MS #16180

Prospective EHR-based clinical trials: The challenge of missing data

Hadi Kharrazi MHI, MD, PhD*, Chenguang Wang, PhD** and Daniel Scharfstein, ScD***
Johns Hopkins University
* Department of Health Policy and Management, Bloomberg School of Public Health
** Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center
*** Department of Biostatistics, Bloomberg School of Public Health

Corresponding Author:

Daniel Scharfstein, ScD

Johns Hopkins Bloomberg School of Public Health

615 North Wolfe Street

Baltimore, MD 21205

dscharf@jhu.edu

This discussion focuses on the challenges of using prospectively collected electronic health record (EHR) data as outcomes in clinical trials, with a particular emphasis on the issue of missing data. Our discussion is motivated by the article in this issue: '*Translating the Hemoglobin A1C with More Easily Understood Feedback: A Randomized Controlled Trial*' by Gopalan *et al*[1]. In the spirit of open science, the authors generously shared their study protocol, statistical analysis plan and analysis dataset. Using their dataset, we conducted analyses to help emphasize important statistical issues. This editorial should not be considered a criticism of their paper; rather, their study is used as a reference to expand on the challenges of missing data in EHRs and to provide suggestions for future studies.

The Rise of EHR and Its Use in Clinical Trials

The HITECH[2] act has empowered and incentivized healthcare providers to adopt EHRs. As a result, there has been a dramatic rise in EHR adoption. The adoption rate among office-based physician practices has increased from 18% in 2001 to 78% in 2013, and for hospitals it has increased from 10% to more than 80%.[3]

The increased adoption of EHRs among providers along with enhanced completeness of clinical data has led researchers to design clinical trials based on prospectively collected EHR-based outcome data that are collected as part of routine clinical practice. Key selling points of such trials are reduced operational demands, patient burden and costs. Further, some have argued that by not enforcing a data collection protocol apart from routine care, the trials will be more reflective of the real-world, a central tenet of "effectiveness" trials. The key tension is that, due to variability of adherence to practice guidelines in routine care, there are likely to be high levels of missing outcome data.

As a case in point, the Gopalan paper reports the results of a randomized controlled trial evaluating a health literacy intervention using baseline and 3 month (when available) hemoglobin A1C (HbA1C) values to improve the 6 month HbA1C for patients with uncontrolled diabetes. The HbA1C's were to be captured as part of routine clinical practice and recorded in EHRs. It is important to note that the ADA recommends HbA1C testing every three months for patients who are not meeting glycemic goals. Despite this recommendation and study reminders to "follow up with your primary care provider regularly and have your diabetes control monitored every 3 months", unexpectedly, 70% and 49% of patients were missing HbA1C values at 3 and 6 months, respectively.

Statistical Implications of Missing Data

While the beauty of randomized controlled trials is to probabilistically ensure that the treatment groups do not differ with respect to measured and unmeasured baseline prognostic factors that can confound study results, this advantage is offset by the presence of missing outcome data. This is because any analysis requires (1) untestable assumptions about the distribution of missing outcomes in relation to the distribution of observed outcomes and (2) testable assumptions about the distribution of the observed data[4]. Gopalan *et al.* make the following assumptions:

1. *Untestable*: within levels of treatment assignment and key baseline covariates, the distribution of HbA1C at 6 months is the same for patients with missing data and those with observed data
2. *Testable*: the conditional distribution of Hb1AC outcome at 6 months among those with observed data is normally distributed with mean depending linearly on treatment assignment and baseline covariates (no interactions) and a homoscedastic variance term; a test we conducted of the adequacy of their model failed (p=0.0035).

The authors used these assumptions to multiply impute five complete datasets. They then "used ANOVA to test for differences in A1C change among the groups in each imputed data set, and then combined the results using standard formulae." While their approach accounts for uncertainty due to missing data, it "buy[s] information with assumptions". [5]

An under-appreciated feature of their approach is that they are borrowing information from one treatment group to impute missing data in another treatment group. Their approach leads to imputation of treatment-specific HbA1C values that are outside the treatment-specific range of observed HbA1C values. The observed range in the standard of care [grade; face] arm was 6.9-13 [6.3-16; 6.5-14.3], with 11% [5%; 15%] and 3% [0%; 0%] of imputed values lower than the minimum and higher than the maximum, respectively. While it may be reasonable to believe that healthier patients (i.e., those with lower HbA1C values at 6 months) are less likely to follow ADA recommendations, there is absolutely no evidence in the observed data to support values outside the range of the observed data.

The bottom line is that statistical methods for imputing missing outcome data are *not* a panacea. Given the increased availability of software, it is tempting to think otherwise. The only answer is better study design.

