

Original Article
Medicine General & Policy



Building Linked Big Data for Stroke in Korea: Linkage of Stroke Registry and National Health Insurance Claims Data

Tae Jung Kim ,¹ Ji Sung Lee ,² Ji-Woo Kim ,³ Mi Sun Oh ,⁴ Heejung Mo ,¹ Chan-Hyuk Lee ,¹ Han-Young Jeong ,¹ Keun-Hwa Jung ,¹ Jae-Sung Lim ,⁴ Sang-Bae Ko ,¹ Kyung-Ho Yu ,⁴ Byung-Chul Lee ,⁴ and Byung-Woo Yoon ¹

¹Department of Neurology, Seoul National University Hospital, Seoul, Korea

²Clinical Research Center, Asan Medical Center, University of Ulsan, Seoul, Korea

³Department of Bigdata, Health Insurance Review and Assessment Service, Wonju, Korea

⁴Department of Neurology, Hallym University Sacred Heart Hospital, Anyang, Korea



Received: Aug 30, 2018

Accepted: Oct 2, 2018

Address for Correspondence:

Byung-Woo Yoon, MD, PhD

Department of Neurology, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.

E-mail: bwoon@snu.ac.kr

*Tae Jung Kim, Ji Sung Lee, and Ji-Woo Kim contributed equally to this work.

© 2018 The Korean Academy of Medical Sciences.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ORCID iDs

Tae Jung Kim

<https://orcid.org/0000-0003-3616-5627>

Ji Sung Lee

<https://orcid.org/0000-0001-8194-3462>

Ji-Woo Kim

<https://orcid.org/0000-0002-4070-3021>

Mi Sun Oh

<https://orcid.org/0000-0002-6741-0464>

Heejung Mo

<https://orcid.org/0000-0001-7810-035X>

Chan-Hyuk Lee

<https://orcid.org/0000-0002-3421-0909>

Han-Young Jeong

<https://orcid.org/0000-0002-3373-118X>

ABSTRACT

Background: Linkage of public healthcare data is useful in stroke research because patients may visit different sectors of the health system before, during, and after stroke. Therefore, we aimed to establish high-quality big data on stroke in Korea by linking acute stroke registry and national health claim databases.

Methods: Acute stroke patients (n = 65,311) with claim data suitable for linkage were included in the Clinical Research Center for Stroke (CRCS) registry during 2006–2014. We linked the CRCS registry with national health claim databases in the Health Insurance Review and Assessment Service (HIRA). Linkage was performed using 6 common variables: birth date, gender, provider identification, receiving year and number, and statement serial number in the benefit claim statement. For matched records, linkage accuracy was evaluated using differences between hospital visiting date in the CRCS registry and the commencement date for health insurance care in HIRA.

Results: Of 65,311 CRCS cases, 64,634 were matched to HIRA cases (match rate, 99.0%). The proportion of true matches was 94.4% (n = 61,017) in the matched data. Among true matches (mean age 66.4 years; men 58.4%), the median National Institutes of Health Stroke Scale score was 3 (interquartile range 1–7). When comparing baseline characteristics between true matches and false matches, no substantial difference was observed for any variable.

Conclusion: We could establish big data on stroke by linking CRCS registry and HIRA records, using claims data without personal identifiers. We plan to conduct national stroke research and improve stroke care using the linked big database.


Keywords: Big Data; Data Linkage; Stroke Registry; National Health Claim Data

INTRODUCTION


Linked health datasets from multiple sources offer a powerful resource for conducting epidemiological and clinical studies.¹⁻⁴ Linkage with public healthcare data is useful in stroke research because stroke patients may enter into different sectors of the healthcare system before, during, and after stroke. Linked large-scale datasets enable researchers and healthcare

Keun-Hwa Jung 


<https://orcid.org/0000-0003-1433-8005>

Jae-Sung Lim 

<https://orcid.org/0000-0001-6157-2908>

Sang-Bae Ko 

<https://orcid.org/0000-0002-9429-9597>

Kyung-Ho Yu 

<https://orcid.org/0000-0002-8997-5626>

Byung-Chul Lee 

<https://orcid.org/0000-0002-3885-981X>

Byung-Woo Yoon 

<https://orcid.org/0000-0002-8597-807X>

Funding

This work was supported by the Ministry of Health and Welfare (HI 16C1078), Korea. The funding organization had no role in the study or in the preparation of this report.

