Published in final edited form as:

Int J Med Inform. 2016 July ; 91: 1–9. doi:10.1016/j.ijmedinf.2016.03.008.

A Data-driven Concept Schema for Defining Clinical Research Data Needs

Gregory W. Hruby, MA[1], **Julia Hoxha, PhD**[1], **Praveen Chandar Ravichandran, PhD**[1], **Eneida A. Mendonça, MD, PhD**[2,3], **David A Hanauer, MD, MS**[4,5], and **Chunhua Weng, PhD**[1,*]

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA

[2]Department of Pediatrics, University of Wisconsin, Madison, WI, USA

[3]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

[4]Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA

[5]School of Information, University of Michigan, Ann Arbor, MI, USA

Abstract

OBJECTIVES—The Patient, Intervention, Control/Comparison, and Outcome (PICO) framework is an effective technique for framing a clinical question. We aim to develop the counterpart of PICO to structure clinical research data needs.

METHODS—We use a data-driven approach to abstracting key concepts representing clinical research data needs by adapting and extending an expert-derived framework originally developed for defining cancer research data needs. We annotated clinical trial eligibility criteria, EHR data request logs, and data queries to electronic health records (EHR), to extract and harmonize concept classes representing clinical research data needs. We evaluated the class coverage, class preservation from the original framework, schema generalizability, schema understandability, and schema structural correctness through a semi-structured interview with eight multidisciplinary domain experts. We iteratively refined the schema based on the evaluations.

RESULTS—Our data-driven schema preserved 68% of the 63 classes from the original framework and covered 88% (73/82) of the classes proposed by evaluators. Class coverage for participants of different backgrounds ranged from 60% to 100% with a median value of 95% agreement among the individual evaluators. The schema was found understandable and structurally sound.

[*]**Corresponding author**, Chunhua Weng, PhD, FACMI, Department of Biomedical Informatics, Columbia University, 622 West 168 Street, PH-20, New York, NY 10032, USA, ; Email: chunhua@columbia.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author contributions

GWH collected data, conducted data analyses and wrote the initial draft of the paper. CW initiated and supervised GWH's research and edited the paper critically in all iterations. All authors made substantial contributions to evaluation designs, results interpretation, and refinement of the paper and approved the final draft.

Conflicts of interest statement

No author has conflict of interest to disclose.

CONCLUSIONS—Our proposed schema may serve as the counterpart to PICO for improving the research data needs communication between researchers and informaticians.

Keywords

medical informatics; comparative effectiveness research; needs assessment; data collection; models, theoretical

1 Introduction

The rich data made available by electronic health records (EHRs) represents a promising resource for accelerating clinical and translational research [1]. However, medical researchers face significant barriers to accessing EHR data, including the articulation of their often abstract and vague data needs without knowing data details and to mapping these needs to fine-grained, contextual lower-level data representations. Two mechanisms for overcoming the barrier to mapping the data need to EHR data representations are self-service query tools [2–4] and common data elements (CDE) [5–7]. The latter are developed for standardizing research data collection and retrieval. However, complex data needs often cannot be specified in the current generation of self-service query tools [8]. At the same time, CDEs have not been widely adopted and suffer from their limited coverage, which is a common problem in clinical terminologies. As such, many medical researchers find existing query formulation solutions inadequate to help them resolve their data needs and hence have to ask an informatician to aid their data retrieval using a process called biomedical query mediation (BQM) process [8, 9]. A big part of the BQM process involves mapping abstract medical concepts to local heterogeneous data representations, while most of these data are not defined using CDEs. Moreover, it is impractical to validate the structural and content comprehensiveness of a research data query using the large number of CDEs. A preferred and more practical approach would be an abstracted concept schema that summarizes key concept classes representing clinical research data needs at a higher level. An unorganized list of many CDEs may be overwhelming to a researcher. In contrast, a concept schema can organize medical concepts commensurate with the way in which medical researchers organize those concepts. This will allow researchers to refer to the concept classes to ensure the comprehensiveness of their data requests without reviewing the extensive lists of all medical concepts.

