What Are Frequent Data Requests from Researchers? A Conceptual Model of Researchers' EHR Data Needs for Comparative Effectiveness Research

Gregory W. Hruby[1], MA, Praveen Chandar[1], PhD, Julia Hoxha[1], PhD, Eneida A. Mendonça[2,3], MD, PhD, David A. Hanauer[4,5], MD, MS, Chunhua Weng[1], PhD
[1]Dept. of Biomedical Informatics, Columbia University, New York, NY; [2]Dept. of Pediatrics; [3]Dept. of Biostatistics and Medical Informatics Univ. of Wisconsin, Madison, WI; [4]Dept. of Pediatrics, [5]School of Information, Univ. of Michigan, Ann Arbor, MI

Introduction

Data request forms are the key communication media linking medical researchers and informaticians.[1] The Carpenter framework serves as a representative mental model for how researchers conceptually organize information used for research.[2] Additionally, this model complements the Patient, Intervention, Control/Comparison, and Outcome nodes of the PICO framework, which is used for medical information retrieval.[3-5] We hypothesized that the semantic structural similarities between the two models suggest the Carpenter model may be well-suited as a standard template for data needs specification. As such, we choose the Carpenter model as a foundation for seeking a conceptual model to represent and organize common data needs of biomedical researchers.

Methods

We extended this work on the following two-fold hypotheses. The Carpenter framework (1) is a well-organized and comprehensive representation of medical concepts used in retrospective comparative effectiveness research (CER) for cancer, however (2) it can be expanded in scope to represent data needs for multiple medical domains. The leaf nodes of the model are used as an initial seed of codes for further processing and annotation of our three datasets: Clinical trial inclusion/exclusion criteria, EHR SQL project queries, and EHR data request logs. These data sources contain an intersection among the types of research requests they represent. Clinical trials dataset exclusively represent cohort identification criteria, whereas EHR SQL project queries almost exclusively represent the creation of complete datasets for retrospective CER, and EHR data request logs are representative of both request types: cohort identification and retrospective CER dataset generation. We parsed each dataset to the sentence/variable level. From the clinical trial inclusion/exclusion criteria and the EHR data request logs, we randomly selected 1,000 and 897 sentences, respectively, for annotation. From the SQL project files, we extracted 1,445 distinct variables for annotation. Each dataset was iteratively annotated with the Carpenter framework by one coder (GH). Each round saw the pruning of concepts from the Carpenter framework when the concept was not identified in the dataset, and the addition of concepts in the framework when new concepts in the data set were identified. Schematic representations of the datasets were produced based on the foundations of the Carpenter framework. The union of these schematic representations produced the proposed model.

Results

Figure 1 represents the combined data elements from the three datasets. Six original leaf nodes were pruned from the Carpenter framework. Many of these pruned abstract concepts, e.g. local disease burden, health norms, care process guidelines, and care systems and coordination, are not well represented in the EHR and, as such, many medical researchers may not specify these concepts in their data need request. 15 additional leaf nodes were added to the representation and are shown by an underlining of the concept within the figure. The Organizational/Provider Characteristics node was moved to the second column in the proposed model, as the data suggested it was more related to the Detection/Diagnostics and Intervention nodes rather than the Patient node.

Discussion

This study contributes an enhanced framework representing the common abstract medical concepts for clinical research. It can serve as both a guideline for medical researchers when specifying their EHR data needs, and a representation to informaticians for how medical researchers organize medical concepts used for CER. Informaticians can use this model to better anticipate the needs of researchers, which may aid in the successful provision of data to meet medical researchers' EHR data needs.

