KJA
Korean Journal of Anesthesiology

Statistical Round

Korean J Anesthesiol 2020;73(3):205-211
https://doi.org/10.4097/kja.20129
pISSN 2005-6419 • eISSN 2005-7563

Received: March 23, 2020
Revised: April 1, 2020 (1st); April 24, 2020 (2nd)
Accepted: May 12, 2020

Corresponding author:
EunJin Ahn, M.D.
Department of Anesthesiology and Pain
Medicine, Chung-Ang University College of
Medicine, 102 Heukseok-ro, Dongjak-gu,
Seoul 06973, Korea
Tel: +82-2-6299-2582
Fax: +82-2-6299-2585
Email: compassion10@gmail.com
ORCID: https://orcid.org/0000-0001-6321-5285

Introducing big data analysis using data from National Health Insurance Service

EunJin Ahn

Department of Anesthesiology and Pain Medicine, Chung-Ang University College of Medicine, Seoul, Korea

Among the different providers of health care big data in Korea, the data provided by the National Health Insurance Database include the medical information of all the citizens who have subscribed to medical insurance. As such, the data have representativeness and completeness. In order to conduct research using these National Health Insurance Database data, it is necessary to understand the characteristics of the claim data to avoid various biases, and to control confounding variables when making various operational definitions in the planning stage of the research. Moreover, without a proper understanding of the big data, it is possible during the analysis and data interpretation to mistakenly interpret the correlation between variables as a causal relationship. Therefore, in order to help advanced medical science, which reflects the medical reality such as medical expenses and number of hospital visits by clearly recognizing and analyzing the characteristics and limitations of health care big data, this author has dealt with the use of data sharing services provided by the National Health Insurance Database.

Keywords: Cohort; Correlation; Customized research database; Database; Korea; National Health Insurance; Operational definition; Public health service; Sample research database; Statement.

Research background and purpose

The digitization of documents has led to an abundance in the health care data, which range from electronic medical records used by hospitals to the data collected at the national level. The International Genome Sample Resource, world's largest human genome variation data released for free on the Amazon web (www.internationalgenome.org), is an example of how to build and use a health care database. The Ontario Institute of Technology, Canada also developed a system to predict pathogen infection by analyzing data from premature infants placed in incubators [1]. The Cleveland Clinic in the United States uses the data collected through its own network to meet the needs of emergency patients.

The Korea National Health and Nutrition Examination Survey, Health Insurance Review and Assessment Service, and National Health Insurance Database are the representatives of health care big data in Korea. Recently, data from four organizations (Korea Center for Disease Control, Health Insurance Review and Assessment Service, National Health Insurance Database, and National Cancer Center) were linked for use in the big data platform for health care (https://hcdl.mohw.go.kr/), providing researchers open access to data for public research purposes. The government is currently devising policies to provide open data for health care research. Through this paper, the author intends to help access this data by understanding the concept of health care big data.

Types and characteristics of health care big data

The data collected through the Korea National Health and Nutrition Examination Survey are different from the data provided by the Health Insurance Review and Assessment Service or the National Health Insurance Database, and are hence suitable for cross-sectional studies with clinical information and blood tests on samples taken from population groups. The Korea Centers for Disease Control and Prevention (https://knhanes.cdc.go.kr/knhanes/main.do) provides open access to data, which can be analyzed with the SPSS statistical packages (SPSS Inc., USA) commonly used by the researchers. On the other hand, the data collected by the National Health Insurance Database and Health Insurance Review and Assessment Service are used for billing medical expenses. Korean nationals are obliged to subscribe for health insurance and pay health insurance premiums as per the insurance subscriber category they belong to. When someone uses medical services, Health Insurance Review and Assessment Service evaluates the medical care expenses incurred by the medical care institution and notifies the industrial complex and that medical care institution of the evaluation results. Since the National Health Insurance Database pays medical care costs to nursing homes based on the information provided by the Health Insurance Review and Assessment Service, it can be considered that the data provided by the National Health Insurance Database and the Health Insurance Review and Assessment Service are almost the same. Data from the Health Insurance Review and Assessment Service can be accessed through the Health Insurance Big Data Open System (http://opendata.hira.or.kr/home.do). The data are paid and are preferably analyzed using statistics processing program such as SAS (SAS institute Inc., USA). Although, compared to the National Health Insurance Database, the data provided by Health Insurance Review and Assessment Service include additional information related to drugs, the sample data have limitations in cohort research. Moreover, it is difficult to access clinical information as there are no screening data.