Information needs assessment is an established research field. For any information-seeking endeavor, users are required to specify their information needs upfront [10]. In the realm of EHR data requests, task-oriented static online query forms have been explored to enable medical researchers to specify their research data needs [11]. Templates, which guide users to specify their information needs with increased specificity, have been shown effective at structuring an information need request and improving the precision and recall of information needs [12]. The best template example in the medical domain is the PICO framework [13], where P standards for population, I for intervention, C for control or comparison, and O for outcome. PICO is an effective technique for expressing information needs free of ambiguity [14] and improves information retrieval accuracy [15, 16]. The PICO framework has been shown to be effective at improving the resolution of information needs for medical literature [12, 17]. The success of PICO inspired us to develop its counterpart for articulating clinical research data needs.

Carpenter et al. developed a conceptual framework to define data needs for cancer research [18] based on semi-structured interviews and focus groups with over 76 stakeholders, including providers, researchers, industry representatives and journal editors. The framework defines data types, such as patient characteristics, diagnosis, treatment, and outcomes, as well as their temporal and association relations. The framework also represents the iterative nature of the cancer care continuum [18]. The framework provides a semi-granular representation of data needs yet remains compact enough to achieve an efficient representation of a complex information space. If able to extend beyond cancer, this framework may serve as a template for defining data requests for medical research in general.

Therefore, this study aims to use a data-driven approach to adapt and extend the Carpenter framework to achieve an enriched concept schema for defining clinical research data needs beyond the cancer domain. Our study validated and extended the Carpenter framework utilizing three data sources that represent researchers' data needs in various medical domains.

2 Methods

The study design is illustrated in Figure 1. Three data sources were processed and analyzed to identify discrete variables for specifying research data needs. We used the Carpenter framework as the starting point for data annotation and iterative schema enrichment. We performed an evaluation with eight multidisciplinary medical researchers and refined the resulting class schema for representing generic clinical research data needs accordingly. This study has received the approval from Columbia University Institutional Review Board.

2.1 Data Sources and Characteristics

Our three data sources include the public clinical trial inclusion/exclusion criteria obtained from ClinicalTrials.gov, EHR data requests submitted to our institutional clinical data warehouse, and EHR SQL queries obtained from the Department of Urology at Columbia University. The data sources represent a diverse set of values across the attributes of (1) data request type, (2) representativeness of all data needs, and (3) granularity of EHR data needs. For example, clinical research eligibility criteria represent high-level research cohort requests that are independent of the knowledge about what is retrievable from the EHR. Therefore, they tend to be vague, ambiguous, or non-granular representations of a researcher's need. In contrast, EHR data requests are expressed by a mixture of narrative descriptions of medical concepts or various terminologies frequently used in EHRs, such as ICD-9 or 10 codes or CPT codes. Finally, SQL queries are translations of EHR data requests into executable database queries. They reflect the needs of researchers based on not only what is retrievable from the EHR but also how these available data elements are encoded. Therefore, they represent the data needs at the lowest level of concept granularity (e.g., a specific representation such as "A1c" or "HbA1c" in discharge summaries or a local code for A1c in lab test results tables). We assumed these three data sources provide a rich and complementary representation for the data needs of medical researchers. Table 1 provides a detailed description of the datasets used for this project. The next section will discuss our sampling strategy for each data source.

2.2 Data Sampling

To obtain a representative sample of sentences from the clinical trial eligibility criteria, we extracted 2,729,525 sentences from 181,356 Clinical Trials downloaded from the public Clinicaltrials.gov on 2/12/2015. We annotated the concepts in these sentences with UMLS sematic types using a previously published method [19]. Using the K-means clustering algorithm [20], we divided all the enriched sentences into 27 classes. To cover sentences from these classes evenly, we sampled 1000 sentences evenly from these clusters for further annotation. For the EHR data requests logs, we randomly sampled 432/1200 data requests submitted to our data request service at Columbia University in the 2014 calendar year. A total of 897 sentences were extracted from these request logs. For the SQL queries, we used the SQL transact code associated with the 204 research projects performed at our institution's Department of Urology over the course of five years (2008–2012). For each project SQL code, we selected the "SELECT* FROM* WHERE*" statements and isolated the "SELECT *" clause for annotation.

