J Korean Med Sci. 2025 Mar 3;40(8):e111
https://doi.org/10.3346/jkms.2025.40.e111
eISSN 1598-6357·pISSN 1011-8934

JKMS

Review Article
Medicine General & Health Policy

Leveraging National Health Insurance Service Data for Public Health Research in Korea: Structure, Applications, and Future Directions

Seung-Ji Lim [1] **and Sung-In Jang** [1,2]

[1]Health Insurance Research Institute, National Health Insurance Service, Wonju, Korea
[2]Institute of Health Services Research, Yonsei University College of Medicine, Seoul, Korea

Received: Feb 13, 2025
Accepted: Feb 16, 2025
Published online: Feb 27, 2025

Address for Correspondence:
Sung-In Jang, MD, PhD
Health Insurance Research Institute, National Health Insurance Service, Wonju 26464, Republic of Korea.
Email: jangsi@yuhs.ac
jangsi@nhis.or.kr

ORCID iDs
Seung-Ji Lim
https://orcid.org/0000-0001-5954-9629
Sung-In Jang
https://orcid.org/0000-0002-0760-2878

Disclosure
The authors have no potential conflicts of interest to disclose.

Disclosure of Artificial Intelligence (AI)-Assisted Technology
Artificial intelligence (AI)-assisted technology was utilized in both the writing and editing processes.

Author Contributions
Conceptualization: Lim SJ, Jang SI.

ABSTRACT

The National Health Insurance Service (NHIS) database serves as a crucial resource for public health research in Korea. As a comprehensive dataset within the single-payer healthcare system, NHIS data provides longitudinal insights into healthcare utilization, disease prevalence, and health outcomes. This review article explores the structure, characteristics, and applications of NHIS data, emphasizing its role in epidemiological studies, health policy evaluations, and clinical research. We discuss key methodological considerations, including data access procedures, outcome measures, and strategies to mitigate bias. Additionally, we highlight future directions, such as integrating NHIS data with other national health datasets and utilizing artificial intelligence for predictive analytics. By leveraging the NHIS database, researchers can enhance evidence-based policymaking and improve public health outcomes in Korea.

Keywords: National Health Insurance Service; Public Health Research; Epidemiology; Big Data; Health Policy; Healthcare Utilization

The National Health Insurance Service (NHIS) database forms a vital foundation for public health research in South Korea. As part of a single-payer healthcare system, the NHIS collects comprehensive, population-level data from diverse operational sources (**Fig. 1**), enabling longitudinal studies and detailed analyses.[1] A notable strength of the NHIS data lies in its health examination dataset, which provides extensive clinical and lifestyle information. Additionally, the individual cohort datasets support targeted research on specific groups, such as older adults, children, and individuals with chronic diseases, thereby fostering tailored public health insights.

This review aims to outline the structure and characteristics of the NHIS database. It also examines methodological considerations, key variables, and analytical approaches for utilizing these databases effectively. By highlighting the central role of the NHIS system and its data, this review underscores its critical contribution to advancing public health research and understanding.

