Adjusting for confounders in outcome studies using Korea National Health Insurance Claim Database: A review of methods and applications

Running title: Adjusting confounders in outcome studies
Abstract

Objectives: Adjusting for potential confounders is key to producing valuable evidence in outcome studies. Although numerous studies have been published using the Korea National Health Insurance Claim Database, no study has critically reviewed the methods to adjust for confounders. This study aimed to review these studies and suggest methods and applications to adjust for confounders.

Methods: We conducted a literature search on electronic databases, including PubMed and EMBASE, from January 1, 2021 to December 31, 2022. In total, 278 studies were retrieved. Eligibility criteria were published in English and outcome studies. Literature search and article screening were independently performed by two authors and finally, 173 of 278 studies were included.

Results: Thirty nine studies used matching at the study design stage, and 171 adjusted for confounders using regression analysis or propensity scores at the analysis stage. Of these, 125 underwent regression analysis based on study questions. Propensity score matching was used most frequently among the propensity score methods. A total of 171 studies included age and/or sex as confounders. Comorbidities and healthcare utilization, including medications and procedures, were used as confounders in 146 and 82 studies, respectively.

Conclusions: This is the first review to address the methods and applications to adjust for confounders in recently published studies. Our results indicate that all studies adjusted for confounders with appropriate study designs and statistical methodologies; however, understanding and careful application of confounding variables are required to avoid erroneous results.

Keywords: Confounder, Risk adjustment, Statistical methodology, Health Insurance Claim Database
Introduction

Use of real-world data (RWD) to support evidence-based decision-making and improve patient health outcomes have been emphasized in health care [1]. Unfortunately, in the process, causal relationships can be misleading because of potential confounders that obscure the real effect. Therefore, adjusting for the potential confounders is the key to producing valuable evidence in outcome studies using RWD.

The National Health Insurance Claim Database (NHICD), a representative RWD in South Korea, is widely used as a data source in health outcome studies. The database is collected primarily for reimbursement by the Health Insurance Review and Assessment Service, and includes the general characteristics of patients and healthcare providers and information on health service utilization. Controlling confounders is a core process when conducting health outcome studies using the NHICD. Several methods can be applied to adjust for confounders at study design and analysis stages [2]. The basic methods used to address confounders at the study design stage include subject restriction, stratification, and matching. In the analysis stage, there are regression analysis and propensity score (PS) methods. Kim et al. compared kidney cancer incidence and mortality of patients who used statins with those who did not use them matched by age and sex [3]. Oh et al. investigated the relationship between endocrine treatment and risk of depression in patients with breast cancer using a Cox proportional hazard model after adjusting for age and comorbidities [4].

Systematic approaches, including well-designed studies and employment of proper statistical methods, are required to translate the data from the NHICD into meaningful real-world evidence. Although numerous studies have been published using the data from the NHICD [5], none has critically reviewed the methods to adjust for potential confounders.
Therefore, in this study, we aimed to review recent studies and suggest methods and applications to adjust for confounders.

**Methods**

**Searching for studies**

We conducted a literature search on PubMed and EMBASE between January 1, 2021 and December 31, 2022. The search strategy was designed to retrieve all published articles without any restrictions. We used the terms (“Health Insurance Claim Data” or “Health Insurance Claim Database”) and (“South Korea” or “Korea”), and 278 studies were retrieved. Inclusion criteria were: (1) originally published in English and (2) outcome studies investigating the causal relationship between exposure and outcomes.

**Selecting studies**

Literature search and article screening were independently performed by KHK and SJH. After removing duplicates, both authors screened the initial search results for relevance using the titles and abstracts. Full texts were reviewed against the eligibility criteria. There was no substantial disagreement between the authors regarding the inclusion. Three studies without full text, three review articles, seven international studies, eighty one epidemiological studies, three economic evaluation studies, and eight other studies such as prediction model development were excluded. Finally, 173 studies were included for analysis (Figure 1).

**Ethics Statement**
Ethical approval for this study was waived by the Institutional Review Board of the Health Insurance Review and Assessment Service (No. 2023-040-001) in accordance with the waiver reason “study using information open to the public.”

