Clinical Pediatric Hematology-Oncology Volume 29 · Number 1 · April 2022 REVIEW ARTICLE

Pediatric Cancer Research using Healthcare Big Data

Hyery Kim

Department of Pediatrics, University of Ulsan College of Medicine, Asan Medical Center Children's Hospital, Seoul, Korea

Health insurance big data provides real-world evidence of unmet needs in clinical practice and breakthroughs in the medical industry that will impact the future of health care. Big data is expected to revolutionize the current medical paradigm and usher in an era of personalized medicine. In Korea, the Health Insurance Review and Assessment Service and the National Health Insurance Service established large-capacity healthcare big data open systems in 2011 and 2013, respectively, and are providing researchers with secured healthcare big data. However, concerns have been raised regarding the quality of big data-based research. Thus, numerous obstacles remain in leveraging big data research to change medical practice. This paper describes the understanding and practical applications of healthcare big data in pediatric cancer research, ranging from clinical research design using health insurance big data to medical writing.

Key Words: Healthcare, Big data, Pediatric cancer, Research

pISSN 2233-5250 / eISSN 2233-4580
https://doi.org/10.15264/cpho.2022.29.1.1
Clin Pediatr Hematol Oncol 2022;29:1~11

Received on March 21, 2022
Revised on April 6, 2022
Accepted on April 13, 2022

Corresponding Author: Hyery Kim
Department of Pediatrics, University of Ulsan College of Medicine, Asan Medical Center Children's Hospital, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
Tel: +82-2-3010-3373
Fax: +82-2-473-3725
E-mail: taban@hanmail.net
ORCID ID: orcid.org/0000-0003-2852-6832

Introduction

Big data is defined as data of a scale exceeding the scope of existing database (DB) management tools for collecting, storing, managing, and analyzing data. Healthcare big data has recently been established in various fields [1]. Big data is mainly divided into electronic medical records, claims, genomic data, and patient-generated health data [2]. Health care data is essential for improving the quality of treatment and reducing medical costs. It encompasses various data types and sources, including all health-related genotypes, family and friend relationships, biological phenotypes, environmental exposures, behaviors, and lifestyles-based on medical data [3].

Big data is divided into primary data (acquired for research) and secondary data (collected for claims) [4]. The use of secondary data enables thorough analyses; however, it carries a considerable risk of systemic and random errors. In Korea, the Health Insurance Review and Assessment Service (HIRA) and National Health Insurance Service (NHIS) established large-capacity health and healthcare big data open systems in 2011 and 2013, respectively, and are providing researchers with various types of secured healthcare big data [4]. The use procedure supports visits or remote access to enable customized data analysis desired by industry and researchers. Recently, a sample cohort DB based on claims data was built and supported, and it is being used in numerous clinical studies. This paper details the step-by-step knowledge required to design a clinical trial, write a thesis, and apply claims data to clinical practice using

healthcare big data.

Current State of Public Big Data in Health and Medicine in Korea

1) Claims data from the NHIS

The representative public big data in the health and medical fields used by Korean researchers are the NHIS and HIRA DB data (Table 1) [5]. The NHIS and HIRA DBs contain the same information. However, there is a slight difference in the research data services provided by these institutions.

Data from 2002 on qualifications and insurance from birth to death, hospital usages, national health exam results, rare incurable cancer registration information, medical benefits, and elderly long-term care data are available from the NHIS DBs (NHIS data sharing service: https://nhiss.nhis.or.kr/bd/ay/bdaya001iv.do). The DBs also include treatment, medical checkups, and medical care history.

Research data is primarily provided as "customized research" and "sample cohort" DBs. The customized research DB refers to data that has been processed and delivered as customized data so that it can be used for policy decisions and academic research. They can be reviewed and analyzed using SAS or R in the data analysis room within the NHIS or other designated local centers. The sample, medical checkup, geriatric, children's screening, and working women cohort DBs are all included in the sample cohort DB, distinguishing the HIRA's data provisioning service. The sample cohort DB 2.0, for example, contains samples from approximately one million people, covers the entire country as of 2006, and gives data from 2002 to 2015, allowing for longitudinal research. The customized data provides only the date of death, however the sample cohort DB offers the date and cause of death.

Because the customized research DB is a representative DB consisting of the medical usage behaviors, diagnostic codes, prescription codes, and drug codes of Koreans, all types of research, including rare diseases, are possible. The control group can be derived based on the control group selection conditions if needed. There are considerable challenges related to access, cost, space, and data analysis for a customized research database. SPSS is difficult to use for analysis, so R statistics or SAS must be used instead.