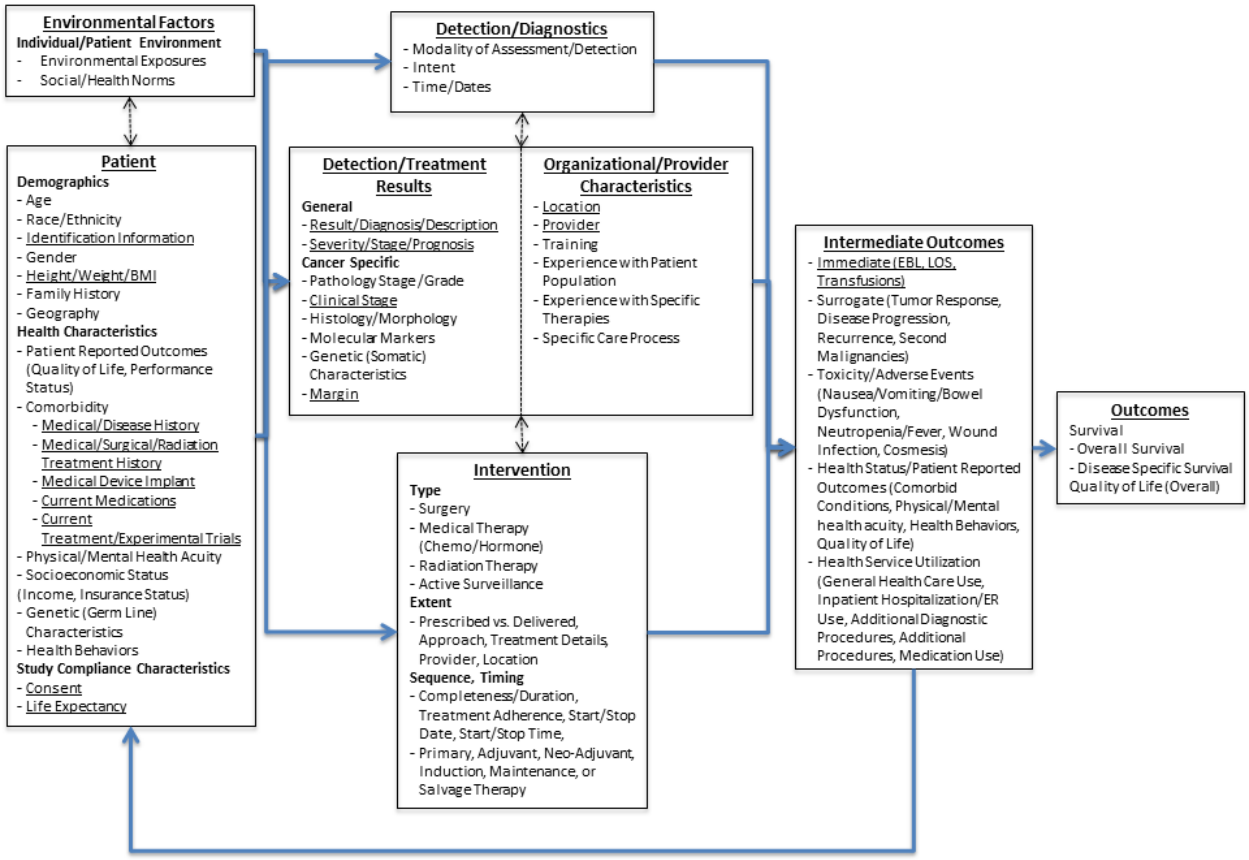

Figure 1. The Union of all three data needs schemas: Clinical Trial Inclusion/Exclusion criteria, EHR Data request logs, and EHR SQL project queries.

Acknowledgements

This study was sponsored by the U.S. National Library of Medicine grant **R01LM009886** (PI: Weng) and U.S. National Center for Advancing Translational Science grant **UL1 TR000040** (PI: Ginsberg).

References

1. Hanauer DA, Hruby GW, Fort DG, Rasmussen LV, Mendonça EA, Weng C. What Is Asked in Clinical Data Request Forms? A Multi-site Thematic Analysis of Forms Towards Better Data Access Support. AMIA Annual Symposium Proceedings: American Medical Informatics Association; 2014.
2. Carpenter WR, Meyer A-M, Abernethy AP, Stürmer T, Kosorok MR. A framework for understanding cancer comparative effectiveness research data needs. Journal of Clinical Epidemiology 2012;65:1150-8.
3. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC medical informatics and decision making 2007;7:16.
4. Villanueva EV, Burrows EA, Fennessy PA, Rajendran M, Anderson JN. Improving question formulation for use in evidence appraisal in a tertiary care setting: a randomised controlled trial [ISRCTN66375463]. BMC medical informatics and decision making 2001;1:4.
5. Vechtomova O, Zhang H. Articulating complex information needs using query templates. Journal of Information Science 2009;35:439-52.