Among the health care big data providers introduced so far, the author would like to take a deeper look at the data sharing service provided by the National Health Insurance Database. Various studies have been conducted using data from the National Health Insurance Database, such as studies that analyzed socio-economic costs of specific diseases [2–4], studies that described prevalence, mortality, or causes of death [5–8], and studies that analyzed the causes of death according to socio-economic level and disease using the 10th percentile variable of health insurance premium income [9]. Other studies have analyzed mortality or morbidity according to specific medical practices (general anesthesia vs. regional anesthesia) [10,11].

Classification of National Health Insurance Database data

The data collected by the National Health Insurance Database, include detailed treatment practices and prescriptions based on the fee-for-service payment model, and the medical information of all the citizens who have signed up for medical insurance in Korea. Data include insurance eligibility and premiums from birth to death of all citizens, medical history of hospitals and national health examination results, rare refractory and cancer registration information, medical benefits data, and elderly long-term care data since 2002. As the data are provided in the form of a cohort, longitudinal studies are possible, and if a customized database (explained later) is used, studies including those on rare diseases with low prevalence are also possible. In addition, the data on cause and date of death are provided by Statistics Korea (http://kostat.go.kr). The National Health Insurance Database data related to the number of hospitalization days or medical expenses can be used by researchers for cost-effective analysis. Moreover, there is no need to collect or build data separately, and if clinical data are needed, research can be conducted using a screening cohort that actually includes clinical data.

The National Health Insurance Database's data sharing service is largely divided into customized and sample database (DB). The customized DB is representative of the transmission data provided by de-identifying health insurance and long-term care insurance data collected by the DB. As mentioned earlier, if the researcher aims to study rare diseases, customized DB is recommended because large volume of data can be accessed using the customized DB than the sample DB.

However, to protect patient information, data analysis is to be conducted in a secured room situated in the industrial complex.

The sample DB extracts high-demand data as a sample and de-identifies it for analysis, providing a sample cohort DB, medical check-up DB, elderly cohort DB, infant medical check-up DB, and working woman cohort DB. The sample cohort DB, medical check-up DB, and elderly cohort DB include data for 14 years from 2002 to 2015. Of these, the sample cohort DB includes data on 1,025,000 subjects, which is 2% of the total population, and provides data on samples representing the whole nation. The Infant medical check-up DB provides 8 years of data from 2008 to 2015 on 84,000 subjects, as a 5% population of infants who have undergone the 1st to 2nd checkups of infants and toddlers at the 1st and 2nd stages of births between 2008 and 2012 (5 years)

https://doi.org/10.4097/kja.20129

which was simple randomized. However, since the date of birth of the subject is not included, it is difficult to know the exact age of the child. The working women cohort DB includes data of 185,000 subjects, collected for 9 years from 2007 to 2015. Till end of December 2007, 5% of female job enrollers aged 15–64 were randomly extracted.

Both customized and sample DB include an eligibility table showing the health insurance subscriber's eligibility, a death table showing the years of birth and death, a table of injuries showing injuries, number of days spent hospitalized, and total medical expenses, and a table of medical history showing hospital activity and amount[1]. In addition, they provide information details of issuance of prescriptions for outpatient use, as well as a sickness history table, which includes all the history of diagnosis received. Furthermore, medical health check-up tables are provided for 54 variables, including smoking, drinking, physical activity, physical measurement, and blood tests. However, information in form of images such as CT scan or ultrasound is not provided. The data for each table are summarized in Table 1.