2.3 Dataset Annotation and Analysis

Author GH annotated the datasets. This coder has 10 years of experience conducting research and 6 years of experience resolving medical researchers' data requests. We did not ask two independent annotators to annotate the datasets and measure inter-rater agreement for the following reasons. First, our goal was not to evaluate the Carpenter framework as an annotation tool, nor the process used to annotate the datasets, but to assess the portability of this framework beyond cancer and its coverage of concepts in other disease domains. Therefore, annotation is a means to achieve our goal, not the end. Second, the purpose of employing two independent annotators followed by a measurement of the inter-rater agreement is to ensure reproducible annotations generated manually. However, previous studies have reported limitations in employing inter-rater agreement for ensuring the reliability of human annotations. An example paper is provided at [21]. In this paper, the authors reported the complexities involved in reporting inter-rater reliability and some simplified inter-rater agreement calculation and reporting methods may not necessarily be reliable. Given such concerns about the limitations in the inter-rater ability assessment itself, we were more inclined to utilize a data-driven approach rather than a human-driven approach to achieve our goal. Therefore, our annotation was a semiautomatic process, which uses NLP-assisted concept recognition followed by manual mapping of each sentence represented by a set of terminology-encoded concepts into a class defined in the Carpenter model. The terminology can be UMLS for clinical research eligibility criteria or ICD-9 codes for EHR SQL queries. Therefore, the classification step performed by the annotator was informed by the rich semantic information in the UMLS concepts, including UMLS semantic types and concept definitions, rather a completely subjective process. Third, this annotator strictly followed a transparent systematic process to perform the annotation, as suggested by the following article on improving the rigor of qualitative study [22]:

1. Recognize all the concepts in the sentences/SQL variables and map each concept to a class in the Carpenter framework semi-automatically using a previously published method.
2. Tag the sentence/SQL variables with the class(es) identified from the Carpenter framework.
3. If a concept within the sentence/SQL variables is unable to be tagged with a class from the carpenter framework, label that sentence/SQL variables with “new class.”
4. Group all “new class” sentences/SQL variables and perform a thematic review to name the “new class”.
5. Review the Carpenter framework and insert new concept classes in the right positions in the hierarchy.
6. Repeat steps 1–5 until no new classes can be identified or relocated in the hierarchy.

We augmented the Carpenter framework by editing a preexisting class, adding a new class, deleting an unused class, or moving a class in the hierarchy. For example, the original class, *Comorbidities*, was expanded with the following subclasses: *Medical/Disease History; Medical/Surgical/Radiation Treatment History; Medical Device Implant; Current Medications; and Current Treatment/Experimental Trials*. Appendix provides the details of the augmentation.

2.4 Evaluation

We assessed the enriched schema using selected measures proposed by Mehmood et al.: concept class coverage, schema generalizability, class preservation, understandability, and structural correctness [23]. Each evaluation metric is further described in Table 2.

The evaluation consisted of two parts. The first part evaluated class preservation through a direct comparison of the enriched schema to the original. The second assessed the metrics of concept class coverage, schema generalizability, understandability, and structural correctness through a semi-structured one-on-one interview with eight clinical researchers (Table 3) identified through a convenience sample. Each interviewee was consented for participation and the interviews were recorded. The semi-structured interview was conducted in three blocks (see appendix). First, an introduction section designed to establish the researcher’s area of research, their cumulative experience conducting research, and the number of data request they submit in a year. Next, we presented each participant with a recent study from his or her lab and asked the participant to list the major types of data needed to conduct the study. Then we introduced our enriched schema to the participant and asked them to map the concepts they listed to the classes in our enriched schema. For example, if the participant listed 10 major types of data needed to conduct the study and they were only able to map these data to seven of the concept classes, and then we would calculate class coverage for this participant at 70% (7/10). To evaluate schema generalizability, we calculated the median of our eight participants’ class coverage. During the concept mapping exercise, we instructed the participants to “think-aloud” their actions and decision-making processes. We followed this with a set of questions addressing difficulties they may have had during the mapping process. We used the transcripts from the think-aloud process and follow-up responses to assess the evaluation metric, understandability. In the third block of questions, we evaluated the metric, structural correctness. Member checking was performed to confirm our interpretation of the evaluation results with each participant. Additionally, augmentations to the enriched schema were made to accommodate constructive feedback we received during the evaluation process.