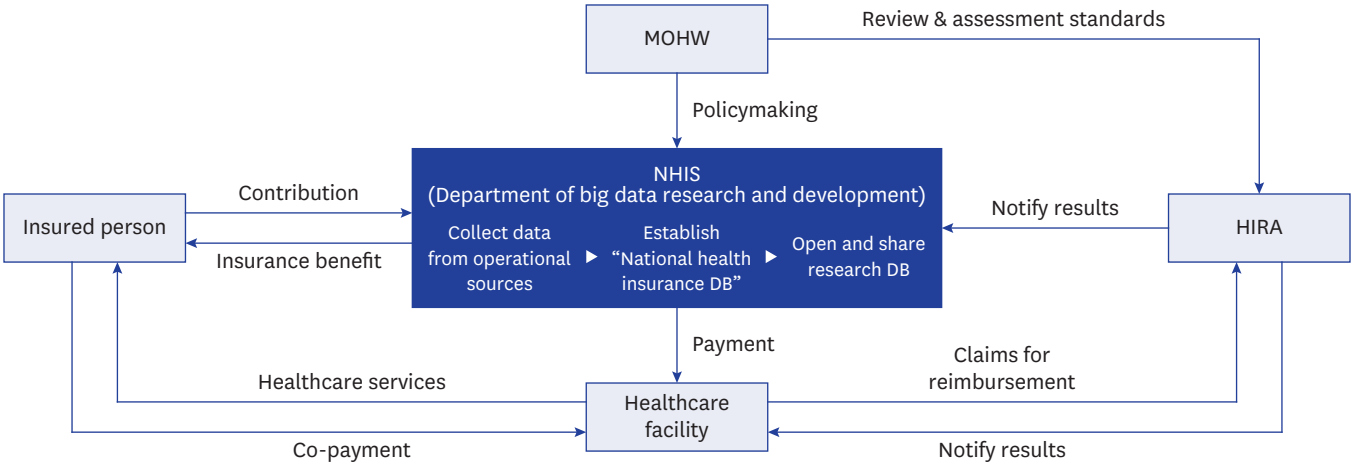

Fig. 1. Data generation and operational process of the NHIS.
NHIS = National Health Insurance Service, MOHW = Ministry of Health and Welfare, HIRA = Health Insurance Review and Assessment Service.

HISTORICAL DEVELOPMENT OF THE NATIONAL HEALTH INSURANCE SERVICE

Before the establishment of the NHIS in South Korea, the country relied on multiple voluntary health insurance cooperatives. These cooperatives, formed regionally and within workplaces under the 1963 National Health Insurance Act, faced significant challenges, including disparities in financial stability and service quality due to variations in income levels and health conditions. Policymakers and academics debated whether to retain the fragmented cooperative model or transition to a unified system. Integration was eventually implemented in phases, with organizational consolidation achieved in 1989 and 2000, and financial unification completed in 2002 (**Fig. 2**).[2]

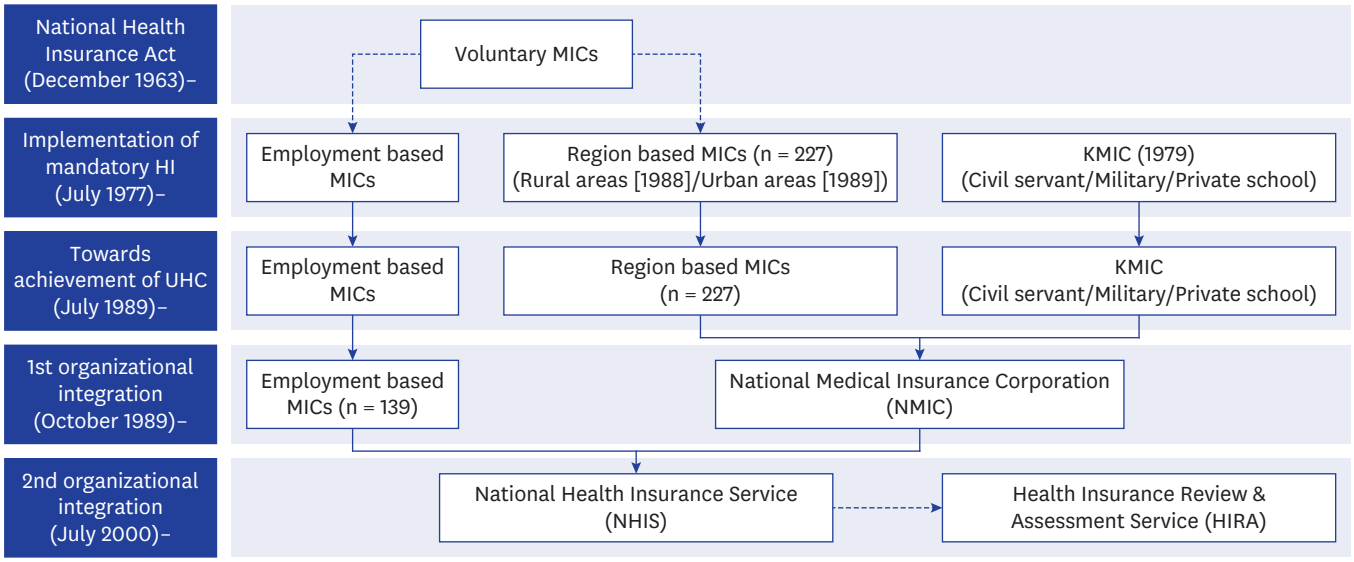

Fig. 2. Timeline of integration for health insurance cooperatives (redrawn from the World Bank material).
MIC = Medical Insurance Cooperative, KMIC = Korea Medical Insurance Corporation.

