AMERICAN ACADEMY
OF OPHTHALMOLOGY®

Editorial

When Does Size Matter? Promises, Pitfalls, and Appropriate Interpretation of "Big" Medical Records Data

Kathryn Rough, ScD - *Mountain View, California*
John T. Thompson, MD - *Baltimore, Maryland*

The analysis of large datasets in ophthalmic research, and medical research more broadly, has become increasingly common. The growth of electronic health records (EHRs) has facilitated passive collection of large quantities of computerized medical data; EHR systems have been adopted nearly universally in the United States.[1,2] Researchers are eager to leverage these data into insights that meaningfully improve clinical care and patient outcomes. The American Academy of Ophthalmology's Intelligent Research in Sight (IRIS) Registry represents one such dataset.

In this issue, Chiang et al[3] (https://www.aaojournal.org/article/S0161-6420(17)32703-3/fulltext) outline the creation of the IRIS Registry and describe the basic characteristics of the physicians and patients included. The IRIS Registry contains an impressive amount of data from ophthalmologic practices; in 2016 alone, IRIS aggregated data on 37 million encounters from 17 million unique patients across 10 000 providers. The utility of the IRIS Registry for clinical research is demonstrated in a study by Park and Lum,[4] (p. https://www.aaojournal.org/article/S0161-6420(17)33178-0/fulltext) also appearing in this issue. The authors describe the 1-year risk of return to the operating room after vitrectomy, with a sample of more than 73 000 eyes with epiretinal membrane and 41 000 eyes with macular holes.[4] This investigation is an example of studies the IRIS Registry may enable, and we expect many others will soon follow.

Yet enthusiasm about the impressive size of large medical datasets should not diminish awareness of their weaknesses; as with all research, it is essential to draw careful conclusions that are well supported by the data. This editorial intends to give an overview of big EHR data, review their potential strengths for medical research, and outline 5 common pitfalls with recommendations on how they can be mitigated.

Big Data Basics

We can think of most "big" datasets as having 2 dimensions: width and length. Wide datasets have information on a large number of variables (e.g., genetic sequence data), and long datasets contain information on a large number of people (e.g., census data). Databases from EHR systems, including the IRIS Registry, contain information on both a large number of variables and a large number of people.

Medical studies using observational data will generally have at least 1 of 3 goals: (1) description, (2) prediction, or (3) causal inference. Descriptive studies yield insight into the patient experience by reporting on the prevalence and incidence of conditions and treatments. The studies by Chiang et al[3] and Park and Lum[4] featured in this issue are primarily descriptive in nature. Predictive studies provide estimates of a patient's risk for an outcome, given their individual characteristics, without any intention to intervene. Examples include risk scores, such as the Framingham Risk Score.[5] Finally, studies of causal inference aim to quantify the impact of an intervention, for example, estimating improvement in visual acuity after treating diabetic macular edema to determine efficacy of different drug treatments.[6] The specific context and goals of studies will dictate which strengths and limitations of large EHR data will be most pertinent.

Promises of Big Electronic Health Record Data

The IRIS Registry aggregates data collected during routine provision of clinical care, making it possible to gain insight from information that would otherwise be inaccessible to researchers from more conventional studies. Although deployment of EHR systems and registries like IRIS require a substantial initial investment of resources to create, they are scalable once in place; the marginal cost and effort required for the addition of new information is low. Compare this with the more traditional paradigm for medical research, where study investigators specify predefined clinical data to be collected outside the course of usual care. The costs (including time, effort, and money) of collecting additional measurements or enrolling new patients prevents collection of original research data on the same extensive scale as EHR systems enable.

As a result, investigators can leverage large, passively collected datasets to perform studies that would be otherwise infeasible; the broad range of clinical variables available creates numerous possibilities for potential analyses. The value of "real-world evidence" from passively collected datasets is being increasingly recognized by institutions including the US Food and Drug Administration[7] and the National Academy of Sciences, Engineering, and Medicine.[8] Research can be performed relatively rapidly, without the delays for data collection. Another tangible advantage of the large size of the IRIS Registry and similar EHR data is the increased absolute number of patients with rare outcomes or in uncommon subgroups; large databases have previously been used to study rare

https://doi.org/10.1016/j.ophtha.2018.04.034
ISSN 0161-6420/18

events, such as infectious endophthalmitis after cataract surgery[9] or systemic complications after anti–vascular endothelial growth factor therapy.[10]