Characteristics of National Health Insurance Database data and how to use it

In order to conduct research using National Health Insurance Database data, it is necessary to understand the various characteristics of the claim data. First, it is difficult to know the details of non-salary items (plastic for cosmetic purposes, preventive care, etc.) because both the National Health Insurance Database and the Health Insurance Review and Assessment Service collect data for billing purposes, rather than for research purposes, and include only the salary details in the bill prescribed by the medical institution. Second, it is difficult to grasp the detailed medical history because the cost of a surgery is fixed which is covered by a diagnosis-related group. Third, when diagnosing a patient at the hospital, the severity of the disease tends to increase to avoid reducing insurance claims. On the contrary, even in case of a severe ailment, the sensitivity of the diagnosis may be lowered because the corresponding code may not be entered when it is not related to the salary.

Therefore, in the planning stage of the research, it is necessary to know how sensitive the diagnostic code of each disease is and how to specify the operational definition. To this end, it is imperative to first understand whether a planned research hypothesis can be identified using the particular data.

Operational definition

Operational definition refers to the condition that best extracts the data of a patient with a specific disease by maximally utilizing information such as diagnosis code, prescription of examination procedural codes, and drugs, which are alternatively available within big data. The degree to which the established operational definition identifies actual target patients and diseases is an important factor in determining the reliability of the study. In the National Health Insurance Database, 20Table provides general information about the wounded and the clinic, and 40Table provides information about all the diagnosis codes including every major and minor diagnosis. The name of the diagnosis is based on the Korean Standard Classification of Disease (6th Revision, KCD-6), which is a revised form of the International Classification of Disease (10th Revision). The National Health Insurance Database continues to study the validity of diagnostic names us-

Table 1. Tables Provided by National Health Insurance Database

Title of the table	Contents of the table
Qualification DB	Gender, age, location, type of subscription, social economic variables of the subject such as income rank, disability, death, etc.
Statement (20Table)	General information of the subject and the clinic, major diagnosis, minor diagnosis, number of days spent hospitalized, total medical expenses, etc.
Treatment details (30Table)	In-hospital activity (prescription, surgery, materials, etc.), cost of hospitalization, etc.
Type of disease (40Table)	Every diagnosis including all major and minor diagnoses (ICD-10 codes)
Prescription details (60Table)	Outpatient prescription details (drug code, dosage, etc.)
Medical health check-up DB	Composed of 54 variables including smoking, drinking, physical activity, body measurement, and blood test.
Clinic DB	Status, facility, equipment, and personnel data of clinics by type, establishment, and location.

DB: database, ICD: International Classification of Diseases.

[1] Statement refers to 20Table, treatment details to 30Table, type of disease to 40Table and prescription details to 60Table which are provided by year basis, statement unit. Therefore, the researcher need to reform the provided tables to make desired data set.

ing various operational definitions[2)]. Therefore, it is important to examine the validity of operational definition by referencing previous studies in the planning stage.

If there are no existing studies on the validity of operational definitions, researchers need to make sensitive operational definitions by making the most of information such as diagnosis code, prescription of examination and procedural codes, and drugs. For example, in [6], in order to diagnose Parkinson's disease, the operational definition of the disease was given as follows:

- The patient should at least once be treated using Parkinson's disease medications such as Levodopa, dopamine agonists (ropinirole, pramipexole, etc.), entacapone, amantadine, selegiline, rasagiline, anticholinergics (trihexyphenidyl HCl, benztropinemesylate, procyclidine) in the general hospital grade neurology department specified in Article 3 of the Medical Code among patients with GCD (Parkinson's disease) claim code in KCD-6.
- Among them, Parkinson's disease and Parkinson's syndrome (G21. Secondary Parkinson's disease, drug-induced secondary Parkinson's disease, vascular Parkinson's disease, etc.; G22. Parkinson's disease in other classified diseases; G23. Excludes patients with extrapyramidal symptoms and tremors) are included in the claim code as minor diagnosis or suspicion.
- However, the diagnosis codes of G21, G22, G23, and G25 are deleted from the final diagnosis during the disease period among excluded patients, and only G20 is included in the diagnosis of Parkinson's disease.