https://doi.org/10.3346/jkms.2025.40.e111

In 2013, the NHIS established the "Health Insurance Big Data Operations Center," marking its initial foray into big data management. This was followed by the creation of the "Department of Big Data Management" in 2014 to systematically oversee data openness, sharing, and integration. Recognizing the growing importance of big data, the NHIS reorganized its structure in 2020, establishing the "Department of Big Data Strategy" staffed with over 100 dedicated personnel to enhance data management and utilization.[3]

TYPES AND CHARACTERISTICS OF NATIONAL HEALTH INSURANCE SERVICE DATA

The NHIS provides two primary types of databases for research: the Sample Research Database and Customized Database. The Sample Research Database includes the Sample Cohort Database, Health Screening Cohort Database, and Elderly Cohort Database. The Working Women Cohort Database and Infant Medical Check-up Cohort were discontinued due to a decline in research demand and the inability to update these databases.[4,5] For specific research needs, researchers are encouraged to apply for the Customized Database, which provides tailored datasets designed to address particular study requirements. Details of these databases are presented in **Supplementary Table 1**. Researchers can apply for access through the National Health Insurance Sharing Service (NHISS) website (http://nhiss.nhis.or.kr). Applications must include a detailed study protocol and receive approval from an institutional review board. Upon submission, the application is reviewed by the Data Provision Review Committee, which notifies the applicant of the outcome. Following approval, researchers are required to pay the associated fees in order to access the requested database. Access to the Customized Database is restricted to designated locations, and the data provided are limited to the variables specified in the approved research protocol. Raw data cannot be directly accessed or transferred. In contrast, remote access and analysis are available for the Sample Research Database.[6] According to NHISS guidelines, the standard review process is expected to take approximately 25 days. However, due to the increasing demand for NHIS data, researchers are advised to anticipate potential delays and plan accordingly when allocating time for their studies.

OUTCOME MEASURES AND COVARIATES

As outlined in **Supplementary Table 1**, NHIS data encompass various categories, including qualification, treatment, health screenings, medical care institutions, and long-term care for the elderly. Qualification data include variables such as age, gender, location, subscription type, and socioeconomic factors like income, disability status, and death. Treatment data covers statements (T20), treatment details (T30), disease type (T40), and prescription information (T60) from medical, dental, oriental, and pharmacy services. Health examination and questionnaire variables have evolved over time, including blood test results and health behavior data.[6]

To conduct clinical research using NHIS data, it is essential to first define how exposures and outcomes are measured and then address strategies to control for confounding factors.[7] Clinical outcomes, such as diagnoses, surgeries, or mortality, are identified using specific codes through operational definitions. These outcomes can be defined solely by Korean Classification of Diseases (KCD) codes or supplemented with biometric data collected during

health examinations. Blood pressure, body mass index, glucose, and cholesterol levels are among the most frequently used health metrics, providing key insights into chronic conditions such as hypertension, diabetes, and dyslipidemia. For example, chronic kidney disease can be defined by an admission record or outpatient visits coded N18 or N19, in combination with an eGFR < 60 mL/min/1.73 m^2.[6] Comorbidities are often quantified using tools like the Charlson Comorbidity Index or the Elixhauser Index to capture the burden of coexisting conditions and their impact on health outcomes. The Charlson Index assigns weights to various comorbid conditions based on their association with mortality,[8] while the Elixhauser Index uses a broader range of comorbidities, focusing on conditions that influence hospital resource utilization.[9] Sociodemographic and lifestyle factors are critical covariates that contextualize clinical data. Lifestyle variables include smoking status, alcohol consumption, and levels of physical activity, which are major risk factors for chronic diseases. Sociodemographic data, such as age, gender, income level, and area of residence, provide insights into health disparities and the influence of social determinants on health outcomes. Medication data, including prescription records (Anatomical Therapeutic Chemical code) and adherence measures like the Medication Possession Ratio and Proportion of Days Covered, provide insights into treatment patterns and adherence. Service utilization data, such as inpatient and outpatient visits and emergency care usage, help assess access to and quality of healthcare services. Additionally, cost-related variables, including total healthcare expenditures, co-payments, and exemptions for severe conditions, enable cost-effectiveness analyses and evaluations of financial barriers to care. Together, these key variables and covariates provide a comprehensive framework for conducting robust analyses using NHIS data, supporting diverse research objectives from epidemiological studies to health policy evaluations.

