ABSTRACT

OBJECTIVES: This scoping review aimed to appraise the existing literature on using the claims databases for epidemiological studies and to draw inferences for using data from Pakistan’s health insurance databases.

METHODS: We conducted a scoping review of literature focusing on health insurance databases, querying MEDLINE, EMBASE and Google Scholar. We used the frameworks proposed by the Joanna Briggs Institute and Arksy and O’Malley for mapping our results.

RESULTS: There was a considerable chronological increase in studies published using data from health insurance databases. Most of the studies in our search were from economically developed countries. Most of the studies (n=84) focussed on chronic non-communicable diseases, while a limited number (n=09) focussed on communicable (infectious) diseases. Our findings suggest that insurance databases could be utilised to study rare diseases, prospects of prolonged follow-up, and minimal research costs. This is especially important for countries like Pakistan, having limited resources to conduct regular, population-level epidemiological studies. Several methodological approaches (for instance, disease, pharmacy or intervention classification codes) were presented in these studies to extract epidemiological data from the insurance database.

CONCLUSION: Health insurance databases are utilised as sources for epidemiological studies, predominantly for chronic illnesses, in economically developed countries. Methodological approaches described in these papers could be used to extract data for epidemiological research from health insurance databases in Pakistan. This could be especially useful for following the patterns of infectious disease in the country.

KEYWORDS: Database (MeSH); Database Management Systems (MeSH); Insurance (MeSH); Insurance, Health (MeSH); Universal Health Coverage (Non-MeSH); Sehat Sahulat Programme (Non-MeSH); Social health protection (Non-MeSH); Sehat insaf (Non-MeSH); Pakistan (MeSH).

INTRODUCTION

Many economically-developing countries have a scarcity of reliable health information to improve their health systems. These countries are implementing large-scale health insurance programmes. These programmes (for instance, the Sehat Sahulat Programme in Pakistan) maintain their claims databases. The claims databases of these programmes present an opportunity to bridge the gaps in the district health information system and health information systems for the vertical programmes. The primary purpose of these databases is billing, but these could serve as a valuable source of secondary data for research. They offer large sample sizes, multiple comparative control groups, opportunity to study rare diseases, prospects of prolonged follow-up, and minimal research costs. Insurance claims’ databases hold rich information related to patients. Studies have noted these databases serve as resources for long-term safety surveillance of medical products and health system research. However, we did not find a comprehensive review of using insurance databases for incidence and prevalence studies. We conducted this scoping review to bridge this gap.

Objectives of our scoping review were: (i) To map the epidemiological studies conducted on health insurance databases by time, geography and disease-groups, (ii) to map methodological considerations of using health insurance databases for epidemiological studies, and (iii) to draw inferences for using the health insurance database in Pakistan for similar studies.

METHODS

Identification of studies for inclusion

We looked for peer-reviewed publications reporting on the incidence and prevalence of diseases from insurance databases. We included studies in all age groups and both public and private health insurance programmes. We excluded studies that (i) compared the outcomes of specific interventions, (ii) reported pharmacovigilance data, and (iii) studies not published in English.

We used the frameworks proposed by the Joanna Briggs Institute and Arksy and O’Malley for mapping our results. We adopted a systematic search strategy for two major biomedical
We included studies published in English from the inception of these digital databases until 31 May 2019. There was no geographical or methodological bar on publications. Initially, we queried MEDLINE through PubMed with Medical Subject Headings (MeSH). The Syntax for PubMed was: (“Insurance” [MeSH]) AND “Database Management Systems” [MeSH].

After piloting the syntax with PubMed, it was modified for EMBASE. For EMBASE, we used the Emtree words (database AND insurance). Additionally, we manually selected studies from reference lists of other publications and through Google Scholar. On EMBASE and Google Scholar, we restricted our search to titles only.

**Selection of studies for inclusion**

One researcher (SAK) ran the queries on MEDLINE and EMBASE. The database queries found 899 studies. Another 313 studies were found through Google Scholar and secondary references. After removing duplicates, we had 883 studies.

After identification, two independent researchers (RSK, AJ) screened the papers. A two-step screening process was adopted. First, the titles and abstracts were screened by the two researchers (RSK, AJ). At the screening stage (title and abstracts), 761 studies were removed. The two researchers arrived at a final list of 122 articles for full-text review. The same researchers reviewed the full-text, excluding 29 for a different reason. The final list had 93 studies. Figure 1 shows the PRISMA flow diagram of our literature screening and selection process.