2) Claims data from the HIRA

Since 2007, the customized dataset has been available, including information on general specifications, medical treatment, disease, outpatient prescriptions, and the state of health care institutions (open data system of HIRA: http://opendata.hira.or.kr/home.do) [6]. Although it is different from the cohort DB of the NHIS, the HIRA also provides sample data, as well as inpatient, total patient, elderly patient, and pediatric patient datasets. In the case of the pediatric patient dataset, the extraction rate of pediatric patients (under 20) after 2009 is 10%; therefore, the data of approximately one million children are provided. Since it is not longitudinal data, it is more suitable for cross-sectional studies such as prevalence. The difference from the NHIS is that HIRA provides additional information related to drugs. It contains data on drug utilization reviews (DURs) and medication distribution management, information on treatment adequacy evalu-

Table 1. Types of health insurance claims data in Korea

Customized research DB	Sample cohort DB
It is extracted, summarized, and processed according to the research goal to be utilized for policy and academic study without revealing the subject's identity. • Very high volume of data • Only the "data analysis room" at the NHIS or the designated centers is used for analysis	It consists of high-demand data sets for research. • De-identified data • Standardized dataset provides for policy and academic research • DB consists of sample, medical checkup, geriatric, children's screening, and working women cohort DBs • Remote research is possible

DB, database; NHIS, National Health Insurance Service.

ation, and some non-reimbursable drugs. The researcher is given remote access rights for some customized datasets, which may then be examined via the HIRA server. Unlike NHIS, HIRA does not offer a date of death.

3) Data limitations

The disadvantages include difficulty obtaining an accurate clinical diagnosis and excluding non-reimbursable prescription medications or treatments. Due to the fact that this information is used to process insurance claims, doctors may enter disease codes that are more severe than the actual condition in order to avoid a reduction in non-reimbursement. In contrast, even for severe diseases, it may be difficult to accurately identify the code if it is not related to reimbursement. In the case of the NHIS data, analysis is only feasible by visiting the main center of NHIS directly or restricted rooms in the regional analysis centers. Additionally, individual patient data cannot be extracted from the analysis center. After processing raw data in a summarized format, it can only be extracted from the original datasets, such as in the form of a table. Since data is stored as of the invoice issuance date, there may be a difference in time from when care was actually administered.

4) Other materials

Another health and medical statistics source is the Korean National Institute of Health's Korean Genome Analysis Project; this data is subject to limited disclosure. Additionally, the Korea Disease Control and Prevention Agency collects the Korea National Health & Nutrition Examination Survey, which is used to determine the current state and trends in people's health and nutritional status, as well as the Community Health Survey, which is used to develop a community health care plan and evaluate health projects. If such healthcare big data are used in various ways, more diverse research can be conducted. However, since each sample datum is not linked with the National Cancer Center and Statistics Korea data, it has limitations as data from a single institution. Cancer registration data containing personal information cannot be used except in exceptional circumstances.

Clinical Research using Healthcare Big Data

Prior to starting research using healthcare big data, it is critical to understand the characteristics of the data and the types of research that could be conducted using it. The case-control and cohort studies are the most frequently used research designs for health insurance data [4]. The two studies differ in where the clinical outcome occurs. A case-control study is designed to ascertain prior exposure to risk factors after grouping participants into disease-prone and non-occurring groups at the time following the clinical outcome. A cohort study evaluates the risk of future disease occurrence between the two groups by dividing them into intervention and non-intervention groups. A relatively simple cross-sectional study among study design types is a method to simultaneously investigate the onset of the targeted disease and risk factors by extracting a sample.

First, select the customized database or a sample cohort for each study design. For diseases with a relatively high prevalence, researchers are recommended to use a sample cohort because most of the research objectives can only be achieved with sampled cohort data. It is advantageous to utilize customized data targeting the entire population for diseases with a low prevalence or incidence.

Researchers unfamiliar with the data have difficulty comprehending the classification and qualities, making it difficult to initiate the investigation. It is necessary to understand the characteristics of the claim code, and until now, analysis was possible only through SAS or R. To understand and analyze the data, a collaborator who can interpret it is essential.

In pediatric cancer research, public healthcare big data includes data from the entire population, so it can solve unmet research needs that are difficult to conclude with limited data. Since it is possible to check the medical records of each individual, it is an excellent data source for researching pediatric cancer survivors. When combined with national cancer registration data, it may

help identify pediatric cancer patients and clarify their diagnostic name and date of diagnosis, allowing for world-class epidemiologic research of pediatric cancer patients.