Disclosure

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Kim TJ, Oh MS, Jung KH, Ko SB, Yu KH, Yoon BW. Data curation: Kim TJ, Kim JW, Oh MS, Jung HY, Lim JS. Formal analysis: Kim TJ, Kim JW. Investigation: Mo H, Jung HY, Lim JS. Methodology: Kim TJ, Kim JW, Mo H. Supervision: Oh MS, Lee CH, Jung KH, Ko SB, Yu KH, Lee BC, Yoon BW. Validation: Lim JS. Writing - original draft: Kim TJ, Lee JS. Writing - review & editing: Ko SB, Yu KH, Lee BC, Yoon BW.

practitioners to obtain a comprehensive view of stroke care and to improve national stroke care systems.^{5,6} There are several advantages to using large-scale medical datasets in Korea because health data are computerized and there is a government-administered universal health insurance system.⁷ However, access to personal identifiers is strictly prohibited by the Personal Information Protection Act (PIPA) in Korea; and many uses of health data are consequently discouraged, especially including linkages to external data using personal identifiers. These confidentiality requirements even apply to research that is conducted with the aim of improving public health. Although the requirements are known to affect the linkage process, large-scale linked administrative datasets are demonstrating increasing importance in epidemiological and clinical stroke research, since their use can improve research quality and transparency.^{8,9} Since 2006, the Clinical Research Center for Stroke (CRCS), a 9-year research project supported by government funding, has established the largest nationwide acute stroke registry in Korea.¹⁰⁻¹² Therefore, we aimed to build a large dataset on stroke in Korea by linking the CRCS stroke registry with the Health Insurance Review and Assessment Service (HIRA) administrative claim database.

METHODS

The CRCS data and the HIRA data

The CRCS registry was started in 2006 to collect data on acute stroke or transient ischemic attack (TIA) (within 7 days after onset) in Korea. The CRCS is supported by the Korea Healthcare Technology R&D Project of the Ministry of Health and Welfare in the Republic of Korea. Using a web-based database, CRCS collects clinical information on all acute stroke patients hospitalized at the neurology departments of a total of 65 participating hospitals. According to predefined protocols, demographic features, risk factors, stroke characteristics, treatment information, National Institute of Health Stroke Scale (NIHSS) scores, and laboratory information were collected at the time of entry into the registry database by stroke physicians or trained nurses. Data quality is monitored and audited regularly.¹⁰⁻¹²

The HIRA collected and managed claims data related to the National Health Insurance (NHI) program in the process of reimbursing healthcare providers in Korea. Accordingly, the HIRA database contains all information on the diagnoses, treatments, and prescribed medications for approximately 50 million Koreans. The information on prescribed drugs includes brand name, generic name, prescription date, duration of administration, and route of administration. In addition, all diagnoses are coded according to the International Classification of Disease, Tenth Revision (ICD-10).^{7,13-15}

Data cleaning and preparation for linking datasets

We initially screened 108,430 stroke registry cases recorded from 65 participating hospitals in 2006 to 2014. These cases were screened based on the CRCS identifier. We excluded case records from 31 hospitals from which the patients were inconsistently registered ($n = 8,709$), patients who visited a hospital more than 7 days after stroke symptom onset ($n = 3,113$), and patients with inaccurate insurance claim data ($n = 31,297$). A total of 65,311 patients from the hospitals were finally included in the dataset that was used for linkage (**Fig. 1**).