**RESULTS**

There is a considerable, chronological increase in studies published using data from health insurance databases. Majority of the studies are from South Korea, the United States, Taiwan, followed by Japan, France, Canada, and Germany. Two studies from Hungary and one from South Africa were included in the full-text review. The temporal and geographical focus of included studies is reflected in Figure 2.

Around 90% of studies (n=84) focussed on chronic non-communicable diseases (NCDs). Only 10.71% studies (n=9) focussed on communicable (infectious) diseases.

**Descriptive analytics**

We grouped the studies according to the International Classification of...
<table>
<thead>
<tr>
<th>Number of studies</th>
<th>ICD-10 Diagnostic Group</th>
<th>Disease/condition studied</th>
<th>Geographical distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Diseases of the musculoskeletal system</td>
<td>Fractures, Osteoporosis, Rheumatoid arthritis, Ankylosing spondylitis, Gout</td>
<td>South Korea, Japan, Taiwan, France, United States, Canada, Hungary</td>
</tr>
<tr>
<td>13</td>
<td>Diseases of the circulatory system</td>
<td>Myocardial infarction, Cardiac arrhythmias, Heart failure, Atherosclerotic CVDs, Cardiomyopathy, Peripheral arterial disease, Kawasaki disease, Pulmonary artery hypertension</td>
<td>South Korea, United States, Taiwan, France</td>
</tr>
<tr>
<td>12</td>
<td>Diseases of the nervous system</td>
<td>Amyotrophic Lateral Sclerosis, Dementia, Multiple sclerosis, Parkinson's disease, Motor neuron disease, Myasthenia gravis, Stroke, Trigeminal neuralgia</td>
<td>United States, France, South Korea, Taiwan</td>
</tr>
<tr>
<td>11</td>
<td>Endocrine, nutritional and metabolic diseases</td>
<td>Diabetes mellitus, Diabetes and cancer, Addison's disease, Metabolic syndrome</td>
<td>Taiwan, South Korea, Germany, Canada, United States</td>
</tr>
<tr>
<td>09</td>
<td>Certain infectious and parasitic diseases</td>
<td>Clostridium difficile, Herpes zoster, Influenza, Invasive pneumococcal disease, Non-tuberculous mycobacterial pulmonary disease, Pyogenic liver abscess, Infectious diseases in Amyotrophic Lateral Sclerosis</td>
<td>United States, South Korea, Japan, France, Taiwan</td>
</tr>
<tr>
<td>08</td>
<td>Mental and behavioural disorders</td>
<td>Depression, ADHD, Alzheimer's disease, Psychiatric illnesses in survivors of critical illnesses</td>
<td>Taiwan, South Korea, South Africa, Germany</td>
</tr>
<tr>
<td>06</td>
<td>Diseases of the skin and subcutaneous tissue</td>
<td>Psoriasis, Chronic wounds, Stevens-Johnson syndrome &amp; toxic epidermal necrolysis</td>
<td>Japan, Taiwan, South Korea, United States</td>
</tr>
<tr>
<td>05</td>
<td>Diseases of the respiratory system</td>
<td>Asthma, Idiopathic pulmonary fibrosis (IPF)</td>
<td>United States, South Korea, Canada</td>
</tr>
<tr>
<td>04</td>
<td>Diseases of the blood and blood-forming organs and immune mechanism disorders</td>
<td>Haemophilia, Hereditary haemorrhagic telangiectasia</td>
<td>Taiwan, United States</td>
</tr>
<tr>
<td>02</td>
<td>Diseases of the genitourinary system</td>
<td>Renal dysfunction in stroke patients, Renal replacement therapy</td>
<td>Taiwan, United States</td>
</tr>
</tbody>
</table>
The outcomes of interest were incidence rates, the incidence of complications, disease prevalence, rate ratio, crude hazard ratio (HR), adjusted HR, adjusted mortality risk and cost of care.

Limitations of database-enabled studies

Studies in our review reported that medical records were not available to confirm the diagnosis. Enrolling patients through only ICD code reported much different incidence rates from the previously known. Enrolling patients through a combination of ICD codes with prescription codes rendered results comparable to previous studies. Using ICD-10 diagnostic code in combination with prescription code would enhance specificity at the cost of sensitivity, by dropping many true positive cases. Using health insurance databases for epidemiological research include the likelihood of underestimating the prevalence of the disease. Also, there might be residual confounders like factors related to lifestyle and disease severity. The secondary data could not account for such confounders.