STUDY DESIGNS AND RESEARCH EXAMPLES

The utilization of NHIS data has grown substantially since it became publicly accessible, largely due to its cost-effectiveness and efficiency in facilitating administrative claims-based research. Given the large size of the database, it provides unique opportunities to study rare diseases, treatment complications, and specific populations, such as the elderly.[6] Nevertheless, certain limitations must be acknowledged, including potential inaccuracies in disease definitions, incomplete datasets, and privacy concerns.[10] Researchers should carefully consider these factors when utilizing the NHIS database. Observational research using NHIS data can be categorized into cross-sectional, case-control, and cohort studies.[11] For instance, if a study begins with an exposure-eg, treatment history (medication use or surgical procedure) or health screening indicators (hypertension or obesity)- and subsequently follows individual over time to assess outcomes-such as disease onset, mortality, or healthcare utilization (hospital admissions or emergency room visits) it is classified as a cohort study. In contrast, if an analytical study begins with an outcome and then looks back in time to identify potential exposures, it is considered a case-control study.[12] Additionally, the NHIS database is suited for other study designs, including survival analysis, time series analysis, and cost-effectiveness studies.

Cross-sectional studies

Cross-sectional studies are designed to examine the presence or absence of a disease and its relationship to an exposure at a single point in time.[12] These studies focus on estimating prevalence, making them suitable for identifying population-level health patterns and associations between health conditions and exposures. However, because both the outcome

and exposure are measured simultaneously, temporal relationships between the two cannot be established.

One study utilizing a customized NHIS database identified an association between vestibular loss and an increased risk of dementia in older adults.[13] Another study, based on the NHIS Sample Cohort database, examined trends and risk factors for severe hypoglycemia in individuals with type 2 diabetes in Korea.[14] These studies are critical for hypothesis generation and for identifying potential areas of intervention, although causal inferences are limited.

Cohort studies

Cohort studies follow groups of individuals over time to examine associations between exposures and outcomes. They are commonly used in claim database research.[15] Cohort studies can be divided into retrospective and prospective types. Retrospective cohort studies look backward in time, utilizing existing NHIS data to examine exposure-outcome relationships and can be useful in evaluating drug safety in specific populations.[16] While prospective cohort studies look forwards in time, as NHIS adds new data with each round of health exams, it allows for the inclusion of updated information over time, enhancing the utility of these studies.

Case-control studies

Case-control studies are observational designs that compare individuals with a specific condition (cases) to those without it (controls), to identify potential risk factors or exposures associated with the condition. These studies are particularly useful for investigating rare diseases or outcomes, as they enable efficient data collection from a smaller sample size compared to cohort studies.[12] A recent case-control study using NHIS data examined the relationship between iron deficiency anemia (IDA) and related disorders in premenopausal women. Researchers identified women diagnosed with IDA as cases and matched them with controls using propensity score matching (PSM). The study found that gynecological diseases, particularly leiomyoma of the uterus and adenomyosis, were significantly more prevalent in the IDA group compared to the control group.[17] This example illustrates how case-control studies can utilize large-scale health claim databases like NHIS to uncover associations between conditions and potential risk factors, aiding in the development of targeted prevention and treatment strategies.

Survival analysis

Survival analysis is used to examine time-to-event outcomes, such as disease onset or mortality, and to assess factors influencing the timing of these events.

For instance, a study explored survival rates among patients with prostate cancer who received primary androgen deprivation therapy (ADT) compared to those who underwent radical prostatectomy. Researchers used survival models to evaluate the time to treatment failure and overall survival between both groups. The study demonstrated that ADT may have a different impact on survival based on patient characteristics such as age and comorbidities. This highlights the utility of survival analysis in comparing treatment strategies and guiding clinical decisions.[18]

Time series analysis

Time series analysis evaluates trends in health metrics or healthcare utilization over time, often in response to policy changes or interventions. A recent interrupted time series

analysis investigated the effects of reimbursement policy changes on the incidence of chronic periodontitis-related procedures using the NHIS Sample Cohort database. The analysis revealed a significant change in the incidence of chronic periodontitis-related procedures following the policy change, demonstrating how time series analysis can effectively evaluate the impact of healthcare policies on treatment patterns.