Five Potential Pitfalls of Big Electronic Health Record Data

Data quality: The efficiency of passive collection of research data from EHRs comes with trade-offs in data quality; data errors can arise from several sources.[11] For databases that aggregate EHR data from multiple sources, including the IRIS Registry, errors may be introduced from the extraction and harmonization of information across systems that store medical data in substantially different formats. Clinicians also may make data-entry errors in the EHR interface or fail to report important information in structured fields, opting to write free-text notes instead. After all, the primary motivation for clinician EHR use is not research, but the support of clinical workflow, satisfaction of administrative requirements, and documentation for reimbursement. As a result, important types of information may not be captured or may be imprecisely measured in EHRs. Ambiguity in the available data often leads researchers to use inclusion criteria, exposure measures, case definitions, and outcomes that are inexact. For instance, Park and Lum[4] identify patients receiving vitrectomy for epiretinal membrane and macular holes using a combination of International Classification of Diseases 9th and 10th Revisions, and Current Procedural Terminology codes. Assumptions were made to prioritize conditions; for a hypothetical patient with diagnostic codes indicating the presence of epiretinal membrane, vitreous floaters, and macular hole, macular hole was labeled as the primary diagnosis. Furthermore, the authors were not able to identify what percentage of macular holes actually closed, only the percentage of patients with macular hole who returned to the operating room.

Although there is little researchers can do to change the quality of EHR data, definitions for key variables, especially outcomes, should be validated whenever possible. Validation compares an inexact definition (e.g., using presence or absence of diagnostic codes) with a gold standard (e.g., full chart review) and reports metrics such as the sensitivity, specificity, positive predictive value, and negative predictive value. Performance can vary widely on the basis of the condition and the sophistication of the definition; one study found diagnostic code–based definitions for 32 conditions yielded positive predicted values ranging from 23% to 100%.[12] Researchers may cite studies that have previously validated the variable definition they use or validation may be performed internally on a subset of their data.

Patient loss to follow-up: Electronic health record databases often do not contain a complete record of a patient's interactions with the medical system. Single-center EHR data, as well as multicenter registries like IRIS, do not capture clinical interactions occurring at nonparticipating facilities. This may lead to substantial undercounting of clinical events after a treatment or procedure. As a result, descriptive studies may report artificially low rates of complications—an issue that should be seriously considered before establishing performance-based quality of care metrics exclusively on registries and other datasets with less than complete follow-up, regardless of their size. In studies estimating causal effects, this can lead to selection bias, particularly when reasons for loss to follow-up are associated with both the exposure and the outcome.[13]

The EHR data may not be ideal for answering questions that require a long, uninterrupted duration of individual patient follow-up. Studies that depend on information collected across patient encounters should report potential loss to follow-up as a limitation. To test the robustness of conclusions to this limitation, authors may consider conducting quantitative bias analyses.[14]

Overemphasis on statistical significance: There are also several important statistical considerations when working with any large database. A by-product of having many patient observations is increased statistical power and precision; this can lead to remarkably small *P* values for hypothesis testing, even when the observed differences between groups are clinically inconsequential.[15] Simply put, statistical significance does not necessarily imply clinical significance. Furthermore, large sample sizes can have the unfortunate side effect of magnifying problems related to statistical model misspecification. Incorrect model assumptions will lead to biased estimates; the narrower the confidence intervals, the more likely the true parameter will be excluded.[16] In addition, the high-dimensionality of EHR data can lead investigators to test an excessive number of hypotheses or manipulate the data in a number of arbitrary ways until arriving at the desired outcome, typically only reporting results that are statistically significant.[17] These practices, sometimes called "p-hacking" or "data-dredging," lead to an inflated risk of false-positive results (also known as "type I errors") and findings that are not reproducible.

To ensure that the audience can fairly judge whether results are meaningful enough to warrant action, *P* values must not be presented in isolation. Authors should report meaningful effect estimates, such as the absolute difference or risk ratio, with corresponding confidence intervals. To prevent misleading results from p-hacking practices, primary analyses can be prespecified. If post hoc testing is performed, the total number of tests and description of the tests should be reported, regardless of whether results were statistically significant.

Confounding: Causal inference studies using big EHR data are vulnerable to the same confounding biases common in other observational studies. Confounding occurs when outcomes and treatment decisions share common causes. For instance, confounding by severity occurs when mild cases of a condition are treated differently than severe cases and have better outcomes, independent of treatment received. A naïve comparison of outcomes for the 2 treatment alternatives will produce misleading results. Investigators have demonstrated that influenza vaccination appears to increase the risk of influenza-related complications when the study design and analysis do not account for the differences in age and pulmonary disease between vaccine recipients and nonrecipients.[18]

When designing observational studies for causal inference, it can be helpful to conceptualize how an ideal randomized trial would be conducted and to make study design choices to mimic the ideal trial.[19] To minimize bias due to confounding, factors that act as a common cause of the outcome and treatment must be accounted for through statistical adjustment, restriction, or matching. Although a complete overview of study design for causal inference is beyond the scope of this article, several textbooks provide detailed guidance.[19,20]