Since such healthcare big data is based on claims for actual treatment, reimbursement-related factors affect data accuracy. As a result, the element of the study that requires the most effort is diagnostic accuracy. Because most research employs ICD-10 codes to define a specific disease, the number of patients with a given condition could frequently be higher or lower than predicted. Thus, extra attempts to support it, such as the use of medications and surgery, might be evaluated concurrently to improve diagnostic accuracy. It may not contain information about the use of essential drugs or procedures.

Health insurance data has the problem that, unlike medical records, it lacks comprehensive clinical information, making it impossible to distinguish between time-based medical practices that can explain a causal relationship. For example, suppose a pediatric cancer patient has surgery because of a perforation caused by a colonoscopy. It is hard to tell whether the operation is for cancer therapy or colonoscopy perforation because only the codes for colonoscopy and surgery can be checked. It is impossible to determine the outcome of a health checkup performed at patients' own expense, which does not include non-covered therapy such as new surgeries or medications. While gender and age can be validated in the claims data, additional physical characteristics such as height, blood pressure, and socioeconomic factors such as drinking history, smoking history, and activity level are not included in the claims data. Therefore, it is impossible to account for multiple risk variables in the data analysis, contributing to the research's lack of precision and trustworthiness.

It is necessary to select an appropriate research topic after thoroughly examining whether data suitable for the research can be extracted with expert advice. Because it is impossible to define operational diagnosis using only a few claim items thoroughly, it is required to confirm the operational diagnosis of a targeted disease with prior research. It is preferable to first validate the accuracy of the operational diagnosis by comparing it to the institution's health records or a cohort built by researchers' organizations.

Research using Korean Public Claims Data

Research using public claims data is performed as a retrospective study. A retrospective database study can generally reflect routine care compared to randomized clinical trials and long-term follow-up of large-scale patient data to determine the clinical effect. It is useful when it is necessary to derive timely research results because research can be carried out in a relatively short time and with minimal cost.

By combining pre- and post-illness data, healthcare big data can be used to generate a disease cohort. An important advantage of such healthcare big data is that no patients drop out halfway through treatment, owing to the nature of Korea's medical system, which is population-based. Because healthcare big data was initially designed to charge for medical treatment and treatment expenses, it enables numerous cost analyses.

1) Establishment of research hypotheses

When establishing a research hypothesis, it is necessary to consider whether it can be elucidated with this data. In studies involving child and adolescent cancer patients, non-reimbursable data such as immunotherapy or targeted anticancer drugs are not appropriate. Clinical results other than death are not confirmed, which needs to be taken into account. It is unknown whether over-the-counter medications are used. It is necessary to consider the limitations owing to insufficient information about the cancer stage and accurate histotype at the time of diagnosis. Since cancer diagnosis claims continue even after treatment is finished, the claims data does not precisely define recurrence or the occurrence of secondary cancer, which is a drawback.

2) Operational definitions for selecting study subject

Research subjects must be chosen carefully to obtain the best results using the claims data and operational

Table 2. Methods for establishing operational definitions in cancer research

The English abbreviation is the name of a searchable variable of the health insurance data
Person with a 'C' code in the primary diagnosis (SICK_SYM1)
Person with a 'C' code in the primary diagnosis (SICK_SYM1) or the secondary diagnosis (SICK_SYM2)
Person with a 'C' code in the entire diagnosis (SICK_SYM1–5)
Person with a 'C' code in the primary diagnosis (SICK_SYM1) among inpatients (FORM_CD)
Person with a 'C' code in the primary diagnosis (SICK_SYM1) or the secondary diagnosis (SICK_SYM2) among inpatients (FORM_CD)
Person with a 'C' code in the entire diagnosis (SICK_SYM1–5) among inpatients (FORM_CD)
Other procedure or treatment codes, and medical expenses of one million won or more

Table 3. Variables used in operational definitions in pediatric cancer research

Types of variables	Actual codes	Code descriptions
KCD-10 codes	C00–C97, D00–D09, D37–D48	Diagnosis for cancer patients
Specific symbolized type (SPCF_SYM_TYPE)	V011	The day of outpatient treatment for pediatric cancer patients under age 18
	V026	The day of outpatient treatment for leukemia patients (C90–C95)
	V027	The day the cancer patient received treatment (C00–C97, D00–D09, D37–D48) (including children under age 18 with cancer and leukemia patients)
	V193	When a cancer-registered patient has received cancer treatment (C00–C97, D00–D09, D32–D33, D37–D48) for five years after registration
	V194	When a cancer-registered patient has received home care for five years after registration (C00–C97, D00–D09, D32–D33, D37–D48)

definitions using codes must be determined.