Data linkage methods

We linked CRCS and HIRA data (2007–2017) via a type of statistical matching that used common variables that were shared and stored in the enrolled hospitals and the HIRA. First, we used the claim data to identify common variables for linking the CRCS and the HIRA

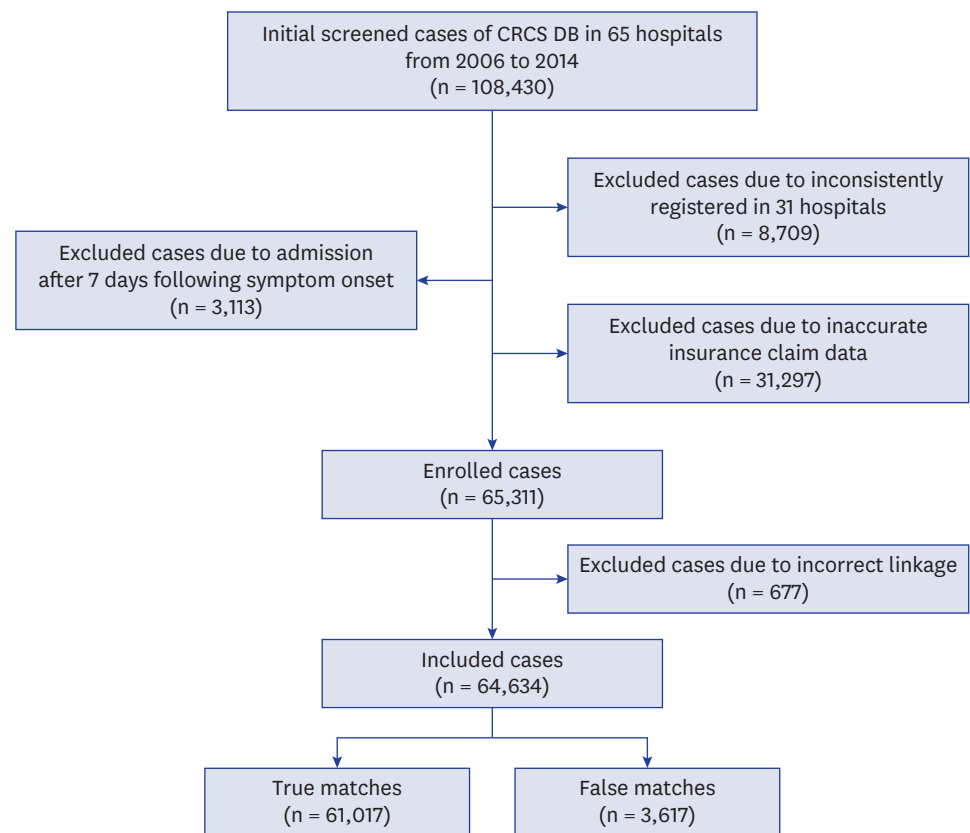


Fig. 1. Flow diagram of included cases for matching.
CRCS DB = Clinical Research Center for Stroke database.

data. The selected common variables needed to be accurate, and there were no missing data for the linking process.¹⁶ From claim data, we choose four variables for matching: provider identification, receiving year, receiving number, and statement serial number. Additionally, we selected gender and date of birth as common variables for linking the two databases. Together, these six variables were used as the matching variables for the data linkage. The matching process was performed in the server for HIRA using Sybase IQ software (Sybase Inc., Dublin, CA, USA). After the linking was completed, all linked data were de-identified before analysis of the dataset.

Analysis of linkage accuracy and statistical analysis using linked data

First, we assess the matching rate (1:1 matching) of CRCS data that had been linked to HIRA data. Second, we evaluated linkage quality and errors during the linkage process. To assess the quality of the linked data, we compared the hospital visiting date in the CRCS data to the commencement date for health insurance care in the HIRA. If the difference between the two dates was 7 or fewer days, we accepted the case as a true match. In addition, we used absolute standardized differences (ASDs) to compare the baseline characteristics of true matches and false matches in linked data. ASD analysis was used because it is expected to be more informative than *P* values for comparing large linked datasets, and may help to identify variables affected by potential bias due to linkage errors.^{3,17,18} For all variables, ASDs that are less than 0.2 represent small standardized differences.^{3,18} Finally, we created a linked CRCS-HIRA database based on the truly matched cases. The purpose of this database was to allow analyses of outcomes after index stroke.