In the above study, operational definition including both the diagnostic code and the use of drugs was specified to diagnose patients with Parkinson's disease. In studies dealing with diseases, it is recommended to define an operational definition with high sensitivity and specificity in parallel with drug use or treatment as in the above study because the diagnosis code may be overextracted or underextracted. If the rheumatoid disease is defined only by serological test results, the specificity is very high, but the sensitivity is low; however, if a patient who is prescribed disease-modifying anti-rheumatic drugs (DMARDs) is searched, sensitivity and specificity can be higher in this operational definition [12].

Research topics so far have mainly dealt with anesthesia and pain medicine including general and regional anesthesia. Data extracted based on procedural code corresponding to the anesthesia type tends to be more accurate compared to data extracted based on diagnosis code which could be overextracted or underextracted. In addition, the total cost of general or regional anesthesia can be utilized to indirectly calculate the anesthesia time, making it easy to apply in research.

How to control confounding variables

When planning a study using confounding variables, external confounding and disturbance factors due to internal variables need to be taken into account in the study plan.

External confounding factors include insurance standards such as medical technology, drug codes, treatment codes, and other standards of salary. For example, if a specific medical procedure can be paid for up to 3 times, the treatment method can be changed regardless of whether or not the patient has improved. In addition, it is necessary to understand the medical use behavior, medical treatment process, and clinical environment of doctors based on the unique characteristics of the disease for diseases that are complex and clinically related to the research hypothesis. For example, if a patient has clinical symptoms, such as cognitive dysfunction after surgery, but the symptoms are temporary and do not meet the patient's medical purpose, the attending doctor can only treat the symptoms without entering the diagnosis code. In the same context, it is also necessary to collect information on the accuracy of the disease as described above, and to verify how well the established operational definition identifies the actual patient or disease.

The internal confounding factor of a study is a factor that can occur in a retrospective study and refers to the relationship of various variables. For example, when the relationship between thyroid cancer and dementia is analyzed and a positive correlation is obtained, a confounding variable that both thyroid cancer and dementia have is a high incidence rate in women. In order to control confounding variables in such cases, the research subject can be 'restricted' to 'women' or when selecting a control group, gender matching can be used to control the disturbance factors according to gender. In a study analyzing the relationship between type of anesthesia and mortality rate in elderly who have undergone hip fracture surgery, age may act as a disturbing factor when the average age of a general anesthesia group is higher than that of a regional anesthesia group. In this case, it is possible to control the confounding factors according to the age by stratifying and analyzing the patient's age in the 60s, 70s, 80s, or higher. Alternatively, internal confounding factors can be controlled through multivariate analysis such as linear regression model, logistic regression model, Poisson, and Cox regression.

When analyzing health insurance data, it is difficult to select an ap-

[2)] Kim D. A study on the operational definition of disease based on health insurance claim data. Goyang: National Health Insurance Institute Ilsan Hospital; 2017 Dec. 88 p. Report No.: 2017-20-029. Available from http://www.alio.go.kr/download.dn?

