Using "big data" to improve health care services and research

Jeff A Dennis PhD.

Abstract

The growth of large databases of health information has accelerated substantially as computers are able to store and process these data with increasing efficiency. Analysis of this growing cache of data has the potential to aid health care providers, improve cost and efficiency, and inform public health policies, yet caution must be exercised in the interpretation of results. Large databases cannot be assumed to be representative of the population from which they are derived simply due to their size, and as with all research, sampling and data quality issues must be carefully considered throughout the research process. This article discusses advances in large database availability and research, including new avenues that the data have opened, as well as potential problems that may arise when results are not appropriately evaluated.

Key words: data sources, data linkage, population statistics, demography, vital statistics, national health and nutrition examination survey

Large databases represent a rapidly growing area of research and analytics, particularly with the increasing propensity for our daily activities, interactions, and transactions to leave digital footprints. Often termed "big data," these databases vary widely in scope, size, and quality, such that no single approach can be used to derive meaningful results from them. In health outcomes and quality improvement research, large databases may be produced via hospital and clinic records, regional or national surveys, vital statistics collection systems, randomized trials, insurance company claims, and myriad other sources, including what health information people search for online. These data can be applied to many analytical areas, including understanding population health outcomes, monitoring costs and efficiency, and identifying risks in individuals or groups of patients. This article discusses uses for large databases, potential linkages with other data, quality issues, representativeness, and potential issues with statistical power that arise with a large sample size.

Corresponding author: Jeff A Dennis PhD
Contact Information: jeff.dennis@ttuhsc.edu
DOI: 10.12746/swrccc2016.0413.177

Potential applications of big data in health care administration, practice, and research are widespread. Hospital or clinic data may be analyzed to identify high risk/high cost patients and work to improve their care and cost efficiency in the health care system.[1] Finding the place of big data as it relates to clinical practice and health care decision-making remains complex, as analytics may identify important individual or group trends impacting patients, yet aggregate analysis remains an imperfect estimator of individual risk and patient differences. As such, big data analysis must evolve to inform clinicians effectively without overreaching its predictive power in any individual case. Similarly, predictive methods must be refined to inform clinicians of relevant points of risk in a clear manner.[2] Among the more publicized big data

findings in recent years, Google Flu Trends estimates influenza prevalence via analysis of internet searches. Yet it has been shown to overestimate prevalence in some years and miss new strains in other years.[3] In short, data availability and computing power may currently outpace our ability to analyze and interpret output with accuracy.

Large databases existed long before the term "big data" entered common vernacular, but advances in computer processing speed and storage in recent years have created new possibilities for quantitative analysis of many different research areas, with health research representing a major area of interest. Further , innovative researchers merge existing databases with additional data points or datasets to improve descriptive power and add context. As an example, the National Center for Health Statistics collects and maintains a large number of health surveys, and although most of these databases are cross sectional in nature, a number of them can be linked to later mortality via the National Death Index, as well as to Medicare, Social Security, and Census data, for neighborhood contextual information.[4,5] Although more sensitive individual data may require special permissions, the possibilities for linking social and environmental data to existing detailed health surveys continue to grow.

Substantial differences exist between clinic data retained for medical care or billing and survey data collected and/or maintained by large national entities such as the Centers for Disease Control (CDC), primarily intended for research or public health purposes. Whereas many of the large national surveys are carefully planned, collected, coded, and "cleaned" by coordinated teams, hospitals and clinics may record information in different systems, and not in an entirely consistent manner. Whatever the source, whether the CDC, a hospital, or an insurance company, one must carefully evaluate what the data contain, as well as why and how it was collected, to understand the validity of various measures, as well as the consistency of coding. Analytical costs for these data differ widely, as structured data, such as the CDC datasets, may be analyzed with a variety of existing and user friendly statistical software packages. However, less structured or unstructured data may require more specialized (and thus costlier) computer science expertise to extract data and set it up for meaningful analysis.[6,7]

Related to health surveys, vital statistics data typically represent the full population for a given outcome or event, rather than more commonly used random sampling. For example, the U.S. Birth Data File contains information collected at birth for about 4 million births per year. These data are useful in their coverage of the population, greatly reducing the potential for error in the estimation of various outcomes. However, the breadth of available variables in this type of data remains reasonably narrow, given that data collected from birth certificates do not include many variables relevant to clinical or social/behavioral research.