Statistical analysis

We presented categorical data as frequencies (proportions) and normally distributed continuous variable data as means \pm standard deviations (SDs) or medians (interquartile ranges [IQRs]), as appropriate. To compare the baseline characteristics between groups, univariate analyses were conducted using either Student's t-tests or Mann-Whitney U tests for continuous variables and χ^2 or Fisher's exact tests for categorical variables. Statistical analyses were performed using the SAS statistical software (Release 9.4; SAS Institute Inc., Cary, NC, USA).

Ethics statement

The study was approved by the Institutional Review Board (IRB) of Seoul National University Hospital, 33 participating hospitals and HIRA (IRB No. H-1608-078-785). Informed consent was waived by the IRB.

RESULTS

Accuracy of data linkage between CRCS and HIRA data

A total of 65,311 cases were processed using the linkage algorithm, of which 677 were unmatched or one-to-many (1:M) matched. In total, 64,634 cases were one-to-one (1:1) matched in the HIRA dataset; the overall matching rate was 99.0%.

As described in the Methods, we classified matches as true or false based on the difference between the hospital visiting date in the CRCS data and the commencement date for health insurance care in the HIRA data. Among the matched records, 61,017 cases (94.4%) were truly matched and 3,617 cases (5.6%) were falsely matched, giving an accuracy rate in the total matched dataset of 94.4% (**Fig. 1**). The baseline characteristics of true matches and false matches are summarized in **Table 1**. When we used ASD values to compare the baseline characteristics of true matched cases and false matched cases, no substantial difference was observed for any variable (**Table 1**).

The characteristics of the true matches were analyzed in detail. The mean age was 66.4 years and 58.4% of the patients were men. Recanalization treatments were received in 13.9% of the cases (intravenous [IV] thrombolysis 67.5%, endovascular treatment 16.0%, and combined IV thrombolysis and endovascular treatment 16.5%) and the median NIHSS score was 3 (IQR 1–7). Of the cases, 91.1% (n = 52,213) were ischemic stroke, 7.0% (n = 3,988) were TIA, and 1.9% (n = 1,113) were hemorrhagic stroke. Among the cases of ischemic stroke, 34.9% were accounted for by large artery atherosclerosis, 24.2% by small vessel occlusion, 18.6% by cardioembolism, 2.6% by other determined factors, and 19.7% by undetermined factors.

DISCUSSION

We have established a large dataset on stroke by linking the CRCS registry and administrative HIRA data. A matching rate of 99.0% was achieved by using data from claims as matching variables, without relying on personal identifiers. Additionally, the accuracy of the linkage was high (94.4%). Moreover, there was no substantial difference in baseline characteristics between the true matches and the false matches.