Results

Methods to adjust for confounders

All 173 studies applied a restriction method with exclusion criteria to preliminarily control for confounders during the study design stage (Table 1). Thirty nine studies used the 1:1 or 1:n matching method with age and/or sex, and some additionally considered comorbidity [6-8] and index date year [7,9-13] as matching variables. Of these 39 studies, only two applied univariate analysis such as Chi-squared test and t-test to a matched study cohort. Choi et al. compared the outcomes of patients using univariate Cox proportional hazard model after the effects of entecavir and tenofovir disoproxil fumarate groups were matched 1:2 based on age, sex, comorbidities, hospital type, and index date year [7]. Yoon et al. investigated the risk of cancer between ranitidine and famotidine users without risk adjustment after matching based on age, sex, diabetes mellitus, and cumulative exposure [8]. The remaining 37 studies controlled for confounders by applying regression analysis and/or PS methods at the analysis stage. Jeong et al. evaluated whether women with moyamoya disease have an increased risk of cerebrovascular disease after delivery [14]. This study matched the cohort by the ratio of 1:10 based on age and parity, and applied Cox proportional hazard model adjusting for confounding variables. Similarly, Kim et al. extracted data on users and non-users of statin in a 1:4 ratio based on age from the NHICD, and used a Cox proportional hazard model adjusting for confounding variables to compare the incidence and mortality of breast cancer between the two groups [15].
Among the 173 studies, 171, excluding the two that applied matching method at the design stage, adjusted for confounders using regression analysis and/or the PS method at the analysis stage (Table 1). Of these, 125 applied regression analysis such as logistic regression, Cox proportional hazard model, and Poisson regression, depending on the study questions. The PS method was used in 46 studies including 34 that applied both regression analysis and the PS method. There are four ways of using the estimated PS: propensity score matching (PSM), stratification on the PS, inverse probability of treatment weighting (IPTW) using the PS, and covariate adjustment using the PS [16]. There were 38 studies that applied PSM, which was used the most among the PS methods, and no studies that used covariate adjustment. Studies using PSM calculated PS using logistic regression adjusted for confounders and performed 1:1 or 1:n greedy nearest-neighbor matching based on the estimated PS. Park et al. applied PSM to balance the aforementioned baseline characteristics between the prophylaxis and non-prophylaxis groups and then, conducted univariate analysis with Chi-squared or Student’s t-test to estimate the incremental burden in the matched cohort [17].

Five studies employed Cox proportional hazard model using IPTW [18], and one used IPTW after applying PSM [19]. Song et al. compared the risk of dyslipidemia between patients diagnosed with thyroid cancer and the general population using IPTW and stratification on PS [20]. They estimated PS for each participant with logistic regression, and divided the data into three strata defined by tertiles of PS. Stratified Cox proportional and Cox proportional hazard models were employed within each stratum using IPTW. Thirty four studies applied regression analysis adjusting for confounding variables in the matched study cohort after performing PSM. Yun et al. used PSM to investigate the risk of osteoporotic fractures after gastrectomy [21]. However, they additionally employed a Cox proportional hazard model adjusting for confounding variables owing to the imbalance between the matched cohorts.
Application of potential confounders

The list and frequencies of confounders used in the study design and analysis stages of the 173 studies are shown in Table 2. A total of 171 studies included age and/or sex as confounders. Two studies without age and sex involved women of childbearing age [22,23]. Comorbidities were applied as confounders in 146 studies. Of these, 123 studies used ad hoc selection method that selects comorbidities based on clinical judgement and 65 studies used comorbidity measurements such as Charlson comorbidity index (CCI), Elixhauser comorbidity measurement (ECM). In 21 studies, both methods were used. Sixty two studies used the CCI, two studies used the ECM, and one used both [24]. Health care utilization including medication, procedures, and history of admission were considered in 82 studies. Thirty six and 26 studies used type of health insurance and characteristics of healthcare providers, including the number of doctors, type and location of hospitals, and volume, respectively. Year of diagnosis and enrollment were employed in 12 studies.