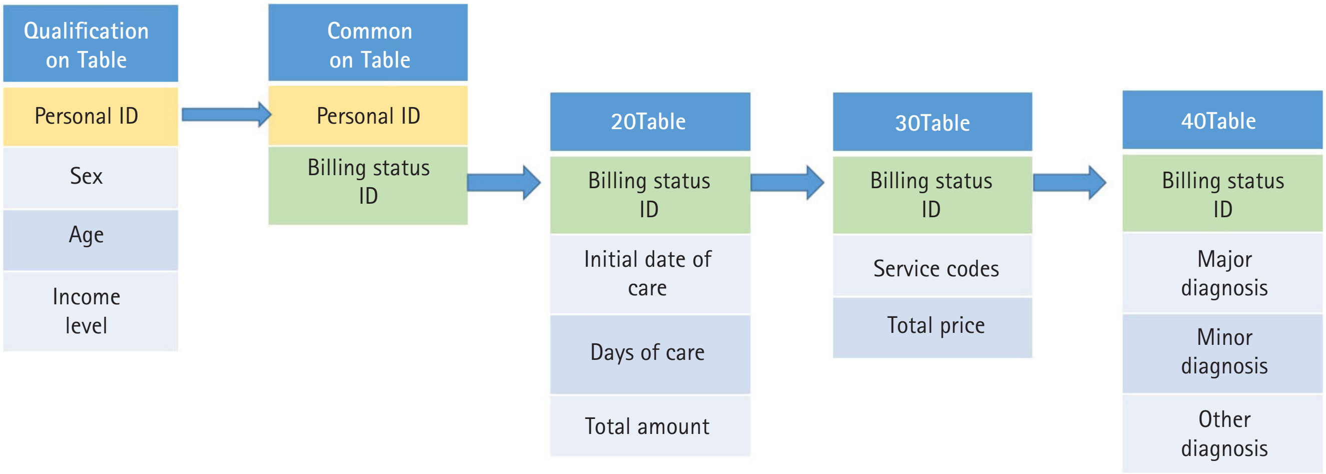

Fig. 1. Process of combining provided tables from National Health Insurance Database by personal ID and billing status ID.

propriate control group. The patient's diagnostic code is used to identify the underlying disease, and the Charlson comorbidity index [13] or Elixhauser comorbidity score [14] is used to match the propensity score. The Charlson comorbidity index and the Elixhauser comorbidity score are based on medical data and are thought to be of great help in further research if a shared disease classification system is established that reflects mortality and fatality in surgical patients.

Analyzing National Health Insurance Database

The data provided by the National Health Insurance Database were simply expressed as shown in Fig. 1. The qualification table provides information such as the patient's ID, age, gender, and income level. If a patient makes multiple claims, initial date of hospitalization, duration of hospitalization, cost of treatment, contents of treatment, and the name of diagnosis can be provided according to each billing unit. However, since the discharge date is not specified in the data provided, it is necessary to rebuild the bill-based data into hospitalization data. For example, if a long-term inpatient has been hospitalized for three months, the bill may be split into three if the patient is billed every month. In this case, if the initial date of hospitalization on the next bill is the same as the date which was added to the duration of hospitalization with initial date of hospitalization in the prior bill, it can be viewed as one hospitalization episode. Furthermore, even if the patient has been transferred to some other hospital, it can be assumed to be one episode if the initial date of hospitalization and the duration of hospitalization is within 2 days of the initial date of hospitalization on the next bill, because it is considered a re-hospitalization for the same diagnosis. This is called an inpatient episode, and when analyzing it, refer to the commands below to help in practice. We would like to show you how to group the inpatient records for each billing unit by the inpatient episode by using the variable for the initial date and the duration of the hospitalization provided in the qualification table (20Table). Instructions were analyzed using SAS v9.4 (SAS institute Inc., USA).

