Editorial

Healthc Inform Res. 2018 October;24(4):251-252.
https://doi.org/10.4258/hir.2018.24.4.251
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Making Sense of the Big Picture: Data Linkage and Integration in the Era of Big Data

Hyejung Chang, PhD
Editor of Healthcare Informatics Research, Kyung Hee University, Seoul, Korea

The advent of the Big Data era has naturally brought with it an undeniable growth of interest in the topic of data. Our explorations of this topic thus far have led to the conclusion that the challenge posed by the volume, variety, and velocity that characterize data must be addressed by data linkage and integration [1]. Analyzing multiple datasets by linking and integrating can answer questions that require large sample sizes or detailed data regarding hard-to-reach populations and can generate evidence with a high level of external validity and applicability for policy making [2].

The strategy of data linkage and integration—the process of combining data from different sources and providing users with a unified view of this data—is not a new concept. Even prior to the Big Data era, efforts were made to adopt and analyze various data to interpret real-world problems in the public and private sector. Researchers, however, realized the limitations of information from an individual dataset and thus proposed the integration of different datasets. Since this proposal, researchers have worked to improve the usefulness of data and the quality of knowledge derived from data analysis.

In the field of health, the acceleration of aging has led researchers to focus on identifying factors related to chronic diseases and symptoms, and data linkage and integration strategies have aided them in this pursuit. Chronic diseases and their associated health services result from various factors ranging from physical and genetic scatters of an individual to demographic, social, psychological, economic, medical system-related, and environmental factors [3]. A variety of quantitative and qualitative data are thus required to identify health risk factors that affect disease, death, and health care utilization and to correlate causes and consequences. Clearly, realizing this goal through a single dataset is not easy. We must thus take on a more comprehensive approach with the data on health and medical care. While most of these sources have been conducted by national surveying and reporting systems, there are also medical records by hospitals and clinics for administrative purposes as well as major disease registration surveillance systems organized by specialized academic societies for research purposes. For example, there are national data on death reports, data on the output of administrative procedures such as health insurance, data from interview surveys on health, and registration surveillance systems of major diseases such as cancer, stroke, and myocardial infarction.

Since each resource was developed to collect data for a particular purpose, the scope of each purpose has sufficient value. However, the analysis of each dataset has limitations in the cause-effect approach and the comprehensive analysis of health problems that controls potential influences. As a way to overcome these limitations, a strategy of linking and integrating different kinds of healthcare databases is important in that it can expand the potential value of the data being used.

In this context, countries such as the United States, Canada, the United Kingdom, Australia, and New Zealand are attempting to link different data within their country in order to maximize the value of their data sets, developing mapping algorithms to increase the probability of linkage. For

example, the National Center for Health Statistics (NCHS) in the United States developed a program to link various external data sources in an effort to maximize the statistical value of population-based survey data held from the Center for Disease Control (CDC). The data sources included death records from the National Death Index (NDI), Medicare/Medicaid registration and medical billing data from the Centers for Medicare and Medicaid Services (CMS), social security data from the Social Security Administration (SSA), End Stage Renal Disease (ESRD) data from the United States Renal Data System (USRDS), and housing data from the Department of Housing and Urban Develoment (HUD) [4]. These linked databases have allowed researchers to study major healthcare issues more effectively and is thus far the only population-based program to combine information on healthcare with social and demographic information in epidemiological and healthcare service research. Outside of the US, organizations in other countries such as Statistics Canada in Canada, the Administrative Data Research Network (ADRN) and the Administrative Data Liaison Service (ADLS) in the UK, the Center for Health Record Linkage (CHeReL) in Australia, and Statistics NZ in New Zealand have also taken active steps to link their data [5-8].

While data linkage has become more common, this strategy is not without its challenges [2]. For example, data linkage relies not only on technology and statistics but also on infrastructures such as laws and regulations. In the NCHS in the Unite States, the Public Health Service Act [9], Systems of Record Notice (SORN) [10], and other provisions were established to protect the privacy and confidentiality of survey respondents. Information provided under a confidentiality pledge is used only for statistical purposes, and information disclosed without such a pledge is also used to ensure a high degree of confidentiality and to take account of confidential information protection and statistical efficiency. Linkage should be used only if there are higher public outcomes than the potential privacy or confidentiality of the individuals being linked.

In December 2016, the Korean government launched a 'Big Data Task Force' to support the proactive utilization of big data [11]. Since the government takes direct advantage of big data and the convergence of public and private data, it may be able to build a successful business model in the future by further enhancing the value of individual data and utilizing advanced data analysis. Such a data linkage strategy is expected to be utilized in various industrial fields in the future. In the healthcare field, analyses using data linkage and integration are expected to maximize the potential of data and allow for the discovery of infinite practical knowledge.

References

1. Dong XL, Srivastava D. Big data integration. Proc VLDB Endow 2013;6(11):1188-9.
2. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data Soc 2017;4(2):2053951717745678.
3. Australian Institute of Health and Welfare. Evidence for chronic disease risk factors [Internet]. Canberra, Australia: Australian Institute of Health and Welfare; 2016 [cited at 2018 Oct 19]. Available from: https://www.aihw.gov.au/reports/chronic-disease/evidence-for-chronic-disease-risk-factors/summary.
4. National Center for Health Statistics. Data linkages resources by NCHS health surveys [Internet]. Atlanta (GA): CDC; c2017 [cited at 2018 Oct 19]. Available from: https://www.cdc.gov/nchs/data/datalinkage/LinkageTable.pdf.
5. Statistics Canada. Social Data Linkage Environment (SDLE) [Internet]. Ottawa, Canada: Statistics Canada; c2017 [cited at 2018 Oct 19]. Available from: https://www.statcan.gc.ca/eng/sdle/index.
6. Elias P. The UK administrative data research network: its genesis, progress, and future. Ann Am Acad Pol Soc Sci 2018;675(1):184-201.
7. Center for Health Record Linkage. How record linkage works [Internet]. North Sydney, Australia: Center for Health Record Linkage; c2018 [cited at 2018 Oct 19]. Available from: http://www.cherel.org.au/how-record-linkage-works.
8. Statistics New Zealand. Linking methodology used by Statistics New Zealand in the integrated data infrastructure project [Internet]. Wellington, New Zealand: Statistics New Zealand; c2014 [cited at 2018 Oct 19]. Available from: http://archive.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/idi-resources/linking-methodology-statsnz-idi.aspx.
9. National Center for Health Statistics. Public Health Service Act [Internet]. Atlanta (GA): Centers for Disease Control and Prevention; c2018 [cited at 2018 Oct 19]. Available from: https://www.cdc.gov/nchs/training/confidentiality/training/page600.html.
10. US Department of Homeland Security. System of Records Notices (SORNs) [Internet]. Washington (DC): US Department of Homeland Security; c2017 [cited at 2018 Oct 19]. Available from: https://www.dhs.gov/system-records-notices-sorns.
11. Park SW. Big data era and data integration. KISDI Res Rep 2018;30(1):1-24.

 https://doi.org/10.4258/hir.2018.24.4.251