databases for clinical research: the good, the bad, and the ugly. *Neurocrit Care* 2018;29:323–5.
9 Bohensky MA, Jolley D, Sundararajan V, *et al*. Data linkage: a powerful research tool with potential problems. *BMC Health Serv Res* 2010;10:346.
10 Harron K, Dibben C, Boyd J, *et al*. Challenges in administrative data linkage for research. *Big Data Soc* 2017;4:205395171774567.
11 Davis KAS, Mueller C, Ashworth M, *et al*. What gets recorded, counts: dementia recording in primary care compared with a specialist database. *Age Ageing* 2021;50:2206–13.
12 Jaakkimainen RL, Bronskill SE, Tierney MC, *et al*. Identification of physician-diagnosed alzheimer's disease and related dementias in population-based administrative data: a validation study using family physicians' electronic medical records. *J Alzheimers Dis* 2016;54:337–49.
13 Lipscombe LL, Hwee J, Webster L, *et al*. Identifying diabetes cases from administrative data: a population-based validation study. *BMC Health Serv Res* 2018;18:316.
14 Stein JD, Lum F, Lee PP, *et al*. Use of health care claims data to study patients with ophthalmologic conditions. *Ophthalmology* 2014;121:1134–41.
15 Russell P, Banerjee S, Watt J, *et al*. Improving the identification of people with dementia in primary care: evaluation of the impact of primary care dementia coding guidance on identified prevalence. *BMJ Open* 2013;3:e004023.
16 Zhu CW, Ornstein KA, Cosentino S, *et al*. Misidentification of dementia in medicare claims and related costs. *J Am Geriatr Soc* 2019;67:269–76.
17 Health AIo, Welfare. *Predicting early dementia using Medicare claims: a feasibility study using the National Integrated Health Services Information Analysis Asset*. Canberra: AIHW, 2021.
18 Wilkinson T, Ly A, Schnier C, *et al*. Identifying dementia cases with routinely collected health data: a systematic review. *Alzheimers Dement* 2018;14:1038–51.
19 Munn Z, Peters MDJ, Stern C, *et al*. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol* 2018;18:143.
20 Peters MDJ, Marnie C, Tricco AC, *et al*. Updated methodological guidance for the conduct of scoping reviews. *JBI Evid Synth* 2020;18:2119–26.
21 Tricco AC, Lillie E, Zarin W, *et al*. A scoping review on the conduct and reporting of scoping reviews. *BMC Med Res Methodol* 2016;16:15.
22 Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005;8:19–32.
23 Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci* 2010;5:69.
24 Khalil H, Bennett M, Godfrey C, *et al*. Evaluation of the JBI scoping reviews methodology by current users. *Int J Evid Based Healthc* 2020;18:95–100.
25 Tricco AC, Lillie E, Zarin W, *et al*. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018;169:467–73.
26 Lin L-Y, Warren-Gash C, Smeeth L, *et al*. Data resource profile: the national health insurance research database (NHIRD). *Epidemiol Health* 2018;40:e2018062..
27 Hsieh C-Y, Su C-C, Shao S-C, *et al*. Taiwan's national health insurance research database: past and future. *Clin Epidemiol* 2019;11:349–58.