Recent analysis of operational definitions in cancer research (Table 2) has shown that both the prevalence and incidence rates showed the most similar results with the actual rates in the case of operational definition as "A person with a 'C' code in the primary diagnosis (SICK_SYM1) among inpatients (FORM_CD)" [7]. In the case of breast, prostate, and cervical cancer, which have a limited incidence according to gender, these definitions are generally sufficient to estimate the real numbers. However, in other representative cancers, the overall prevalence tends to be overestimated.

The operational definition of childhood and adolescent cancer can be defined as a method of cross-searching the specific symbolized type, which is a mark that provides selective insurance benefits for cancer patients, together with the KCD-10 code corresponding to cancer (Table 3). However, because entering the diagnostic codes alone does not allow for the selection of patients who have actually received chemotherapy, a method of selecting a subject by including the entire drug codes of chemotherapeutic agents or treatment codes used when administering an anticancer drug can be considered (Table 4).

Table 4. Codes for chemotherapy injection

Codes for injection of chemotherapy	
Anticancer drugs for injection	J0041, KK059, KK151–KK156, KK158, KK159, AP502

3) Analysis

Disease, drug, treatment, material, and other codes should be used to define treatment, outcome variables, and confounding variables. In the case of a study that includes drug or treatment codes, changes in an insurance policy must be considered because changes in the new drug or treatment codes appear following changes in the insurance policy. Claim codes for hematopoietic stem cell transplantation, for example, vary by age, donor, and billing period, and some codes are no longer valid (Table 5). As a result, if the researchers want to add patients from previous periods, they should include outdated codes. To select patients who have received chemotherapy, it may be more effective to use the treatment

Table 5. Treatment codes related to hematopoietic stem cell transplantation

Category	Subcategory	Code details	Treatment codes	Status
Allogeneic	Bone marrow	Allogeneic BM stem cell transplantation	X5011	Discontinued
		Allogeneic BM stem cell transplantation (children <6 years old)	X5012	Discontinued
		Allogeneic BM stem cell transplantation	X5013	Discontinued
		Allogeneic BM stem cell transplantation (children <6 years old)	X5014	Discontinued
		Stem cell infusion (allogeneic BM)	X5131	Active
		Stem cell transplantation/infusion (allogeneic BM) (<1 year old)	X5131100	Active
		Stem cell transplantation/infusion (allogeneic BM) (1 year≤age<6 years old)	X5131600	Active
	Peripheral blood	Allogeneic PB stem cell transplantation	X5041	Discontinued
		Allogeneic PB stem cell transplantation (children)	X5042	Discontinued
		Stem cell infusion (allogeneic PB)	X5133	Active
		Stem cell transplantation/infusion (allogeneic PB) (<1 year old)	X5133100	Active
		Stem cell transplantation/infusion (allogeneic BM) (1 year≤age<6 years old)	X5133600	Active
Autologous	Bone marrow	Autologous BM stem cell transplantation	X5021	Discontinued
		Autologous BM stem cell transplantation (children)	X5022	Discontinued
		Stem cell infusion (autologous BM)	X5132	Active
		Stem cell transplantation/infusion (autologous BM) (<1 year old)	X5132100	Active
		Stem cell transplantation/infusion (autologous BM) (1 year≤age<6 years old)	X5132600	Active
		Stem cell transplantation/infusion (autologous PB) (<1 year old)	X5134100	Active
		Stem cell transplantation/infusion (autologous PB) (1 year≤age<6 years old)	X5134600	Active
	Peripheral blood	Autologous PB stem cell transplantation	X5023	Discontinued
		Autologous PB stem cell transplantation (children)	X5024	Discontinued
		Stem cell infusion (autologous-PB)	X5134	Active
Cord blood	Cord blood	Cord blood stem cell infusion	X5032	Discontinued
		Cord blood stem cell infusion (children)	X5033	Discontinued
	Allogeneic	Stem cell infusion (allogeneic CB)	X5135	Active
		Stem cell transplantation/infusion (allogeneic CB) (<1 year old)	X5135100	Active
		Stem cell transplantation/infusion (allogeneic CB) (1 year≤age<6 years old)	X5135600	Active
	Autologous	Stem cell infusion (autologous CB)	X5136	Active
		Stem cell transplantation/infusion (autologous CB) (<1 year old)	X5136100	Active
		Stem cell transplantation/infusion (autologous CB) (1 year≤age<6 years old)	X5136600	Active

BM, bone marrow; PB, peripheral blood; CB, cord blood.

code required for anticancer drug administration rather than the diagnostic codes (Table 4).