Variables not recorded in the NHICD were frequently used as confounders. Household income level extracted from the Health Insurance Premium Database was the most used (77 studies). Fifty one studies used regions (urban and rural areas) extracted from the Korea Resident Registration Database. Lifestyle factors, body mass index, and clinical test results, which are core variables reflecting a patient’s health status and severity of the condition, were used in 44, 30, and 19 studies, respectively. These data were extracted from the National Health Screening Database. Family history, length of stay, and costs were classified as "Other."

Discussion
Adjusting for potential confounders is key to providing valuable evidence in health outcome studies using the NHICD. All 173 articles reviewed in this study adjusted for confounders at both study design and analysis stages. In particular, all applied the restriction method with exclusion criteria to control for confounders during the study design. Statistical analysis approaches, including regression analysis and the PS method, were frequently used to control for confounders; age, sex, and comorbidities were mainly applied as confounders.

Confounders were controlled more in the analysis than in the design stage, and regression analysis was used more often than the PS method. Regression analysis aims to investigate the relationship between dependent and independent variables, and is widely used in outcome studies. PS methods are also an increasingly common method for controlling confounders. The PS methods are reported to be less biased and more robust than regression analysis [25]. Amoah et al. suggested that PSMs are generally more favorable than regression analysis to estimate the causal effect, however, they do not compensate for poor study design or data accuracy [26]. Thirty four (19.7%) of the 173 studies used double adjustment. This can be used to eliminate confounding factors if an imbalance exists after PSM. Standardized differences of less than 0.1 can be practically considered as a sign of balance in baseline characteristics of study cohorts.

We often face a failure to achieve a balance of confounders between study cohorts after PSM. Nguyen et al. recommended reiterating PS modeling until standardized differences below 0.1 are achieved for most confounders, and the double adjustment might be worth considering if the imbalance remains [27].

Various variables, including age, sex, and comorbidities, were used as confounders. Researchers must understand the characteristics of the variables to apply them accurately. Otherwise, erroneous results may be derived than when potential confounders are not controlled. There are several implications to use the confounding variables. First, comorbidities
defined as the co-occurrence of more than one disease in the same patient and is associated with worse health outcomes [28], were the second most commonly used confounders. Some guidelines and comparative studies for collecting comorbidities when using the NHICD have been reported for specific conditions [29-31]. However, comprehensive guidelines are required to increase comparability among studies and accurately identify comorbidities.

Second, the type of health insurance and household income level were used as a proxy for socioeconomic status. In particular, the household income level was used in 77 (44.3%) studies because it reflects socioeconomic status more clearly than the type of health insurance. However, it has limitations that need to be addressed. Health insurance in Korea covers self-employed and employer-insured individuals. The insurance premium is levied based on the proportion of income, vehicles, and other assets for self-employed individuals, and the proportion of monthly income for employer-insured individuals. Therefore, comparing the relative rankings of insurance premiums without distinguishing between the two groups does not accurately reflect income levels. There are three methods to deal with these issues in the analysis step. First, stratified analysis can be used for self-employed and employer-insured individuals. Second, while insurance premiums are generally divided into 5 grades, a method that divides them into 10 grades, taking into account both self-employed and employer-insured individuals, can be employed. Lastly, if PSM are applied, we can calculate PS separately for self-employed and employer-insured individuals, then integrate and use the matched data as one dataset.

Third, the emergency status needs to be actively used when appropriate; however, no studies have used it. It is a meaningful variable that reflects a patient’s health status and severity of condition, although additional work is required to create it using fee codes for emergency care management.
Lastly, efforts to improve the quality of the NHICD are required. Diagnosis is the core information used to select the study population and identify comorbidities and health services. The problem with the accuracy of the diagnoses recorded in the NHICD is that the diagnoses are for reimbursement, not for clinical purposes. Present on admission (POA), a diagnosis-timing flag, is core information to differentiate between adverse events and comorbidity. A recent study reported that the agreement on POA between the NHICD and reference standard data was poor [32]. Interventions such as monitoring coding practices are required to improve the data quality including reliability and accuracy of the diagnoses and POA.

This study restricted the literature search to English-language publications in 2021 and 2022 and it may cause the exclusion of meaningful studies. However, this is the first review to address the methods and applications to adjust for confounders in studies using the NHICD. As advanced methodologies to adjust for confounders are being developed and applied, further studies using a comprehensive literature search strategy are required.

Adjustment for confounding factors is a critical component of outcome studies. Our results indicate that all studies adjusted for confounders with appropriate study designs and statistical methodologies; however, understanding and careful application of confounding variables are required to avoid erroneous results. The review also suggests the need for methodological guidance to improve the quality of outcome studies depending on how confounders, that can distort the observed causal relationship between exposure and outcome, are adjusted.
References


17. Park S, Han S, Suh HS. The disease burden of migraine patients receiving prophylactic


Table 1. Methods for adjusting confounding in the study design and analysis stage (n=173)

<table>
<thead>
<tr>
<th>Study design</th>
<th>Total</th>
<th>Regression with PSM</th>
<th>Univariate analysis with PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restriction</td>
<td>173(100%)</td>
<td>125(72.3%)</td>
<td>5(2.9%)</td>
</tr>
<tr>
<td>Matching</td>
<td>39(22.5%)</td>
<td>34(19.7%)</td>
<td>5(2.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2(1.2%)</td>
<td>2(1.2%)</td>
</tr>
</tbody>
</table>

Abbreviations: PS (Propensity score), PSM (Propensity score matching), IPTW (Inverse probability of treatment weighting)

* Of the 173 studies, two studies applied only matching method at the design stage.

** Other: matching and IPTW(n=1), Matching, IPTW and stratification(n=1)

Table 2. List of confounders for adjusting confounders in study design and analysis stage

<table>
<thead>
<tr>
<th>Confounders</th>
<th>Data source</th>
<th>n</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age and sex</td>
<td>NHICD</td>
<td>171</td>
<td>(98.3)</td>
</tr>
<tr>
<td>Comorbidity</td>
<td>NHICD</td>
<td>146</td>
<td>(83.9)</td>
</tr>
<tr>
<td>Ad hoc selection</td>
<td></td>
<td>123</td>
<td>(70.7)</td>
</tr>
<tr>
<td>Comorbidity measurement</td>
<td></td>
<td>65</td>
<td>(37.4)</td>
</tr>
<tr>
<td>Health service including medication, procedures, etc.</td>
<td>NHICD</td>
<td>83</td>
<td>(47.7)</td>
</tr>
<tr>
<td>Type of health insurance</td>
<td>NHICD</td>
<td>36</td>
<td>(20.7)</td>
</tr>
<tr>
<td>Characteristics of health care providers including type of institutions, volume, et al.</td>
<td>NHICD</td>
<td>26</td>
<td>(14.9)</td>
</tr>
<tr>
<td>Year at diagnosis or enrollment</td>
<td>NHICD</td>
<td>12</td>
<td>(6.9)</td>
</tr>
<tr>
<td>Household income level</td>
<td>HIPD</td>
<td>77</td>
<td>(44.3)</td>
</tr>
<tr>
<td>Resident</td>
<td>HIPD</td>
<td>51</td>
<td>(29.3)</td>
</tr>
<tr>
<td>Lifestyle factors including drinking, smoking, etc.</td>
<td>NHSD</td>
<td>44</td>
<td>(25.3)</td>
</tr>
<tr>
<td>Body mass index</td>
<td>NHSD</td>
<td>30</td>
<td>(17.2)</td>
</tr>
<tr>
<td>Results of clinical test including blood pressure</td>
<td>NHSD</td>
<td>19</td>
<td>(10.9)</td>
</tr>
<tr>
<td>Others</td>
<td>NHSD</td>
<td>32</td>
<td>(18.4)</td>
</tr>
</tbody>
</table>

Abbreviations: NHICD (National Health Insurance Claim Database), HIPD (Health Insurance Premium Database), NHSD (National Health Screening Database)
278 published studies identified via electronic database search.

3 studies without full text excluded.

275 studies

302 studies excluded:
- 81 epidemiology studies
- 5 international studies
- 3 economic evaluation studies
- 3 review studies
- 8 others

173 included studies