Table 1. Baseline characteristics of matched cases according to linkage status

Variables	True matches (n = 61,017)	False matches (n = 3,617)	P value	ASD
Age, mean \pm SD, yr	66.4 \pm 12.7	66.7 \pm 11.8	0.155	0.023
Gender, men, No. (%)	35,631 (58.4)	1,990 (55.0)	< 0.001	0.068
HTN, No. (%)	42,934 (70.4)	2,497 (69.0)	0.089	0.029
DM, No. (%)	20,411 (33.5)	1,221 (33.8)	0.705	0.006
HL, No. (%)	17,805 (29.2)	846 (23.4)	< 0.001	0.132
Previous stroke Hx., No. (%)	10,662 (17.5)	759 (21.0)	< 0.001	0.089
Coronary heart disease, No. (%)	4,544 (7.4)	177 (4.9)	< 0.001	0.106
A.fib, No. (%)	10,592 (17.4)	526 (14.5)	< 0.001	0.077
Smoking, No. (%)	23,720 (38.9)	1,196 (33.1)	< 0.001	0.121
Initial NIHSS, median (IQR)	3 (1–7)	3 (1–7)	0.120	0.021
Types of stroke, No. (%)			< 0.001	
Ischemic stroke	52,213 (91.1)	3,020 (91.1)		0.001
Hemorrhagic stroke	1,113 (1.9)	100 (3.0)		0.069
TIA	3,988 (7.0)	194 (5.9)		0.045
Stroke mechanisms, No. (%)			0.149	
LAA	18,236 (34.9)	1,018 (33.7)		0.026
SVO	12,617 (24.2)	784 (26.0)		0.041
CE	9,736 (18.6)	535 (17.7)		0.024
Other determined	1,337 (2.6)	83 (2.7)		0.012
Undetermined	10,287 (19.7)	600 (19.9)		0.004
Recanalization treatment, No. (%)	8,457 (13.9)	347 (9.6)		
IV thrombolysis	5,706 (67.5)	221 (63.7)	< 0.001	0.122
Endovascular treatment	1,353 (16.0)	76 (21.9)	0.644	0.008
Combined IV thrombolysis and endovascular treatment	1,398 (16.5)	50 (14.4)	< 0.001	0.068

ASD = absolute standardized difference, SD = standard deviation, HTN = hypertension, DM = diabetes mellitus, HL = hyperlipidemia, Hx. = history, A.fib = atrial fibrillation, NIHSS = national institute of health stroke scale, IQR = interquartile range, TIA = transient ischemic attack, LAA = large artery atherosclerosis, SVO = small vessel occlusion, CE = cardioembolism, IV = intravenous.

Administrative big data from healthcare databases could be an important resource for clinical studies. In the present study, the healthcare big data involved population-based databases and charge information. It provides easy-to-access longitudinal data from the period after the patient has left the hospital, for follow-up and trend analyses.^{4,5,19} The HIRA data include rich and diverse information on healthcare services, such as diagnosis, treatments and procedures; histories of prescribed medications; costs; and medical care institutions that belong to the universal insurance coverage system. HIRA data could cover 98%–99% of the total population of Korea. Moreover, the HIRA data contain both in-hospital and out-patient clinics information, allowing longitudinal follow-up of utilization histories across the full range of health care services.^{7,8,15} However, the HIRA data do not include clinical information, such as on management during acute stage, severity of disease, laboratory findings, risk factors, or clinical history from the time of disease occurrence. The CRCS registry includes a variety of clinical characteristics, such as clinical history, treatment, severity of stroke, and laboratory information from acute stroke management. However, it is difficult to investigate and analyze stroke outcomes using the CRCS data because they do not include longitudinal outcomes or follow-up information after the index stroke. Therefore, linkage of the HIRA and clinical data, such as the CRCS data from hospitalizations, could serve as a powerful research resource to study stroke prognosis and healthcare service utilization, from acute to chronic stages of stroke.

Despite the importance of linking big data, the linkage of public health data (even for the public good) is limited in Korea because access to personal identifiers is severely regulated by PIPA even for research.^{8,14,15} Therefore, instead of using personal identifiers, we linked the CRCS registry data and the HIRA data using data from claims for the linkage variables.