Study Design Recommendations

It is essential that the study design allows researchers to draw reliable and robust inferences about the effect of treatment on the primary outcome in a hypothetical world in which there is no missing data. This can be achieved by (a) designing studies that minimize missing data, (b) retrieving outcome data for a random sample of patients with missing data, (c) randomizing a subset of patients to a more reliable outcome collection scheme, and (d) crafting endpoints that are less reliant on EHR data collected at patient encounters. The tension with (a)–(c) is that they involve patient engagement outside the naturalistic interactions between patient and provider. Some would then argue that the study is not reflective of the real world. The tension with (d) is whether such endpoints are considered clinically meaningful. In the context of the Gopalan study, an example of an outcome meeting the criteria in (d) would be whether or not the patient adhered to the ADA testing guideline.

Our view is that a "poison" must be picked. We do not believe that researchers can rely solely on unpredictable clinical encounters to collect primary outcome data on all patients. If a research study, be it a randomized or observational, is to rely on non-intervened EHR data collection, here are some suggestions:

1. *Consider local data collection routines:* Although various quality and clinical guidelines mandate certain data collection processes, using these guidelines to predict completeness of data in an EHR system can be misleading. Often localized clinical guidelines generate different data completeness rates [6]. In the Gopalan study, the researchers incorrectly assumed that, due to ADA guidelines, HbA1C would be tested at least once within a 6 month period.
2. *Learn from EHR data patterns:* Historical EHR data can provide valuable information about the potential missing rate of prospective EHR data[i]. In the Gopalan study, the researchers could have used the historical records to estimate the missing data rate of HbA1Cs among their eligible population and use these estimates to inform their study design.
3. *Use additional data sources:* Researchers can utilize additional data sources to compensate for missing data. One can utilize aggregated records of patients across a healthcare delivery system or use non-EHR data to capture data collected outside of the provider's network (e.g., insurance claims or Health Information Exchange data). The Gopalan study involved EHR data aggregated from three outpatient clinics, but did not use other sources of clinical data.
4. *Learn from similar studies:* Other studies have already revealed missingness rates of certain data elements in EHRs. For example, various studies have shown a considerable missing data rate for laboratory values in ambulatory settings· [7, 8, 9] One study has shown a high missing data rate but high accuracy for HbA1C data [10]; and a diabetes RCT has recognized and planned accordingly for missing HbA1C data. [11]

Take-Home Message

EHR adoption has skyrocketed. There is increased interest in conducting real world RCTs with prospectively collected EHR outcome data. There are major data completeness challenges in using prospectively collected EHR data, which cannot be solved by imputation alone. We have provided suggestions to improve study planning and design. We have not addressed other data quality issues such as data accuracy and timeliness, which can also have a major impact on inferences drawn from RCTs.

References

1. Gopalan A, Tahirovic E, Moss H, et al. Translating the hemoglobin A1C with more easily understood feedback: A Randomized Controlled Trial. *Journal of General Internal Medicine.* 2014 DOI: 101007/s11606-014-2810-4.

2. *Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA)* Feb 17 2009.

3. Hsiao C, Hing E. Use and characteristics of Electronic Health Record systems among office-based physician practices: United States, 2001–2013. *Centers for Disease Control and Prevention.* 2014. http://www.cdc.gov/nchs/data/databriefs/db143.pdf. Accessed Jan 20, 2014.

4. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *The New England Journal of Medicine.* Oct 2012;367(14):1355-1360.

5. Coombs CH. *A Theory of Data*. New York: John Wiley and Sons; 1964.
6. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: The nature of patient care information system-related errors. *Journal of American Medical Informatics Association*. 2004;11:104-112.
7. Wahls TL, Cram PM. The frequency of missed test results and associated treatment delays in a highly computerized health system. *BMC Family Practice*. May 2007;8(32):1-8.
8. Smith PC, Araya-Guerra R, Bublitz C, et al. Missing clinical information during primary care visits. *Journal of American Medical Association*. Feb 2005;293(5):565-571.
9. Köpcke1 F, Trinczek B, Majeed RW, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: A retrospective analysis of element presence. *BMC Medical Informatics and Decision Making*. Mar 2013;13(37).
10. Goulet JL, Erdos J, Kancir S, et al. Measuring performance directly using the Veterans Health Administration electronic medical record: A comparison with external peer review. *Medical Care*. Jan 2007;45(1):73-79.
11. Albu J, Sohler N, Matti-Orozco B, et al. Expansion of electronic health record-based screening, prevention, and management of diabetes in New York city. *Preventable Chronic Diseases*. Jan 2013;10(E13).