Cost-effectiveness studies

These studies combine cost data with clinical outcomes to assess the effectiveness of healthcare interventions. For instance, a study compared the cost-effectiveness of laparoscopic versus open pancreatic resection by examining both medical costs and the quality-adjusted life years gained. This enables policymakers to determine whether the higher initial cost of the laparoscopic approach is warranted by its long-term health benefits.[19]

METHODOLOGICAL CONSIDERATIONS

Bias is a critical concern in research utilizing secondary healthcare databases, as it can significantly undermine the validity and generalizability of findings. Methodological rigor is essential to identify, mitigate, and address various types of biases that arise in such studies. With the increasing reliance on secondary data in research, numerous studies emphasize the need for caution due to the heightened risk of introducing biases (confounding, selection, information) during study design, data analysis, and interpretation.[20,21] Selection bias arises when the participants included in the study are not representative of the target population, leading to systematic differences between selected and non-selected individuals. Randomly assigning participants to different groups is the best way to prevent selection bias. However, it is not possible to randomize insurance data because participants have already chosen their treatments. Additionally, it is challenging to avoid differences between individuals who have been health screened and those who have not, which can further introduce selection bias. Confounding occurs when an extraneous variable related to both the exposure and the outcome distorts the observed association between them.[22] Strategies like PSM and inverse probability of treatment weighting (IPTW) can be employed to address confounding in observational studies. PSM involves pairing treated and untreated subjects with similar propensity scores, effectively balancing covariates between groups. IPTW, on the other hand, assigns weights to subjects based on the inverse probability of receiving the treatment they actually received, creating a synthetic sample where treatment assignment is independent of observed baseline covariates.[23] Another consideration when using secondary data is the validation of diagnosis codes and related algorithms. Since NHIS data often relies on administrative codes to capture diagnoses, procedures, and treatments, ensuring the accuracy of these codes is critical for valid analysis. A key step in improving the utility of the NHIS database is refining the algorithms used to define specific conditions by incorporating both diagnostic and treatment codes.

FUTURE DIRECTIONS AND LIMITATIONS

One promising direction for the NHIS database is its integration with other datasets, such as cause-of-death data from Statistics Korea, and cohorts like Korean Genome and Epidemiology Study (KoGES), Korea Nurses Health Study (KNHS), Korean Neonatal Network (KNN), Korean Frailty and Aging Cohort Study (KFACS), and the Community

Health Survey. This integration enables multidimensional analyses, offering deeper insights into long-term health outcomes, social determinants of health, and disease progression. For instance, linking NHIS data with mortality records enhances understanding of disease trajectories, while combining it with the Community Health Survey provides insights into the influence of socioeconomic factors. Moreover, the application of big data analytics and artificial intelligence (AI) further holds transformative potential, enabling predictive modeling and the advancement of precision medicine. AI-driven analyses can help identify high-risk populations, predict disease onset, and refine healthcare delivery strategies.

However, enhancing data reliability remains crucial. Standardizing health examination protocols can help reduce variability in measurement methods across different facilities and time periods. Additionally, expanding validation studies that compare NHIS data with clinical records is essential to assess and improve the validity of variables such as diagnosis codes, laboratory results, and prescription data.

Despite its strengths, research using NHIS data has limitations. Privacy concerns increase as data integration expands, necessitating robust anonymization and governance frameworks. Policy changes, such as restrictions on data sharing or updates to screening guidelines, can impact data availability and research scope. Observational studies using administrative data are prone to confounding, making causal relationships difficult to establish. Selection bias is another issue, as health screening participants may differ systematically from non-participants, skewing results. The exclusion of uninsured services further limits analytical comprehensiveness. Addressing these challenges through careful study design and advanced statistical techniques is essential for enhancing the validity of NHIS-based research. By overcoming these limitations, the NHIS database can continue to advance public health research and inform healthcare policy in Korea and beyond.

CONCLUSION

The NHIS database provides invaluable resources for diverse research objectives, enabling comprehensive analyses of clinical, socioeconomic, and public health factors. Future studies should leverage advanced methodologies and technology while addressing ethical and policy challenges to maximize the potential of NHIS data for public health research.

SUPPLEMENTARY MATERIAL

Supplementary Table 1
DB type and contents

REFERENCES

1. Seong SC, Kim YY, Khang YH, Heon Park J, Kang HJ, Lee H, et al. Data resource profile: the National Health Information Database of the National Health Insurance Service in South Korea. *Int J Epidemiol* 2017;46(3):799-800. PUBMED
2. Wee HS, Jung S, Lee J. *Republic of Korea - World Bank Group Partnership On COVID-19 Preparedness and Response: How Korea's National Health Insurance (NHI) Responded to the COVID-19 Pandemic - Policy Note (English)*. Washington, D.C., USA: World Bank Group; 2023.