3 Results

3.1 Data-Enriched Schema

We identified 1064, 1970, and 1892 concepts from the clinical trial eligibility criteria, the clinical research data requests, and the SQL statements, respectively. These concepts were mapped to 72 classes in the enriched schema. Figure 2 is a Venn diagram displaying the union and intersections for the 72 classes across the three data sets. Figure 3 displays the data enriched schema. The notable structural change was to associate "Organizational/ Provider Characteristics" with "Detection/Diagnosis" and "Intervention" instead of the "Patient" section. In the appendix, we provide definitions and examples for the 72 classes presented in Figure 3.

3.2 Evaluation

With regard to class coverage, the schema contains 89% (73/82) of the concept classes used by our participants. For generalizability, the schema accurately identified concept classes from diverse medical domains with a median accuracy rate of 95% (60–100%). For the metric of preservation, Table 4 displays the schema's preservation of the entities from the Carpenter framework. Overall, 79% (70/89) of the entities within our enriched schema originated from the original Carpenter framework. Table 5 shows the participant breakdown of concept preservation. The participant from Pediatrics, infectious disease reported the lowest class coverage (60%).

Table 6 presents the subjective metrics evaluated. For each metric, we identified themes derived from the interviews. We organized themes into quotes that support or oppose the data-enriched schema and provided counts for the number of times at which those themes occurred. In addition, Table 6 provides representative quotes for each theme. For the metric of understandability, the majority of the positive sentiments surrounded the organization of the classes and the schema's effect to stimulate additional medical concepts needed for research. However, the participants found significant ambiguity in the enriched schema; they described the enriched schema containing overlaps between classes from different sections. Even though the participants were able to map 89% (73/82) of the concepts they identified, they still noted missing classes. For structure, the majority found the temporal and interaction relationships between the sections of the enriched schema to be sound, with the exception of the temporal edge conveying the iterative nature of the care continuum.

3.3 Participant-Enriched Schema

Figure 4 is the participant-enriched schema based on our evaluation. This representation is a significant departure from the original Carpenter framework. We will first describe the major structural changes followed by granular class changes and their justifications, respectively. First, many evaluators expressed confusion with the directed temporal edge that made the conceptual graph cyclic. We removed this edge to simplify the intended temporal information conveyed by the directed edges. Second, many participants expressed difficulty following the temporal pattern. The original framework presented many sections connected in a parallel temporal circuit. While, this representation is probably more accurate of the clinical process, we decided to serialize the major sections in an attempt to better illustrate the temporal pathway a patient follows. Additionally, we increased the border thickness for the major sections of this temporal process: Patient, Pre-Treatment Diagnosis, Treatment, and Outcomes.

Furthermore, we renamed the major sections to better align with clinical terminology. For example, we changed the sections “Detection/Diagnostics” and “Intervention” to “Pre-Treatment Diagnosis” and “Treatment” as this better reflects clinical care documentation. This alteration is a direct change based on the following quote,

> “If you want to be more generic and applicable to screening procedures in general, one heading that proceeded the EMR, back when it was all on paper, operative notes had a ‘Pre-procedure diagnosis’. So, I wander if data elements would be better organized that way…That would guide, the clinician would immediately know which box to go to for those two things.”

The traditional language used to describe the clinical course of a patient is a key component. We felt the language used by physicians to describe the clinical course is best used to represent the sections of the schema. We combined “Survival Outcomes” and “Non-Survival Outcomes” into “Outcomes.” The original framework is based on the cancer care continuum and as such probably over emphasizes the survival outcomes from the cancer domain. