How authors select covariates in the multivariate analysis of cancer studies in 10 oncology journals in Korea: a descriptive study

Mi Ah Han¹, Hae Ran Kim², Sang Eun Yoon³, Sun Mi Park¹, Boyoung Kim⁴, Seo-Hee Kim¹,³, So-Yeong Kim¹

Departments of ¹Preventive Medicine and ²Nursing, Chosun University College of Medicine, Gwangju; ³Department of Public Health, Chosun University Graduate School, Gwangju; ⁴Chonnam National University College of Nursing, Gwangju, Korea

Abstract

Purpose: Cancer is the leading cause of death in Korea, leading many investigators to focus on cancer research. We present the current practice of variable selection methods for multivariate analyses in cancer studies recently published in major oncology journals in Korea.

Methods: We included observational studies investigating associations between exposures and outcomes using multivariate analysis from 10 major oncology journals published in 2021 in KoreaMed, a Korean electronic database. Two reviewers independently and in duplicate performed the reference screening and data extraction. For each study included in this review, we collected important aspects of the variable selection methods in multivariate models, including the study characteristics, analytic methods, and covariate selection methods. The descriptive statistics of the data are presented.

Results: In total, 107 studies were included. None used prespecified covariate selection methods, and half of the studies did not provide enough information to classify covariate selection methods. Among the studies reporting selection methods, almost all studies only used data-driven methods, despite having study questions related to causality. The most commonly used method for variable selection was significance in the univariate model, with the outcome as the dependent variable.

Conclusion: Half of the included studies did not provide sufficient information to assess the variable selection method, and most used a limited data-driven method. We believe that the reporting of covariate selection methods requires improvement, and our results can be used to educate researchers, editors, and reviewers to increase the transparency and adequacy of covariate selection for multivariable analyses in observational studies.

Keywords

Causality; Confounding factors; Epidemiologic; Multivariate analysis; Neoplasms
Introduction

Background
The long-term survival rates of cancer patients continue to increase because of early detection and advances in cancer treatment and care in Korea. However, approximately 243,000 cancers are diagnosed annually, and cancer has been the most common cause of death in Korea since 1983, when cause-of-death statistics began to be collected in Korea [1]. Due to the burden of cancer, researchers and health authorities have focused on epidemiological investigations, including the causes of cancer incidence, mortality, treatment effects, and health outcomes in patients with cancer.

If feasible, randomized controlled trials might provide the best answer for causal inference between an exposure or intervention and outcomes. However, for ethical and practical reasons, observational studies have provided important evidence regarding the epidemiology of cancer. In particular, the long induction period of several decades or more for cancer occurrence or death due to potential risk factors has underscored the crucial role of observational studies in the field of cancer epidemiology.

However, the interpretation of causal inferences in observational studies has a fatal flaw—namely, prognostic factors can frequently differ systematically between the exposure/intervention group and the control group. This systematic imbalance in prognostic factors is known as a confounder. To control for confounders, researchers have employed various methods, including multivariate analysis, stratification, and propensity score matching [2–4].

Selecting appropriate covariates for multivariate analysis is an important part of epidemiological studies for various reasons, including the justification and reproducibility of causal inferences. However, the choice of the best method for this selection remains a matter of debate, and determining which variables to include can be challenging. Although researchers and methodologists have recognized the importance of covariates in multivariate analysis, previous literature reviews have shown that variable selection methods differ, and most authors do not adequately report information on how covariates were selected.

A previous study on 488 articles with multivariate analyses published in five major medical journals, including the New England Journal of Medicine and The Lancet, found that 48% reported variable selection methods unclearly, 16% used data-driven methods, and 36% used knowledge-based methods. The authors found that 10.5% of the studies misused variable selection methods (defined as the use of a data-driven method in a study with causal questions) [5]. A substantial proportion of the studies in specific health journals did not report the justification or method of variable selection. In 193 orthopaedic studies with multivariate regression analyses, 65.8% selected variables based on nonstatistical methods (including all available variables without any interpretation of causality), and only 16% selected variables based on causal inference [6]. Of 150 nutritional studies, 94% did not select covariates a priori, and 63.3% did not report the selection criteria [7].

Objectives
The current practice of selecting covariates for multivariate models lacks well-documented methods for specific selection. Therefore, this study aimed to evaluate covariate selection methods in multivariate models by reviewing studies published in major oncology journals in Korea.

Methods

Ethics statement
Ethics approval was not required because we only used data from published papers. Our protocol has been registered on the Open Science Framework (OSF) Registry [8].

Journal selection
We included the following 10 oncology journals with the highest impact factors in the Korea Medical Citation Index (KoMCI): Asian Oncology Nursing, Brain Tumor Research and Treatment, Cancer Research and Treatment, Clinical Pediatric Hematology-Oncology, Immune Network, Journal of Breast Cancer, Journal of Cancer Prevention, Journal of Gastric Cancer, Journal of Gynecologic Oncology, and Radiation Oncology Journal.

KoMCI provides citation data for individual journals and subject categories in Korean medical journals [9]. We searched KoreaMed (https://koreamed.org/SearchBasic.php), a resource provided by the Korean Association of Medical Journal Editors (KAMJJE) that provides an assessment of articles published in Korean medical journals, for literature published in 2021. The search strategies used in the references were as follows: ("Asian Oncol Nurs"[JTI]) OR ("Brain Tumor Res Treat"[JTI]) OR ("Cancer Res Treat"[JTI]) OR ("Clin Pediatr Hematol Oncol"[JTI]) OR ("Immune Netw"[JTI]) OR ("J Breast Cancer"[JTI]) OR ("J Cancer Prev"[JTI]) OR ("J Gastric Cancer"[JTI]) OR ("J Gynecol Oncol"[JTI]) OR ("Radiat Oncol J"[JTI]) AND (2021[DPY]).

Inclusion criteria
We included observational studies that investigated the associations between exposures or interventions and health outcomes using multivariate analyses. We focused on studies involving human participants (e.g., patients, caregivers, health
volunteers, or healthcare practitioners). We included a broad definition of health outcomes, including those of direct importance to patients (e.g., mortality, morbidity, and quality of life) and surrogate outcomes (e.g., laboratory measures and radiological findings). The study design included cross-sectional, case-control, case-cohort, and cohort studies.

Since we focused on studies in which multivariate model outcomes were of primary interest, we selected studies that reported conducting multivariate analyses (e.g., multiple linear or logistic regression analysis) or presented regression coefficients (e.g., $\beta$ coefficient, adjusted odds ratio, or hazard ratio) in both the abstract and main results sections.

Randomized controlled trials, reviews, meta-analyses, pooled analyses, letters, commentaries, economic studies, and animal studies were excluded.

**Study selection process**

We reviewed all studies published in the journals mentioned above during the study period. For references identified from the electronic database search, title or abstract screening was performed. A study was considered potentially eligible if the title or abstract contained a description of the regression method (e.g., multiple linear or logistic regression analysis) or reported regression coefficients (e.g., $\beta$ coefficient, adjusted odds ratio, or hazard ratio). For potentially eligible studies, we obtained the full text and determined whether it satisfied the inclusion criteria. After calibration exercises, a team of two reviewers conducted the study selection process independently and in duplicate, and resolved any discrepancies by discussion or consultation with a third reviewer.

**Data extraction**

We conducted calibration exercises to ensure optimal accuracy and consistency and extracted data using a prepiloted data extraction form with written guidelines. Paired data extraction reports for each study were included independently and in duplicate. Any discrepancies were resolved by discussion or consultation with a third reviewer if necessary.

We examined the Methods section of the included studies to investigate whether the authors provided information supporting the causal inferences behind the multivariate model. We also scrutinized the Methods and Results sections to identify the method the authors used for covariate selection. In addition, we examined the study objective in the Introduction and Discussion sections to determine whether the authors performed a multivariate analysis with causal intent. Detailed information on data extraction is as follows.

**Study characteristics**

The study characteristics included journal name, first author, publication year, study design (cross-sectional, case-control, and cohort), number of participants, primary exposures or interventions investigated, primary outcomes investigated, type of cancer investigated, language of publication (English or Korean), study questions (causation or association), study protocol, and adherence to reporting guidelines.

We regarded a study as having a causal intent if it described the study questions using causal language (e.g., terms such as “impact,” “effect,” or “causal relationship”) in the Introduction section or study objectives. We also regarded the study as having causal intent if the exposures investigated were pharmacological, surgical, or behavioral treatments, or if they suggested adopting or avoiding exposure to improve the outcome of interest in the Discussion section. Studies with other types of questions, such as prognosis or prediction studies, were regarded as association studies.

**Analytic methods**

We collected data on analytic models (linear, logistic, Cox proportional hazards, and others). If a study reported more than one multivariate model, we regarded the results in the first table as the study’s primary results and reviewed them accordingly.

**Covariate selection methods**

We collected reports of covariate selection methods (not described vs. described), prespecification of covariate selection methods (not described vs. described), prespecification of covariates (not described vs. described), and covariate selection methods in the final analytical model (knowledge-based only, data-driven only, both, or not described).

**Knowledge-based methods**

We regarded covariates as selected by a knowledge-based method if a study stated that the covariates included in the multivariate model were previously known as potential confounders or if the authors hypothesized this. For studies using knowledge-based methods, we collected the following information: source of prior knowledge or hypothesis for covariates (published review, literature search conducted by the authors of the study, primary studies, expert opinion, not described, or others) and covariate selection methods (factors associated with the outcome of interest, factors associated with the exposure of interest, factors associated with both the exposure and outcome of interest, factors associated with either the exposure or outcome of interest, and others).

**Data-driven methods**

We considered a study to have used data-driven methods for covariate selection if it used the analyzed data. For studies us-
ing data-driven methods, we collected the detailed selection method as follows: covariate selection methods (effect estimate change; forward, backward, or stepwise selection; significant in the univariate model with the outcome as the dependent variable [e.g., \( P < 0.05 \)]; significant in the univariate model with exposure as the dependent variable [e.g., \( P < 0.05 \)]; significant in the multivariate model [e.g., \( P < 0.05 \)]; significant in the univariate model first and then significant in the multivariate model; significant in either the univariate or multivariate model; and others). This classification is not mutually exclusive because a single study may use multiple covariate selection methods.

**Bias**
There was no bias in searching and selecting the target literature.

**Study size**
It was not required to estimate the sample size. All target journals were included.

**Statistical analysis**
We presented the basic characteristics of the studies, including descriptive statistics, as numbers and percentages. Next, we reported the proportion of covariate selection methods and their associated characteristics. We compared covariate selection methods according to the study question (causation vs. association). All analyses were performed using the SAS ver. 9.4 (SAS Institute Inc).

**Results**

**Characteristics of the included studies**
A total of 494 articles published in 2021 were identified from the 10 oncology journals. Among them, 123 were included in title or abstract screening. After excluding 16 studies without multivariate analysis in the full-text screening, a total of 107 observational studies were included. About 90% were cohort studies, and about 74% included fewer than 1,000 participants in their study. The most common primary exposure was multiple risk factors, and the primary outcome was mortality. The most common types of cancer investigated were gastric (20.6%) and cervical (11.2%), and eight studies (7.5%) analyzed cancer in general (i.e., any type) (Table 1). The list of the included studies are provided in Suppl. 1.

**Variable selection methods**
Approximately 65% of the studies conducted Cox proportional hazard regression, and 26.2% conducted logistic regression. None used prespecified covariate selection methods and covariates. Only seven studies selected covariates using knowl-

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No. of studies (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of study design</td>
<td></td>
</tr>
<tr>
<td>Cohort</td>
<td>96 (89.7)</td>
</tr>
<tr>
<td>Cross-sectional</td>
<td>6 (5.6)</td>
</tr>
<tr>
<td>Case-control</td>
<td>5 (4.7)</td>
</tr>
<tr>
<td>Total number of participants included</td>
<td></td>
</tr>
<tr>
<td>( \leq 1,000 )</td>
<td>79 (73.8)</td>
</tr>
<tr>
<td>1,001–5,000</td>
<td>11 (10.3)</td>
</tr>
<tr>
<td>( \geq 5,000 )</td>
<td>17 (15.9)</td>
</tr>
<tr>
<td>Primary exposure</td>
<td></td>
</tr>
<tr>
<td>Multiple exposures [e.g., prognostic factors]</td>
<td>39 (36.5)</td>
</tr>
<tr>
<td>Therapeutic clinical intervention [e.g., behavior change facilitation, drug therapy]</td>
<td>29 (27.1)</td>
</tr>
<tr>
<td>Biophysical status [e.g., blood pressure, blood lipids, body weight]</td>
<td>17 (15.9)</td>
</tr>
<tr>
<td>Morbidity [e.g., cardiovascular disease, cancer, diabetes]</td>
<td>6 (5.6)</td>
</tr>
<tr>
<td>Health behavior [e.g., smoking, alcohol consumption, physical activity, diet]</td>
<td>4 (3.7)</td>
</tr>
<tr>
<td>Other</td>
<td>12 (11.2)</td>
</tr>
<tr>
<td>Primary outcome</td>
<td></td>
</tr>
<tr>
<td>Mortality [e.g., all-cause mortality, disease-specific mortality]</td>
<td>64 (59.8)</td>
</tr>
<tr>
<td>Morbidity [e.g., cardiovascular disease, cancer, hospitalization]</td>
<td>22 (20.6)</td>
</tr>
<tr>
<td>Quality of life [e.g., overall, disease-specific quality of life]</td>
<td>5 (4.7)</td>
</tr>
<tr>
<td>Biophysical status [e.g., blood pressure, blood lipids, body weight]</td>
<td>5 (4.7)</td>
</tr>
<tr>
<td>Other</td>
<td>11 (10.3)</td>
</tr>
<tr>
<td>Type of cancer investigated</td>
<td></td>
</tr>
<tr>
<td>Any</td>
<td>8 (7.5)</td>
</tr>
<tr>
<td>Gastric</td>
<td>22 (20.6)</td>
</tr>
<tr>
<td>Cervix</td>
<td>12 (11.2)</td>
</tr>
<tr>
<td>Breast</td>
<td>10 (9.4)</td>
</tr>
<tr>
<td>Uterus</td>
<td>8 (7.5)</td>
</tr>
<tr>
<td>Lung</td>
<td>6 (5.6)</td>
</tr>
<tr>
<td>Colorectum</td>
<td>6 (5.6)</td>
</tr>
<tr>
<td>Brain</td>
<td>5 (4.7)</td>
</tr>
<tr>
<td>Ovary</td>
<td>4 (3.7)</td>
</tr>
<tr>
<td>Leukemia/lymphoma</td>
<td>4 (3.7)</td>
</tr>
<tr>
<td>Other</td>
<td>22 (20.6)</td>
</tr>
<tr>
<td>Language of publication</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>100 (93.5)</td>
</tr>
<tr>
<td>Korean</td>
<td>7 (6.5)</td>
</tr>
</tbody>
</table>

(Continued on the next page)
edge-based methods, and 57 selected covariates using data-driven methods. The most common data-driven method was significance in the univariate model with the outcome as the dependent variable (e.g., P < 0.05) (Table 2 and Dataset 1).

According to the study question, Cox proportional hazard regression was most frequently used for both types of study questions. Studies with associational questions more frequently described covariate selection than studies with causal questions. Even in studies with causal questions, approximately half did not clearly report the variable selection method (Table 3).

**Discussion**

**Key results**

Multivariate analysis is a commonly used method of control-
ling for the effects of confounders in observational studies to determine the relationship between exposures and outcomes. This systematic review presented the current variable selection and reporting practices in leading oncology journals in Korea, and it is the first study to present this information. None of the included studies used prespecified covariate selection methods, and half of the studies did not provide sufficient information to allow a classification of their methods.

Comparison with previous studies
Half of the included studies did not provide sufficient information to evaluate the covariate selection methods. This finding is similar to the results of previous studies. A descriptive review of variable selection methods in four major epidemiology journals reported that 37% of the included studies did not provide sufficient details to allow the determination of variable selection methods [10]. Similarly, the variable selection method was unclear in 48% of the studies in the five medical journals [5]. This might be due to well-known associations between variables that do not require citations, lack of space in the paper, the authors’ ignorance, or the journal’s editorial policy [5]. However, in the absence of a widely agreed-upon method for variable selection, a complete and clear description of the statistical methods, including variable selection, would provide evidence for judgment about the uncertainties associated with the interpretation of the results. Therefore, authors and journals should focus on transparent and clear descriptions of variable selection.