The quality of research can be improved if various data sources are combined with the claims data [3]. It is impossible to ascertain the exact date of diagnosis, cancer stage, and pathology results while researching cancer patients using claims data. If this information is required for the study, the National Cancer Center's cancer registration data can be used by requesting a service that integrates it with insurance claims data. It is possible to specify only actual patients with confirmed cancer by complementing the limitation of the diagnostic codes, which cannot accurately specify a patient even with an operational definition. However, only national cancer registration data from 2011 can be used for combination data, and only data up to 10 years old can be requested.

If the outcome variable is in-hospital death, it can be defined using a medical outcome variable or a diagnostic code. However, the definition is unclear when death occurs outside of a medical institution and cannot be identified by claims data. If the current death data from the Statistics Korea and customized research data are fused, the exact date of death and cause of death listed in the death certificate can be utilized.

Methodology for Research using Healthcare Big Data

A confounding variable that distorts the relationship between treatment and outcome variables can be defined as a confounding factor related to treatment and affects outcomes [8-10]. It is necessary to define confounding variables related to treatment and apply a proper analysis method that controls the defined confounders [11]. The usual methods of controlling the confounding variables in study design are restriction and matching [5,9,12]. The restriction approach limits the study subjects by using appropriate inclusion/exclusion criteria to ensure the homogeneity of subjects included in the study. The matching process involves selecting a comparison group so that the distribution of the confounding variable is the same or similar to the reference group so that the two groups to be compared have identical characteristics.

The methods of controlling the confounding variables in terms of analysis are stratification, multiple regression models, and propensity score matching [9,10]. The limiting, matching, and stratification methods are appropriate when the number of confounding variables is small. If the number of confounding variables is large, it is more suited to apply the matching or stratification method by summarizing the confounding variables with propensity scores [10].

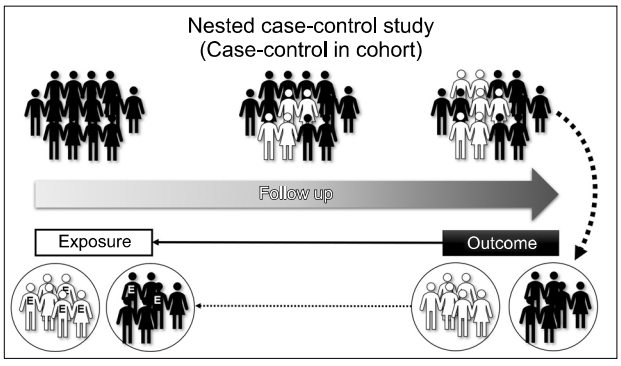

Fig. 1. Nested case-control study: A nested case-control study is conducted by extracting case-control data according to the disease state from the cohort data. E, exposure.

1) Nested case-control study

Case-control studies are frequently used as the initial step in research to evaluate whether exposure factors enhance disease risk [10,13]. They are less expensive than cohort studies and can be completed more quickly. It is also useful when the disease to be investigated is rare [10].

A nested case-control study is conducted by extracting case-control data according to the disease state from the cohort data (Fig. 1). For example, suppose that 1,000 out of 100,000 people developed acute lymphoblastic leukemia after ten years of follow-ups in a cohort study. The differences in gene expression can be seen by properly extracting cases and controls, which is an example of a nested case-control study. If you build a gene expression data set in the early stage of the cohort, you need to extract the blood of 100,000 people and make a gene expression DB. If you use the nested case-control study method, you can save time and money by extracting controls 1-4 times larger than the case and analyzing gene expression.

2) Retrospective cohort study

Research using healthcare big data is a representative observational study, and since it is already established data, it is a retrospective cohort study [10]. In other words, after observation of the subject is completed, the results of observation are constructed as data, and then the research begins with the data (Fig. 2). It was the most frequent design among studies using the National Health

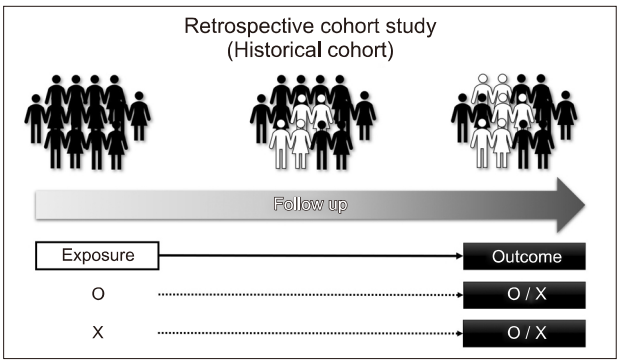

Fig. 2. Retrospective cohort study.

Information DB in Korea, accounting for 56.7% of the total reports [10]. The study time frame is after, not before, cohort recruitment, similar to the nested case-control research. A retrospective cohort study is essentially a cohort study, except that the research group is not chosen based on disease status, as in a case-control study.