Appropriate reporting: Observational studies including research from large EHR datasets can have important implications for both clinical practice and broader health policy. Investigators using these datasets have a responsibility to be circumspect, thoughtfully considering the strengths and limitations of their approach, faithfully reporting their analytic methods, and drawing responsible, unexaggerated conclusions. All authors should prioritize clear reporting and, at minimum, follow the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines[21] for observational research and RECORD (Reporting of Studies Using Observational Routine-Collected Health Data) statement for routinely collected health data.[22]

Conclusions

Most of these limitations are not unique to studies using "big" data; yet, the size of the dataset alone will not compensate for them. Data from EHRs, including the IRIS Registry, will open the door to research investigations that would have been otherwise infeasible or impossible. As members of the medical research community continue to gain access to unprecedented amounts of data, we would be wise to remember that the use of "big" EHR data comes with both substantial promise and potential pitfalls.

References

1. Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2015. The Office of the National Coordinator for Health Information Technology Data Brief 35. Available at: https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php; May 2016. Accessed January 22, 2018.
2. Healthcare Information and Management Systems Society (HIMSS). Essentials Brief: 2016 Outpatient Practice Management and Electronic Health Record Solutions Study. July 2016; Available at https://www.himssanalytics.org/essentials-brief/essentials-brief-2017-outpatient-pm-ehr-study.
3. Chiang MF, Sommer A, Rich WL, et al. The 2016 IRIS Registry (Intelligent Research in Sight) Database: characteristics and methods. *Ophthalmology. 2018*; 2018 Jan 13. pii: S0161-6420(17)32703-3 https://doi.org/10.1016/j.ophtha.2017.12.001.
4. Park DW, Lum F. Return to the operating room after macular surgery. IRIS Registry analysis. *Ophthalmology*. 2018 Feb 14. pii: S0161-6420(17) 33178-0.
5. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117:743−753.
6. Diabetic Retinopathy Clinical Research Network, Well JA, Glassman AR, et al. Aflibercept, bevacizumab or ranibizumab for diabetic macular edema. *N Engl J Med*. 2015;372: 1194−1203.
7. US Food and Drug Administration. Use of real-world evidence to support regulatory decision-making for medical devices: guidance for industry and Food and Drug Administration Staff. Available at: https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm513027.pdf; 2017. Accessed February 23, 2018.
8. National Academy of Sciences, Engineering, and Medicine. *Examining the Impact of Real-World Evidence on Medical Product Development*. Washington DC: National Academies Press; 2018. Available at: https://www.ncbi.nlm.nih.gov/books/NBK481618/pdf/Bookshelf_NBK481618.pdf. Accessed February 23, 2018.
9. Haripriya A, Chang DF, Ravindran RD. Endophthalmitis reduction with intracameral moxifloxacin prophylaxis. *Ophthalmology*. 2017;124:768−775.
10. Moja L, Lucenteforte E, Kwag KH, et al. Systemic safety of bevacizumab versus ranibizumab for neovascular age-related macular degeneration. *Cochrane Database Syst Rev*. 2014;9, CD011230.
11. Bowman S. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspect Health Inf Manag*. 2013;10(1c). eCollection 2013.
12. Quan H, Li B, Saunders LD, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Serv Res*. 2008;43:1424−1441.
13. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615−625.
14. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York: Springer; 2009.
15. American Statistical Association. The ASA's statement on statistical significance and p-values. *Am Stat*. 2016;70: 129−133.
16. van der Laan M. Statistics as a science, not an art: the way to survive in data science. AMSTAT News 2015. Available at: http://magazine.amstat.org/blog/2015/02/01/statscience_feb2015/. Accessed February 22, 2018.
17. Nuzzo R. Scientific method: statistical errors. *Nature*. 2014;506:150−152.
18. Hak E, Verheij TJM, Nichol KL, Hoes AW. Confounding by indication in non-experimental evaluation of vaccine effectiveness: the example of prevention of influenza complications. *J Epidemiol Community Health*. 2002;56:951−955.
19. Hernan MA, Robins JM. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming. Available at: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. Accessed January 23, 2018.
20. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
21. von Elm E, Altman DG, Egger M, et al. STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335:806−808.
22. Benchimol EI, Smeeth L, Guttman A, et al. The Reporting of Studies Using Observational Routine-Collected Health Data (RECORD) Statement. *PLoS Med*. 2015;12:e1001885.

Footnotes and Financial Disclosures

Financial Disclosure(s): The author(s) have made the following disclosure(s): K.R.: Employed Google LLC for research work unrelated to the IRIS Registry. This work was done as part of the Google AI Residency program.

Correspondence:
John T. Thompson, MD, 6569 North Charles St., Suite 605, Baltimore, MD 21204. E-mail: Jthompson@retinaspec.com.