3. NHIS. 2024 Booklet for the Introduction of National Health Insurance System. https://www.nhis.or.kr/english/wbheaa03500m01.do?mode=view&articleNo=10840421&article.offset=0&articleLimit=10. Updated December 27, 2023. Accessed January 20, 2025.
4. NHISS. https://nhiss.nhis.or.kr/. Updated 2025. Accessed January 15, 2025.
5. Park I. How to use health insurance data effectively for healthcare research. *J Health Inform Stat* 2022;47 Suppl 2:S31-9. CROSSREF
6. Kim MK, Han K, Lee SH. Current trends of big data research using the Korean National Health Information Database. *Diabetes Metab J* 2022;46(4):552-63. PUBMED | CROSSREF
7. Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report--Part II. *Value Health* 2009;12(8):1053-61. PUBMED | CROSSREF
8. Jørgensen TL, Hallas J, Land LH, Herrstedt J. Comorbidity and polypharmacy in elderly cancer patients: the significance on treatment outcome and tolerance. *J Geriatr Oncol* 2010;1(2):87-102. CROSSREF
9. Yurkovich M, Avina-Zubieta JA, Thomas J, Gorenchtein M, Lacaille D. A systematic review identifies valid comorbidity indices derived from administrative health data. *J Clin Epidemiol* 2015;68(1):3-14. PUBMED | CROSSREF
10. Ahn EK. A brief introduction to research based on real-world evidence: considering the Korean National Health Insurance Service database. *Integr Med Res* 2022;11(2):100797. PUBMED | CROSSREF
11. Kim S, Kim MS, You SH, Jung SY. Conducting and reporting a clinical research using Korean Healthcare Claims Database. *Korean J Fam Med* 2020;41(3):146-52. PUBMED | CROSSREF
12. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. *Lancet* 2002;359(9300):57-61. PUBMED | CROSSREF
13. Lim SJ, Son S, Chung Y, Kim SY, Choi H, Choi J. Relationship between vestibular loss and the risk of dementia using the 2002–2019 national insurance service survey in South Korea. *Sci Rep* 2023;13(1):16746. PUBMED | CROSSREF
14. Lee SE, Kim KA, Son KJ, Song SO, Park KH, Park SH, et al. Trends and risk factors in severe hypoglycemia among individuals with type 2 diabetes in Korea. *Diabetes Res Clin Pract* 2021;178:108946. PUBMED | CROSSREF
15. Fujinaga J, Fukuoka T. A review of research studies using data from the administrative claims databases in Japan. *Drugs Real World Outcomes* 2022;9(4):543-50. PUBMED | CROSSREF
16. Kim D, Yang PS, Sung JH, Jang E, Yu HT, Kim TH, et al. Risk for osteoporotic fractures in patients with atrial fibrillation using different oral anticoagulants. *International Journal of Arrhythmia* 2021;22(1):4. CROSSREF
17. Lee HJ, Pak H, Han JJ, Chang MH. Comprehensive analysis of iron deficiency anemia and its related disorders in premenopausal women based on a propensity score matching case control study using National Health Insurance Service Database in Korea. *J Korean Med Sci* 2023;38(37):e299. PUBMED | CROSSREF
18. Ha US, Choi JB, Shim JI, Kang M, Park E, Kang S, et al. Is primary androgen deprivation therapy a suitable option for Asian patients with prostate cancer compared with radical prostatectomy? *J Natl Compr Canc Netw* 2019;17(5):441-9. PUBMED | CROSSREF
19. Lee JS, Oh HL, Yoon YS, Han HS, Cho JY, Lee HW, et al. Cost-effectiveness of open versus laparoscopic pancreatectomy: a nationwide, population-based study. *Surgery* 2024;176(2):427-32. PUBMED | CROSSREF
20. Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC Med Res Methodol* 2019;19(1):53. PUBMED | CROSSREF
21. Hoffman SR, Gangan N, Chen X, Smith JL, Tave A, Yang Y, et al. A step-by-step guide to causal study design using real-world data. *Health Serv Outcomes Res Methodol* 2024. CROSSREF
22. Hyman J. The limitations of using insurance data for research. *J Am Dent Assoc* 2015;146(5):283-5. PUBMED | CROSSREF
23. Chesnaye NC, Stel VS, Tripepi G, Dekker FW, Fu EL, Zoccali C, et al. An introduction to inverse probability of treatment weighting in observational research. *Clin Kidney J* 2021;15(1):14-20. PUBMED | CROSSREF