Potential and Pitfalls for Big Data in Health Research

Karim S. Ladha, MD, MS[a], Vishal S. Arora, AB[b,c],
Richard P. Dutton, MD, MBA[d,e], Joseph A. Hyder, MD, PhD[f,*]

[a]Department of Anesthesia, Pain and Critical Care Medicine, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, MA 02114, USA; [b]Department of Public Health and Policy, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK; [c]European Observatory on Health Systems and Policies, Houghton Street, London WC2A 2AE, UK; [d]Anesthesia Quality Institute, 1061 American Lane, Schaumburg, IL 60173, USA; [e]Department of Anesthesiology, University of Chicago, 5841 S Maryland Avenue, Chicago, IL 60637, USA; [f]Department of Anesthesiology, Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, USA

Key points

- Studies using large observational data sets can provide valuable clinical insights with speed and cost efficiency not possible with clinical trials.
- Findings from studies using such data sets must be interpreted with a skepticism and attention to methodological details not often required when data are acquired intentionally, as with a prospective observational study or trial design.
- Methods to analyze these databases are evolving, and anesthesiologists should familiarize themselves with these techniques in order to draw appropriate conclusions from the results of these studies.

INTRODUCTION

The term big data is a ubiquitous catch phrase to describe large databases that are not easily analyzed with traditional statistical approaches. In health care, the adoption of electronic medical and billing records has allowed researchers to perform analyses on thousands, even millions, of patients. Such studies are increasingly encountered in the literature, but their advantages and disadvantages compared with traditional observational studies and randomized trials are not always well understood [1,2]. Readers may feel compelled to accept findings based on large cohorts or might be confused by unfamiliar methods

---

*Corresponding author. *E-mail address*: Hyder.Joseph@mayo.edu

used to analyze these data. This article addresses the promises and perils of big data with examples from perioperative care. First it describes and explains big data as existing data rather than prospectively ascertained data. It then reviews the major challenge of big data studies with attention to determining causal inference in nonexperimental study designs and the key important differences between observational studies and randomized trials. Third, it explains how statistical methods are used to address these limitations. In addition, it presents some potential uses of big data in clinical research.

BIG DATA, SECONDARY DATA SOURCES, AND THE IMPORTANCE OF ASCERTAINMENT

Big data used in clinical and health policy research are secondary data sources, or what are referred to in this article as found data. These data often have been collected for a purpose other than scientific investigations and can include the following sources: administrative, or billing data; electronic medical records; data gathered for quality improvement; insurance claims data; the intraoperative anesthesia record; or a combination of all of these data sources.

ADVANTAGES OF EXISTING DATA

Using existing data has several advantages for research. Compared with prospective, intentional data collection, the costs of data collection are low, and using found data permits the rapid assembly of very large sample sizes, which, in theory, permits the study of rare diseases and the power to detect small effect sizes. In addition, the removal of identifiable patient information means that there are few, if any, risks to the patients who contributed data, so patients are potentially protected from the risks of participation in research and delays caused by approval for human subjects research are minimal.

In addition, use of big data may mean a potential democratization of clinical investigation. Historically, disease-specific clinical investigations were limited to costly and lengthy prospective studies, such as the Framingham Study or the Rancho Bernardo Study [3,4]. For rare diseases or treatments, only a few large groups with high-volume practices, such as the Mayo Clinic, could evaluate clinical outcomes.

Perhaps most important is that big data allow researchers to create research hypotheses while taking into account comorbidities and disease patterns that are reflective of everyday medical practice. In contrast, most current randomized controlled trials often can reasonably test only a single intervention and limited number of outcomes in a group of patients who are less medically diverse than the population average.

WHAT CANNOT BE FOUND IN FOUND DATA

Nonetheless, large data sets carry limitations when used for clinical inquiry or health services research. Most notably, such data are not specifically ascertained for research purposes. In prospective observational studies such as the Framingham Study, variables of interest were specified before data collection. Such studies