Suppose you spent ten years following miners working in a mine. A retrospective cohort study design can be used if we wish to compare these miners' lung cancer death rate to that of the general population. With this cohort of miners, nested case-control research is one in which a case-control is extracted based on lung cancer, and then exposure is compared to this case-control.

3) Propensity score matching

Propensity score matching (PSM) is a method of selecting a comparison group with the same distribution of the confounding factor to the standard group (Fig. 3) [10,14]. The two groups to be compared have similar characteristics. Confounding factors are usually selected as matching variables. Among all studies using the National Health Information DB in Korea, 18.58% of the studies used the matching method, and among them, the PSM study was the most commonly used in 68.25% [10].

The PSM method is a non-parametric method that creates similar conditions in observational studies where randomized trials are impossible [14]. PSM is a method for minimizing selection bias in control selection by controlling the influence of confounding variables through multiple covariates. The covariate should not be a parameter, and the propensity scores between the two groups should have many overlapping parts. When there are many unexposed subjects in a cohort format, PSM matches unexposed patients approximately 1-10 times of exposed patients having similar demographic and socioeconomic characteristics, medical evaluations, and comorbidities of exposed participants. Measuring the incidence and mortality during the follow-up period is similar to the cohort study.

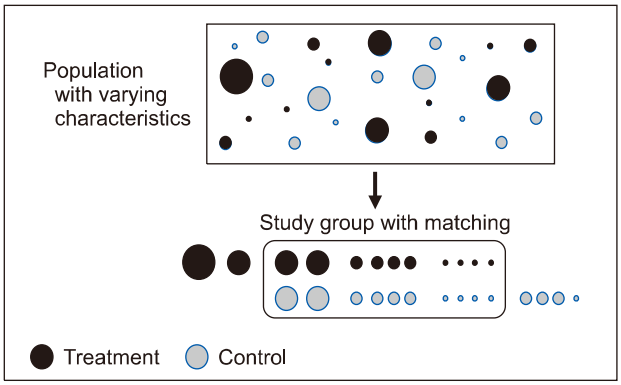

Fig. 3. Propensity score matching method. Propensity score matching is a method of selecting a comparison group with the same distribution of the confounding factor to the standard group.

Writing a Paper using Healthcare Big Data

In research studies using big data, it is necessary to set a topic that is difficult to draw conclusions from existing small-scale studies and sufficiently describe the justification for using big data [15]. The operational definition for big data analysis is the most critical component of big data research. If there is potential for error or confusion, the analysis results will be difficult to trust. Therefore, it is critical to validate the trustworthiness of the big data observational study by comparing the identification code or algorithm used to match the target patient to actual hospital data. [16,17]. If there are studies that validate the criteria for patient selection in previously published articles, these can be substituted for references. The method for determining risk factor exposure, clinical outcome, operation or surgery, or administration of drugs employed in the study should be detailed. Patient data initially extracted according to the selection and exclusion criteria should be presented in a flow chart for easy understanding [18].

Secondary Health DB Research Bias

Owing to the retrospective cohort nature of the secondary healthcare DB, there is a possibility of numerous biases emerging when performing pharmacoepidemiological investigations (Table 6) [19]. Bias is primarily classified into confounding, selection, measurement, and time-related biases. Confounding factors impair the ac-

Table 6. Bias treatment codes related to hematopoietic stem cell transplantation

Category Subcategory	Description of the bias
Confounding	The measure of association between treatment and outcome is distorted by the effect of variables, which are risk factors for the targeted outcome.
Confounding by indication	The clinical condition that determined the prescription of the treatment is associated with the effect, acting as a confounding factor.
Time-dependent confounding	A time-dependent variable acts as a confounding factor between the current exposure and outcome, and as an intermediary between prior and current exposure.
Unmeasured/residual confounding	There is not enough information about all the relevant confounding factors known, unknown or difficult to measure. If confounding cannot be controlled, the residual confounding effect of some factors remains in the effect that is observed.
Healthy user/adherer effect	Access to health care resources is associated with a higher education and health-seeking behavior. Patients who comply with prolonged treatment periods tend to be healthier.
Selection bias	The sample population is not representative of the population to which the results will be extrapolated.
Protopathic bias	The treatment is associated with subclinical disease stages (an early manifestation of the still undiagnosed condition under study gives rise to prescription of the treatment).
Losses to follow-up (informative censoring)	The mechanism that triggers discontinuity of treatment is associated with the risk of observing the outcome of interest.
Depletion of susceptibles (prevalent user bias)	The inclusion of prevalent instead of incident users entails insufficient verification of the adverse effects that occur at the beginning of treatment (those susceptible to the effect have interrupted the treatment).
Missing data	In multivariate analyses, observations that lack some values of a variable included in the model tend to be eliminated.
Measurement bias	Data on true exposures, outcomes, and other variables are recorded in the form of indicators that do not accurately reflect reality.
Misclassification bias	The association between treatment and outcome is distorted by errors, owing to the way variables of interest are measured in comparison groups.
Misclassification of exposure	The measure of exposure of a given treatment is not an exact reflection of its real use.
Misclassification of outcome	There is an error in the diagnosis.
Time-related bias	Follow-up time and exposure status are inadequately considered in the study-design or analysis stages.
Immortal time bias	A period during which the study event cannot occur is included in the follow-up or is excluded from analysis due to an incorrect definition at the start of follow-up.
Immeasurable time bias	A period during which follow-up is ignored and thus misclassified as an unexposed period, since outpatient prescriptions that define exposure cannot occur.
Time-window bias	Using time-windows of different lengths between cases and controls to define time-dependent exposures prevents subjects from having the same opportunity time to receive prescriptions.
Time-lag bias	Treatment comparisons are conducted at different stages of the disease, which introduces bias related to disease duration and progression.

