NIH PRAGMATIC TRIALS
COLLABORATORY
Rethinking Clinical Trials®

Key Issues in Extracting Usable Data from Electronic Health Records for Pragmatic Clinical Trials

Background

Missing data present a universal challenge in clinical research. In pragmatic clinical trials (PCTs), missing data can have an even bigger impact than usual, because patients taking part in PCTs are typically not followed as closely as in traditional clinical trials and may drop out of the study for various reasons, such as leaving the participating health plan or care facility. Moreover, in PCTs, outcome ascertainment and data collection are typically integrated with routine healthcare system data collection (e.g., electronic health records [EHRs]; administrative claims data) to reduce costs and provide access to large numbers of study participants.

Use of EHR and claims data in PCTs

Data gathered from EHRs and claims data are initially collected for non-research purposes and are therefore subject to various limitations. These include data lag, coding errors, and incomplete or inaccurate capture of healthcare conditions and medications. In addition, because the United States lacks a universal healthcare system and fully interoperable and interconnected medical record systems, a more dangerous type of missing data can occur when patients participating in a PCT receive care or medications from other facilities that are not participating in the trial. In such a scenario, missing data go undetected and are not accounted for in the statistical analysis.

Cluster dropout in pair-matched studies

In pair-matched cluster-randomized experiments, it is possible for an entire cluster to drop out of the study, potentially forcing analysts to discard information from the matched cluster and thus reducing the number of matched pairs by one [1]. Such an occurrence can have a large impact on effect estimation, especially in trials with a small number of large clusters.

Missing data mechanisms

There are three missing data mechanisms:

1. **Missing Completely At Random (MCAR)** [2]. In the case of MCAR, the probability of missingness does not depend on any covariates or the outcome, and the complete cases (subjects with complete data) constitute a representative sample of the study population.

2. **Missing At Random (MAR)** [3]. In the case of MAR, the probability of missingness does not depend on unobserved elements, conditional on the observed data.
3. **Missing Not At Random (MNAR)** or non-ignorable missing [4]. MNAR applies when neither MCAR nor MAR holds.

Under MAR, the joint likelihood of measurement process and of missing data process factorizes into two separate likelihoods with variation-independent parameter sets [5]. Multiple statistical methods have been developed to account for MAR, including the popular multiple imputation approach [6], the likelihood-based parametric approaches [7,8], and the inverse-probability-weighting and doubly-robust approaches [9-11].
Under MNAR, the parameters of interest typically cannot be identified using observed data unless additional unverifiable parametric assumptions are imposed. Therefore, sensitivity analysis is usually conducted to assess the robustness of the inference to the imposed unverifiable assumptions [4,12,13]. Challenges remain with regard to adapting these methods into the analyses of cluster-randomized experiments in order to account for clustering, stratification/matching, and missing data in a unified manner.

Next steps

There is no single statistical technique that can address all missing data issues. It is important to assess the causes of missing data and their corresponding mechanisms and then select the appropriate statistical techniques to handle missing data. More advanced methodological research is needed, especially in the case of the MNAR mechanism.

Resources

1. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. Stat Med 1996;15(11):1069–92. PMID: 8804140.
2. Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: John Wiley & Sons; 1987.
3. Rubin DB. Inference and missing data (with discussion). Biometrika 1976;63:581–92. doi: 10.1093/biomet/63.3.581. Available at: http://biomet.oxfordjournals.org/content/63/3/581. Accessed May 14, 2014.
4. Diggle P, Kenward MG. Informative drop-out in longitudinal data-analysis. J R Stat Soc Ser C Appl Stat 1994;4(1)3:49–93.
5. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc Ser B (Methodol) 1977;39(1): 1-22. Available at: http://links.jstor.org/sici?sici=0035-9246%281977%2939%3A1%3C1%3AMLFIDV%3E2.0.CO%3B2-Z. Accessed May 14, 2014.
6. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8:3–15. PMID: 10347857
7. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: A comparative review. J Am Stat Assoc 2005;100(469):332–46. Available at: http://www.jstor.org/stable/27590542. Accessed May 14, 2014. doi: 10.1198/016214504000001844.

8. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madr) 2009;18(1):1–43. PMID: 21218187. doi: 10.1007/s11749-009-0138-x. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3016756/. Accessed May 14, 2014.
9. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 1994;89:846–66. doi: 10.2307/2290910. Available at: http://www.jstor.org/stable/2290910. Accessed May 14, 2014.
10. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 1995;90(429):106–21. doi: 10.2307/2291134. Available at: http://www.jstor.org/stable/2291134. Accessed May 14, 2014.
11. Robins JM, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. Statist Sci 2007;22(4):544–59. Available at: https://projecteuclid.org/euclid.ss/1207580169. Accessed May 14, 2014.
12. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Stat Assoc 1999;94(448):1096–120. doi: 10.2307/2669930. Available at: http://www.jstor.org/stable/2669930. Accessed May 14, 2014.
13. Vansteelandt S, Rotnitzky A, Robins J. Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. Biometrika 2007;94(4):841–60. doi: 10.1093/biomet/asm070. Available at: http://biomet.oxfordjournals.org/content/94/4/841.abstract. Accessed May 14, 2014.