Looking at the SAS command below, since the discharge date is not provided, the hypothetical discharge date (end_date) is set as the initial date of hospitalization (start_date) + duration of hospitalization (mdcare_dd_cnt). If the hypothetical discharge date (end_date) and the next billing start date (start_date2) are within

```
libname sample ‘D:sample’;

data episode1;
set d20;
run;

Data episode2;
Set episode1;
mdcare_start_date = input(put(mdcare_strt_dt,8.),yymmdd8.);
Run;
data episode3;
set episode2;
mdcare_end_dt = mdcare_start_date+mdcare_dd_cnt;
keep rn_indi rn_key mdcare_strt_dt mdcare_dd_cnt mdcare_end_dt mdcare_start_date;
run;

proc sort data = episode3;
by rn_indi mdcare_start_date;

Data episode4;
Set episode3;
By rn_indi;
Format start_date start_date2 end_date end_date2 yymmdd8.;
start_date = mdcare_start_date;
```

```
end_date = mdcare_end_dt;
Retain start_date2 0 end_date2 0 c 0;
If first.rn_indi then do;
C = 1;
start_date2 = start_date;
end_date2 = end_date;
end;
else if 0 < = start_date- end_date2 < = 2 then do;
c = c;
start_date2 = min(start_date, start_date2);
end_date2 = max(end_date, end_date2);
end;
else do;
c+1;
start_date2 = start_date;
end_date2 = end_date;
end;
run;

proc sort data = episode4;
by rn_indi c;
run;

data episode;
set sample.episode4;
by rn_indi c;
if last.c;
run;
```

2 days, the same "c" value is given. Finally, the start date of care (start_date2) + the duration of hospitalization (mdcare_dd_cnt) of the invoice with the last c value for each patient identification number is defined as the actual discharge date (end_date2). For better understanding, the result data before and after applying the last c-values were arranged in an Excel file (Supplementary 1).

Note of caution during big data analysis and data interpretation

Correlation defines the relationship between two variables by observing whether a variable increases in the same direction or decreases in the opposite direction when the other variable increases. If the two variables move in the same direction, they are said to be in positive correlation, and if they move in opposite directions, they are in a negative correlation. Correlation only indicates the degree of closeness between the two variables; it does not specify the cause and effect of this closeness. On the other hand, causation or causality defines the cause and effect relationship between two variables. For example, it can be seen that there is both a correlation and a causal relationship between height and weight because weight increases with height. However, as weight increases, height does not increase; hence, there is a correlation between weight and height, but there is no causal relationship. Big data has limitation in that it cannot be used to prove the causal relationship, and interpreting the correlation as a causal relationship can result in a huge error. Therefore, in order to not make such an error, it is necessary to make good operational definitions to avoid various biases, and to properly control disturbance variables.

Fisher proposed using the P value as a tool to see how well the data fit the null hypothesis [15]. However, big data includes a vast number of samples, and the P value can be significant for almost all the variables. Therefore, big data analysis has another limitation in that it is difficult to consider the P value as a correct scale for determining statistical significance. Therefore, when analyzing the results, it is better to disclose the CI and effect size rather than relying only on the P value [16]. The P value can only be interpreted as 'with or without a significant difference', but the effect size actually shows how much the difference is, and unlike the P value, the effect size is not affected by the number of samples [16,17]. Though reliance on the P value alone is still debatable, data in recent times has become vast and complicated. There are many cases where reproducibility is difficult only with the P value, and there is a movement to apply the P value more strictly to 0.005 [18]. Alternatively, the null hypothesis and the alternative hypothesis can be compared and tested using the Bayesian statistical hypothesis test [19].

Summary

In Korea, the National Health Insurance Database's data sharing service can provide data from a vast number of cohorts. The collected data are based on insurance claims, and have certain limitations. If the researchers clearly recognize and analyze the characteristics and limitations of these data, it is expected to greatly help medical science through research that reflects the medical realities such as medical expenses and number of hospital visits.

Conflicts of Interest

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

Supplementary Materials

Supplementary 1. Result data before (sheet1) and after (sheet2) applying the last c-values. RN_INDI refers to personal ID. RN_KEY refers to billing status ID.