curacy of experimental design and outcomes when they are not controlled. Therefore, it is critical to identify confounding variables and appropriately correct for bias.

Among the time-related biases, the immortal time bias, a representative bias, is a bias to be aware of in cohort studies. Immortal time means the follow-up time

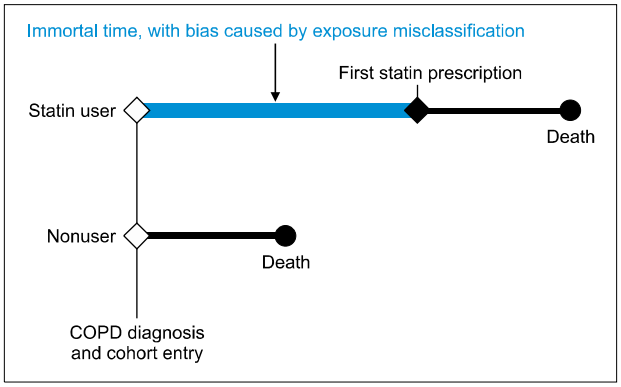

Fig. 4. Immortal time bias in statin study with COPD patients. For statin users, the time between the diagnosis of COPD and the first statin prescription is the immortal time. Therefore, if a patient's death occurs before receiving a statin prescription, the case becomes part of a non-user group. There is a bias in that the mortality rate is higher in the non-user group.

during which death (or specific event) cannot occur according to the study design (Fig. 4).

As follow-up time increases, treatment decisions are often delayed or followed without treatment. Therefore, as the observation period elapses, the subjects of the patient and control groups defined at the start of treatment may change. Immortal time bias includes misclassification bias and selection bias.

The immortal bias can be reduced by using the Landmark method or the Mantel-Byar method (Fig. 5) [20,21]. In the Mantel-Byar method, the time starts at the moment of therapy initiation with all patients in the "non-response" state. Those who eventually respond to therapy enter the "response" state at the time of response and remain there until death or censoring, and those who do not respond always remain in the "non-response" state. This method removes the bias as patients are compared according to their response status at various periods during their follow-up [21]. Landmark analysis is a method of analyzing patients who survived after the landmark period. In this method, time starts being measured at a fixed time after the initiation of therapy. This fixed time is arbitrary but must be clinically meaningful [21].

Conclusion

Globally, interest in and use of big data in health care

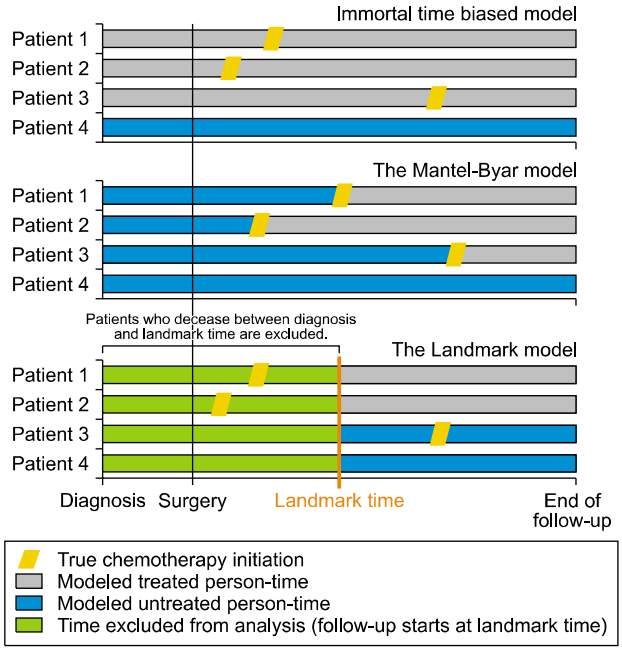

Fig. 5. Study design of the Landmark method and the Mantel-Byar strategy that can complement immortal time.