The private health insurance databases (like in the US) are not a truly random sample of the population. Based on appearance ICD diagnostic codes, the validity of records depends on the physician’s diagnosis and accuracy of administrative coding for the encounter. People who did not seek medical care were missed.

DISCUSSION

Health information systems are one of the key areas for strengthening health systems. But economically developing countries, including Pakistan, have limited capacities to gather real-time information on the population’s health status.

These databases’ primary purpose is billing and insurance administration. Insurance databases map insurance claims to disease treated and interventions performed. Indirectly, insurance databases capture high-quality clinical data, which can be used to help improve patient outcomes. Therefore, in countries like Pakistan, insurance databases could be a valuable source of information. For instance, in 2021 alone, one provincial insurance programme in Pakistan reported half-a-million hospital admission with their data electronically captures.

The European Medicines Agency (EMA) and the United States’ Food and Drug Administration (FDA) authority consider claims databases a good source of data for drugs’ safety surveillance. This practice could be adopted in Pakistan too. The insurance programmes in Pakistan cover high-cost medical devices like cardiac stunts, pacemakers and chemotherapeutic agents. Similarly, the Social Insurance Medical Fee Payment Fund (MFPP) in Japan utilises its claims data for non-accounting purposes. The MFPP extracts valuable, population-level epidemiological information from insurance claims data.

Considering the programmes in Pakistan also cover infectious diseases, these could be used to follow disease spread (epidemics) across time and space.

In other countries, efforts are underway to link insurance databases to other data sources like vital statistics and disease registries. These linkages can extend the key areas for strengthening health systems.
Using health insurance databases for epidemiological research: A scoping review

The range of available data elements. The FDA under its sentinel initiative has linked the FDA certification databases with that of claims databases and disease registries to monitor the long-term safety of FDA-regulated medical products. In Pakistan, linking the health insurance databases with the poverty database of the Benazir Income Support Programme could provide valuable insights into the socioeconomic gradient of disease spread.

Though our findings suggest that the insurance databases in Pakistan might enable similar epidemiological studies, the researchers could not ascertain if these programmes capture the finer granular data. Our study did not focus on what data fields the databases in each country capture. Therefore, we could not describe how the insurance database in Pakistan should be structured. Future researchers should look at the data models and structures of the best insurance database worldwide to draw inferences for the database in Pakistan.

**CONCLUSION**

This review shows that the epidemiological studies on health insurance databases come from a limited number of economically-developed countries. Most of the literature arises from the national health insurance databases in Asia (Japan, Taiwan, and South Korea), Europe (France and Germany) and the private databases in the United States. Our review found that most studies come from economically developed countries, mainly concerned with chronic non-communicable diseases.

We did not find any studies from economically developing countries. Besides, we observed the lack of literature based on health insurance databases on infectious diseases facing economically developing countries. Nonetheless, with the emergence of insurance programmes as a key strategy in many developing countries to improve access to health care, working with their claims database could fill this gap. For instance, as highlighted in the discussion, the insurance databases in Pakistan could help us see the socioeconomic gradient in diseases and follow the patterns of infectious diseases across time and space.

**REFERENCES**


general_and_Khyber_Pakhtunkhwa_education_of_their_proximity_to_Universal_Health_Coverage


15. Chen YC, Yeh HY, Wu JC, Haschler I, Chen TJ, Wetter T. Taiwan’s National Health Insurance Research Database: Administrative health care database as study object in bibliometrics. Scientometrics
10.3346/jkms.2018.33.e60
https://doi.org/10.3346/jkms.2018.33.e60


37. Son MK, Lim N-K, Cho M-C, Park H-Y. Incidence and risk factors for atrial fibrillation in Korea: the National


123. Hertzberg J. Disease management and disease registries: opportunities for MCOs. Manag Care Interface 2001;14(10):70,75.


AUTHOR'S CONTRIBUTION
Following authors have made substantial contributions to the manuscript as under:

SAK: Conception & study design, analysis and interpretation of data, drafting the manuscript, critical review, approval of the final version to be published

RSK & AJ: Acquisition, analysis and interpretation of data, drafting the manuscript, critical review, approval of the final version to be published

Authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICT OF INTEREST
Authors declared no conflict of interest

GRANT SUPPORT AND FINANCIAL DISCLOSURE
Authors declared no specific grant for this research from any funding agency in the public, commercial or non-profit sectors

DATA SHARING STATEMENT
The data that support the findings of this study are available from the corresponding author upon reasonable request

This is an Open Access article distributed under the terms of the Creative Commons Attribution-Non Commercial 2.0 Generic License.