References

1. Khazaei H, McGregor C, Eklund JM, El-Khatib K. Real-time and retrospective health-analytics-as-a-service: a novel framework. JMIR Med Inform 2015; 3: e36.
2. Jung J, Seo HY, Kim YA, Oh IH, Lee YH, Yoon S-J. The economic burden of epilepsy in Korea, 2010. J Prev Med Public Health 2013; 46: 293-9.
3. Kang IO, Lee SY, Kim SY, Park CY. Economic cost of dementia patients according to the limitation of the activities of daily living in Korea. 2007; 22: 675-81.
4. Kim J, Rhee CK, Yoo KH, Kim YS, Lee SW, Park YB, et al. The health care burden of high grade chronic obstructive pulmonary disease in Korea: analysis of the Korean Health Insurance Review and Assessment Service data. Int J Chron Obstruct Pulmon Dis 2013; 8: 561-8.
5. Jung HS, Nho JH, Ha YC, Jang S, Kim HY, Yoo JI, et al. Incidence of osteoporotic refractures following proximal humerus fractures in adults aged 50 years and older in Korea. J Bone Metab 2019; 26: 105-11.
6. Lee JE, Choi JK, Lim HS, Kim JH, Cho JH, Kim GS, et al. The prevalence and incidence of Parkinson's disease in South Korea: a 10-year nationwide population-based study. J Korean Neurol Assoc 2017; 35: 191-8.
7. Park J, Suh B, Shin DW, Hong JH, Ahn H. Cause of death in Korean men with prostate cancer: an analysis of time trends in a nationwide cohort. J Korean Med Sci 2016; 31: 1802-7.
8. Yoo JI, Ha YC, Park KS, Kim RB, Seo SH, Koo KH. Incidence and mortality of osteoporotic refractures in Korea according to nationwide claims data. Yonsei Med J 2019; 60: 969-75.
9. Cho KH, Nam CM, Lee EJ, Choi Y, Yoo KB, Lee SH, et al. Effects of individual and neighborhood socioeconomic status on the risk of all-cause mortality in chronic obstructive pulmonary disease: a nationwide population-based cohort study, 2002-2013. Respir Med 2016; 114: 9-17.
10. Ahn EJ, Kim HJ, Kim KW, Choi HR, Kang H, Bang SR. Comparison of general anaesthesia and regional anaesthesia in terms of mortality and complications in elderly patients with hip fracture: a nationwide population-based study. BMJ Open 2019; 9: e029245.1
11. Suh B, Shin DW, Park Y, Lim H, Yun JM, Song SO, et al. Increased cardiovascular risk in thyroid cancer patients taking levothyroxine: a nationwide cohort study in Korea. Eur J Endocrinol 2019; 180: 11-20.
12. Cho SK, Sung YK, Choi CB, Kwon JM, Lee EK, Bae SC. Development of an algorithm for identifying rheumatoid arthritis in the Korean National Health Insurance claims database. Rheumatol Int 2013; 33: 2985-92.
13. Glasheen WP, Cordier T, Gumpina R, Haugh G, Davis J, Renda A. Charlson Comorbidity Index: ICD-9 update and ICD-10 translation. Am Health Drug Benefits 2019; 12: 188-97.
14. Li B, Evans D, Faris P, Dean S, Quan H. Risk adjustment performance of Charlson and Elixhauser comorbidities in ICD-9 and ICD-10 administrative databases. BMC Health Serv Res 2008; 8: 12.
15. Fisher KA. Statistical tests. Nature 1935; 136: 474.
16. Lee DK. Alternatives to P value: confidence interval and effect size. Korean J Anesthesiol 2016; 69: 555-62.
17. Nahm FS. Understanding effect sizes. Hanyang Med Rev 2015; 35: 40-3.
18. Ioannidis JP. The proposal to lower P value thresholds to .005. JAMA 2018; 319: 1429-30.
19. Halsey LG. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? Biol Lett 2019; 15: 20190174.