Although the matching rate and accuracy may be low in comparison with deterministic linkages using personal identifiers, the accuracy and matching rates that we obtained were similar to previous reports using probabilistic linkage (86%–99%), based on a type of statistical matching method using common variables in our study.^{1,5,20–22} Moreover, we assessed the accuracy of the matched data and evaluated the characteristics of the two datasets. In the HIRA database, the commencement date for health insurance care was generated from data that had been submitted from each center for reimbursement for medical services. Therefore, it may differ from the actual hospital visiting dates in the CRCS. We evaluated the accuracy of the matched data based on the difference between the two dates (≤ 7 days), in consideration of this discrepancy. The linkage of HIRA data to the CRCS registry data during acute stroke management could provide access to outcome data. In addition, the linkage of administrative data in the HIRA can be a significant resource for updating missing registry data. Medical histories are important for evaluating vascular risk factors in stroke patients prior to the index stroke; therefore, it is important to update medical histories of registered patients. Moreover, we can evaluate and follow the outcomes and complications following strokes based the linked data. Longitudinal administrative data linkage can also improve estimates of stroke recurrence, mortality, and health services monitoring.

This study had several limitations. First, the accuracy of our linked data may be lower than deterministic linkages method using personal identifiers. Moreover, we did not evaluate the sensitivity and specificity of our linked dataset, because it is difficult to evaluate the unlinked cases (such as false negatives or true negatives) in our study due to the PIPA regulations in Korea. Although there were limitations to linking method and accuracy, the matching rate and the data accuracy in the present study are similar to those from previous studies.^{1,20} Therefore, the linkage method and accuracy of linked data of our study are reliable. Second, of all the patients, 3,617 were falsely matched, therefore these missing cases and linkage errors could have led to biased result. However, no substantial differences were observed when comparing the baseline characteristics of true matches and false matches to evaluate potential sources of bias. Therefore, the bias deriving from falsely matched data might be minimal. Third, our linking method is only possible in studies with access to information from hospitals. We linked data using the common claim data in each hospital record and in the HIRA. Moreover, data linkage accuracy is dependent on the quality of matching variables. Therefore, it would be difficult to use our method for linking data from cohort studies that do not have access to claims. Fourth, information associated with non-covered healthcare service are absent in the linked data because the claims data in the HIRA are collected for reimbursement of healthcare services under the insurance system in Korea. Despite these limitations, we have built a linked, large data source on stroke in Korea that has a high matching rate, without using personal identifiers. Moreover, using this linked stroke data, we expect to perform several nationwide stroke studies, including epidemiological analyses, comprehensive assessments of the national stroke care system, and research directed at the goal of improving stroke care.

In conclusion, our study shows the feasibility of linking the administrative data and registry data with high accuracy despite the limitations imposed by confidentiality requirements in Korea. Moreover, we constructed a large-scale linked stroke database that should allow the development of prognostic prediction models and the analysis of longitudinal outcomes following stroke. Using this resource, we expect that it will be possible to conduct a broader range of nationwide stroke research and improve stroke care. Further studies and efforts are needed to improve the accuracy of data linkage in Korea.

ACKNOWLEDGMENTS

We would like to acknowledge Professor Byung-Joo Park of the Department of Preventive Medicine, Seoul National University College of Medicine. We are grateful for his valuable comments on this study.

