Identifying and Understanding Data Quality Issues in a Pediatric Distributed Research Network

Levon Utidjian, MD[1], Ritu Khare, PhD[1], Evanette Burrows, MS[1], Greg Schulte, MS[2], Kevin Murphy, BS[1], Sara Deakyne, MPH[2], Richard Hoyt, BS[3], Nandan Patibandla, MS[4], Byron J Ruth, BS[1], Aaron N Browne, BA[1], Megan Reynolds, BBA[3], Keith Marsolo, PhD[5], Michael G Kahn, MD, PhD[2], L. Charles Bailey, MD, PhD[1]

[1]Children's Hospital of Philadelphia, Philadelphia, PA; [2]Children's Hospital Colorado, Aurora, CO; [3]Nationwide Children's Hospital, Columbus, OH; [4]Boston Children's Hospital, Boston, MA; [5]Cincinnati Children's Hospital Medical Center, Cincinnati, OH

Abstract

A prerequisite for building a clinical data research network is that the quality of the aggregated data be well-understood. As part of a newly-formed EHR-based pediatric research network, a set of systematic data quality checks were implemented and executed on the data. This study contributes by providing a detailed account of the types and sources of encountered issues, and a longitudinal distribution of issues across the partner sites in the network.

Introduction

Collaborations across multiple institutions are very essential to achieve adequate cohort sizes in pediatrics research[1]. PEDSnet is a newly established clinical data research network (CDRN) that aggregates electronic health record (EHR) data from eight of the nation's largest children's hospitals[2,3]. With the ultimate goal of supporting a variety of comparative effectiveness research, a prerequisite in PEDSnet is to ensure that the network's data is "high quality." The prominent challenges include the lack of EHR data's fitness for immediate research use, semantic heterogeneity across systems, and data peculiarities in pediatrics[1,4]. While previous studies have presented frameworks and techniques to validate the EHR-derived data, the process of data quality assessment in CDRNs continues to remain "behind the scenes" with no published empirical results[4,5,6]. In this study, we implement a comprehensive set of validity checks in PEDSnet to identify, understand, and report a range of data quality issues (see Table 1)[4,6].

Methods

The PEDSnet network uses the OMOP Common Data Model (CDM), a widely accepted schema for observational medical data[7]. Each partner site prepared an instance of the CDM by performing the extract-transform-load (ETL) operations from its EHR according to network-wide conventions. In this study, we focused on attribute-level data quality assessment, and developed data analysis scripts to ensure adherence to the CDM, perform data domain checks, and compute frequency distributions[5,6]. The output report comprises a visual summary (e.g. bar graphs) and a list of automatically detected data quality issues (e.g. out-of-range values) for each attribute. We executed the scripts on each site's data, and reviewed the graphs to identify additional issues (e.g. unusual shape or peaks). Next, we classified each issue as an "ETL issue" that could be resolved by fixing the ETL logic or a "provenance issue" that exists due to an anomaly, data characteristic, or an error in the EHR[8,9]. Finally, each issue was communicated to the originating site, which was responsible for validating the cause of the issue and resolving the issue.

Results

At the current stage of this project, we have collected data representing 4.6 million children and over 90 million encounters. Table 1 summarizes the total number of issues by the various data quality dimensions. The fidelity dimension corresponded to the cases where the distribution of the EHR values did not match with that in the CDM, due to ETL errors or due to data characteristics, such as differences in granularities between the coding systems used in the source (e.g. ICD-9) and the CDM (e.g. SNOMED CT). The consistency dimension included the cases where the attribute values were not aligned with the conventions (ETL issue), or were abnormal, e.g. a gestational age value greater than 42 weeks (provenance issue). The accuracy dimension included discrepancies between sites, e.g. significant differences in variation of body weights. Finally, the feasibility dimension corresponded to missing data in the EHR or incomplete ETL mappings.

Table 1. Data Quality Dimensions in PEDSnet and Issues Identified to Date

Dimension	Definition	Number of Issues
Fidelity (i.e. reliability)	the degree to which PEDSnet data correctly reflects source systems data	26
Consistency (i.e. internal validity)	the degree to which a specific type of information is recorded in the same way in the different data sources contributing to PEDSnet data	357
Accuracy (i.e. external validity)	the degree to which PEDSnet data accurately reflects the clinical characteristics of patients	11
Completeness (i.e. feasibility)	the degree to which a given type of information is actually collected and available in PEDSnet	192

Conclusion

A key challenge in building a CDRN is to define and achieve an optimal degree of data quality. In this study, we have conducted a data quality assessment of the PEDSnet network using rigorous data checks and manual reviews of statistical plots. While this study only focused on attribute level analysis, we have learned several important lessons. In spite of defining network-wide conventions, hosting a shared repository of ETL scripts, and organizing regular web conferences across sites, we identified 586 quality issues across eight sites[8]. This strongly suggests that proactive project management and documentation are not sufficient to ensure data validity in a CDRN. We found that nearly 20% of the identified issues were ETL mistakes, and despite the diversity of experiences and backgrounds of various teams, the site-wise contributions to these issues were significantly similar. This reinforces that a formal data quality assessment process is critical in building a CDRN.

References

1. Bailey LC, Milov DE, Kelleher K, Kahn MG, Del Beccaro M, Yu F, et al. Multi-Institutional Sharing of Electronic Health Record Data to Assess Childhood Obesity. PLoS ONE. 2013;8(6):e66192.
2. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. Journal of the American Medical Informatics Association. 2014 Jul;21(4):602–6.
3. Forrest CB, Margolis P, Seid M, Colletti RB. PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. Health Aff; 2014 Jul;33(7):1171–7.
4. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association. 2013 Jan 1;20(1):144–51.
5. Kahn MG, Raebel MA, Glanz JM, Riedlinger K. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;50:S21–9.
6. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013 Aug;51(8 Suppl 3):S22–9.
7. Common Data Model. Observational Medical Outcomes Partnership [Internet]. omop.org. [cited 2015 Mar 11]. Available from: http://omop.org/CDM
8. Browne AN, Pennington JW, Bailey C. Promoting Data Quality in a Clinical Data Research Network Using GitHub. AMIA Joint Summit on Clinical Research Informatics. 2015.
9. Brown AB, Patterson DA. To err is human. In Proceedings of the First Workshop on evaluating and architecting system dependability, EASY'01. July, 2001.