Of the studies reporting variable selection, most selected covariates using only data-driven methods, even when the studies had causal study questions. Previous studies have employed various data-driven methods. Among 292 studies in four epidemiology journals, 146 selected variables based on prior knowledge and 69 selected variables using data-driven methods, and the change in estimate approach was the most common [10]. Among 287 studies in two Chinese epidemiology journals, 163 selected variables using bivariate analyses and 45 selected variables based on prior knowledge or personal judgment [11]. Among 488 articles, variable selection was knowledge-based in 176 and data-driven in 78, and univariate selection was the most common [5]. Data-driven methods are known to be suitable for association questions such as prognosis or predictive research, while data-driven methods for causal inference studies are known to pose a risk of bias [4]. Although various data-driven methods have been introduced, only a few were used in the included studies.

Implications for future studies
None of the included studies used prespecified covariate selection methods and covariates. It is best to specify this in the research protocol or statistical analysis plan; however, a previous study addressing the registration practices of observational studies noted that the preparation of protocols for observational studies was very limited [12]. Observational study protocols are necessary for the qualitative aspects of studies, such as covariate selection and prevention of selective reporting; however, the research community has not been able to reach a consensus regarding this issue, and only a few institutions or reporting guidelines suggest it [13].

Guidelines for observational studies, such as STROBE (Strengthening the Reporting of Observational Studies in Epidemiology), indicate that it is necessary to “describe all statistical methods, including those used to control for confounding,” but only one included study directly stated that they followed the reporting guidelines. Since compliance with reporting guidelines is greatly influenced by journal editorial policy, the peer review process or journal guidance for authors should reflect this for the completeness of the quality of research reporting.

Strengths
The strengths of this study include its adherence to a standard methodology. We conducted independent and duplicate reference screening and data extraction after the calibration exercises. Additionally, our study selected representative and major oncology journals and provided a comprehensive picture of the current practice of covariate selection methods.

Limitations
A potential limitation of this study is our reliance on journal reports to evaluate the choice of covariates. The authors may not have accounted for covariates because of their relatively low importance or the journal’s word count limits. Nevertheless, we reviewed all the authors’ descriptions, including the protocol and appendices, if available.

The process of extracting information from the studies may have required subjective judgment by a reviewer. For example, even if the study did not explicitly state that causation was of interest, if it recommended changes in exposure to improve the health outcome of interest, we regarded the study question as causation, depending on the reviewer’s interpretation. We conducted training and calibration exercises for the reviewers with documented instructions to ensure a high degree of agreement.

Conclusions
We provided an overview of the covariate selection methods used in articles published in major Korean oncology journals. None of the included studies used prespecified covariate selection methods, and half of the studies did not provide enough

https://www.escienceediting.org
information to classify the methods. As there is currently no single agreed-upon method, clearly and completely describing the methods is important for the interpretation of results and judgment of uncertainty. Our results inform the research community about controlling for confounders. In addition, they can be used to educate researchers, editors, and reviewers to increase the transparency and adequacy of covariate selection in multivariate analyses in observational studies.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Funding

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Korean Ministry of Education (No. 2021 R1I1A3041301) and the Korean Ministry of Science and ICT (No. NRF-2022R1A5A2030454).

Data Availability

Dataset file is available from the Harvard Dataverse at https://doi.org/10.7910/DVN/DC2F1B.

Dataset 1. Research data of covariates selection methods.

Supplementary Materials

Supplementary materials are available from https://doi.org/10.7910/DVN/DC2F1B.

Suppl. 1. List of the included studies.

References

8. Han MA, Jung JJ, Hwang EC. Variable selection methods in major oncology journals in Korea: a protocol of a systematic survey [Internet]. Open Science Framework (OSF); 2022 [cited 2024 Feb 14]. Available from: https://doi.org/10.17605/OSF.IO/K584B