REFERENCES

1. Silveira DP, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. *Rev Saude Publica* 2009;43(5):875-82.
[PUBMED](#) | [CROSSREF](#)
2. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data linkage: a powerful research tool with potential problems. *BMC Health Serv Res* 2010;10(1):346.
[PUBMED](#) | [CROSSREF](#)
3. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol* 2017;46(5):1699-710.
[PUBMED](#) | [CROSSREF](#)
4. Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. *Annu Rev Public Health* 2011;32(1):91-108.
[PUBMED](#) | [CROSSREF](#)
5. Ido MS, Bayakly R, Frankel M, Lyn R, Okosun IS. Administrative data linkage to evaluate a quality improvement program in acute stroke care, Georgia, 2006–2009. *Prev Chronic Dis* 2015;12:E05.
[PUBMED](#) | [CROSSREF](#)
6. Zingmond DS, Ye Z, Ettner SL, Liu H. Linking hospital discharge and death records--accuracy and sources of bias. *J Clin Epidemiol* 2004;57(1):21-9.
[PUBMED](#) | [CROSSREF](#)
7. Kwon S. Thirty years of National Health Insurance in South Korea: lessons for achieving universal health care coverage. *Health Policy Plan* 2009;24(1):63-71.
[PUBMED](#) | [CROSSREF](#)
8. Kim L, Kim JA, Kim S. A guide for the utilization of health insurance review and assessment service national patient samples. *Epidemiol Health* 2014;36:e2014008.
[PUBMED](#) | [CROSSREF](#)
9. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: issues, methods, and directions for the future. *Health Serv Res* 2010;45(5 Pt 2):1468-88.
[PUBMED](#) | [CROSSREF](#)
10. Kim BJ, Park JM, Kang K, Lee SJ, Ko Y, Kim JG, et al. Case characteristics, hyperacute treatment, and outcome information from the clinical research center for stroke-fifth division registry in South Korea. *J Stroke* 2015;17(1):38-53.
[PUBMED](#) | [CROSSREF](#)
11. Park TH, Ko Y, Lee SJ, Lee KB, Lee J, Han MK, et al. Gender differences in the age-stratified prevalence of risk factors in Korean ischemic stroke patients: a nationwide stroke registry-based cross-sectional study. *Int J Stroke* 2014;9(6):759-65.
[PUBMED](#) | [CROSSREF](#)
12. Hong KS, Bang OY, Kang DW, Yu KH, Bae HJ, Lee JS, et al. Stroke statistics in Korea: Part I. Epidemiology and risk factors: a report from the Korean stroke society and clinical research center for stroke. *J Stroke* 2013;15(1):2-20.
[PUBMED](#) | [CROSSREF](#)
13. Kim HA, Kim S, Seo YI, Choi HJ, Seong SC, Song YW, et al. The epidemiology of total knee replacement in South Korea: national registry data. *Rheumatology (Oxford)* 2008;47(1):88-91.
[PUBMED](#) | [CROSSREF](#)
14. Shin JY, Choi NK, Jung SY, Lee J, Kwon JS, Park BJ. Risk of ischemic stroke with the use of risperidone, quetiapine and olanzapine in elderly patients: a population-based, case-crossover study. *J Psychopharmacol* 2013;27(7):638-44.
[PUBMED](#) | [CROSSREF](#)

15. Kim JA, Yoon S, Kim LY, Kim DS. Towards actualizing the value potential of Korea Health Insurance Review and Assessment (HIRA) data as a resource for health research: strengths, limitations, applications, and strategies for optimal use of HIRA data. *J Korean Med Sci* 2017;32(5):718-28.
[PUBMED](#) | [CROSSREF](#)
16. D'Orazio M. Statistical matching and imputation of survey data with StatMatch. <https://www.researchgate.net/publication/263888033>. Updated 2014. Accessed August 1, 2018.
17. Ford JB, Roberts CL, Taylor LK. Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. *Paediatr Perinat Epidemiol* 2006;20(4):329-37.
[PUBMED](#) | [CROSSREF](#)
18. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009;28(25):3083-107.
[PUBMED](#) | [CROSSREF](#)
19. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014;2(1):3.
[PUBMED](#) | [CROSSREF](#)
20. Capuani L, Bierrenbach AL, Abreu F, Takecian PL, Ferreira JE, Sabino EC. Accuracy of a probabilistic record-linkage methodology used to track blood donors in the Mortality Information System database. *Cad Saude Publica* 2014;30(8):1623-32.
[PUBMED](#) | [CROSSREF](#)
21. Harron K, Dibben C, Boyd J, Hjern A, Azimae M, Barreto ML, et al. Challenges in administrative data linkage for research. *Big Data Soc* 2017;4(2):2053951717745678.
[PUBMED](#) | [CROSSREF](#)
22. Park BJ, Stergachis A. Automated databases in pharmacoepidemiologic studies. In: Hartzema AG, editor. *Pharmacoepidemiology and Therapeutic Risk Management*. Cincinnati, OH: Harvey Whitney Books, 2008, 519-544.