has risen in recent years. In Korea, the use of big data in health care is accelerating due to the opening of big data through the NHIS and HIRA open data systems. However, because data collecting aims to perform reviewing for reimbursement, medical research requires a reprocessing procedure. There are restrictions, such as on linking with external data (from other institutions, nations, etc.) and the lack of non-reimbursement information. However, a comprehensive study is attainable by using a proper operational definition to define pediatric cancer patients and combine accurate cancer diagnosis data from national cancer registration data with mortality data from the Statistics Korea.

Acknowledgment

This work was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI21C2046).

Conflict of Interest Statement

The author has no conflict of interest to declare.

References

1. Tresp V, Overhage JM, Bundschus M, Rabizadeh S, Fasching PA, Yu S. Going digital: a survey on digitalization and large-scale data analytics in healthcare. Proceedings of the IEEE 2016;104:2180-206.
2. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA 2014;311:2479-80.
3. Chung H, Kim S, Kim H. Clinical research from a health insurance database: practice and perspective. Korean J Med 2019;94:463-70.
4. Park JS, Lee CH. Clinical study using healthcare claims database. J Rheum Dis 2021;28:119-25.
5. National Health Insurance Sharing Service. Wonju, Korea: 2022. (Accesed February 11, 2022, at https://nhiss.nhis.or.kr/bd/ay/bdaya001iv.do)
6. HIRA - Healthcare Bigdata Hub. Wonju, Korea: Health Insurance Review & Assessment Service, 2022. (Accesed February 11, 2022, at https://opendata.hira.or.kr/home.do)
7. Kim DW, Lee SM, Lim HS, et al. A study on the operational definition of disease based on health insurance claim data. Goyang, Korea: NHIS Ilsan Hospital Institute Health Insurance & Clinical Research, 2017;88.
8. Meuli L, Dick F. Understanding confounding in observational studies. Eur J Vasc Endovasc Surg 2018;55:737.
9. Kahlert J, Gribsholt SB, Gammelager H, Dekkers OM, Luta G. Control of confounding in the analysis phase - an overview for clinicians. Clin Epidemiol 2017;9:195-204.
10. Lim HS, Oh HC, Jang JH, et al. Research on the development of an analysis method inspection tool to improve the quality of big data research using the National Health Information DB - Methodology review of the literature on the use of the National Health Information DB. Goyang, Korea: NHIS Ilsan Hospital Institute Health Insurance & Clinical Research, 2020-20-015.
11. Jagsi R, Bekelman JE, Chen A, et al. Considerations for observational research using large data sets in radiation oncology. Int J Radiat Oncol Biol Phys 2014;90:11-24.
12. Jager KJ, Zoccali C, Macleod A, Dekker FW. Confounding: what it is and how to deal with it. Kidney Int 2008;73:256-60.
13. Partlett C, Hall NJ, Leaf A, Juszczak E, Linsell L. Application of the matched nested case-control design to the secondary analysis of trial data. BMC Med Res Methodol 2020;20:117.
14. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. Pharmacoepidemiol Drug Saf 2005;14:465-76.
15. Ryu DR. Introduction to the medical research using national health insurance claims database. Ewha Med J 2017;40:66-70.
16. Benchimol EI, Manuel DG, Guttmann A, et al. Changing age demographics of inflammatory bowel disease in Ontario, Canada: a population-based cohort study of epidemiology trends. Inflamm Bowel Dis 2014;20:1761-9.
17. Cheng CL, Lee CH, Chen PS, Li YH, Lin SJ, Yang YH. Validation of acute myocardial infarction cases in the national health insurance research database in taiwan. J Epidemiol 2014;24:500-7.
18. Benchimol EI, Smeeth L, Guttmann A, et al. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. PLoS Med 2015;12:e1001885.
19. Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. BMC Med Res Methodol 2019;19:53.
20. Weberpals J, Jansen L, Silversmit G, et al. Comparative performance of a modified landmark approach when no time of treatment data are available within oncological databases: exemplary cohort study among resected pancreatic cancer patients. Clin Epidemiol 2018;10:1109-25.
21. Delgado J, Pereira A, Villamor N, López-Guillermo A, Rozman C. Survival analysis in hematologic malignancies: recommendations for clinicians. Haematologica 2014;99:1410-20.