Large databases provide the opportunity to explore a wide array of topics, and particularly are often well suited to comparisons of rarer groups. Whereas a prospective study of 100 individuals in a sleep lab may not have sufficient numbers to compare age and gender differences in sleep disorders, the National Health and Nutrition Examination Study (NHANES) provides detailed sleep questionnaire data for a nationally representative sample of nearly 20,000 individuals between the years 2005-2008, combined with detailed behavioral and demographic information.[8] The sleep lab study allows for careful control and monitoring of specific treatments or outcomes, yet generalizations about the broader population may be difficult. Alternatively, NHANES provides a large, nationally representative sample for comparisons of various demographic groups, but with the limitation that data have already been collected so researchers cannot alter questionnaires or protocol. In some cases, limited data linkages exist, with mortality follow-up or Census tract information, but individualized follow-up is generally not possible. More important, all large datasets are not automatically representative of any given population, including those populations

that appear to be represented in the data, and must be regarded as such. Kaiser Permanente and the Veteran's Administration, for example, maintain very large databases of patient information, which are likely representative of the populations they serve, but not necessarily of the U.S. population more broadly.[9]

Although large databases provide sample sizes that allow for comparisons of many subgroups, these analyses must carefully evaluate standards of statistical significance. Confidence intervals for statistical testing become smaller as sample size increases, and with very large datasets, p-values will often be low, such that the null hypothesis may be rejected even for very small absolute differences.[10] Therefore, researchers are responsible for understanding and interpreting effect sizes beyond mere statistical significance. Trends or disparities that show statistical significance in large datasets may highlight subtle changes that, upon evaluation, have little relevance in practice. Relative differences may often be reported in findings from large datasets, yet these comparisons can obscure effect sizes that are small and inconsequential. For example, where a hypothetical disease incidence changes from 6% to 8% in a large population, the relative increase in incidence is 33.3%, despite the fact that absolute change is only 2%; the latter raises far less alarm than the former. A 2% difference can be highly relevant, however, and as such, the challenge is that data must be interpreted for relevance beyond statistical testing to understand the importance of changes and trends.

In conclusion, health data are being collected, aggregated, and stored at unprecedented rates, and the future holds many possibilities for how this data may be used to improve health care systems and population health, respectively. As our actions become increasingly digitized, down to personal devices that record our heart rate, exercise, sleep patterns, and other variables on a continual basis, we must work to improve our capacity to analyze these data effectively and use results in the appropriate manner. With seemingly limitless possibilities in the growth of large databases, we must be critical and discerning in how we collect and interpret this data, as the sheer magnitude lends itself to both intentional and unintentional misuse.

Author affiliation : Jeff A Dennis is a statistician in the Department of Internal Medicine at Texas Tech University Health Sciences Center in Lubbock, TX.
Submitted: 11/24/2015
Accepted: 12/8/2015
Conflict of Interest Disclosures: None
Published electronically: 1/15/2016

References

1. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar, G. Big data in health care: using analytics to identify high risk and high-cost patients. Health Affairs 2014; 33:1123-1131.
2. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013. 309: 1351-1352.
3. Lazer D, Kennedy R, King G, Vespignani A. The parable of google flu: traps in big data analysis. Science 2014; 343:1203-1205.
4. http://www.cdc.gov/nchs/nhis/nhis_products.htm
5. http://rdc/geocodes/geowt_nhis.htm
6. Jee K, Kim GH. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Health Inform Res 2013; 19:79-85.
7. Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. JMIR Med Inform 2014; 2(1):1-11.
8. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2005-2008. http://www.cdc.gov/nchs/nhanes.htm.
9. Crump C, Sundquist K, Winkleby MA. Transnational research partnerships: leveraging big data to enhance U.S. health. J Epidemiol Community Health 2015; Online first 3 Mar 2015.
10. Granger CWJ. Extracting information from mega-panels and high-frequency data. Statistica Neerlandica 1998; 52: 258-272.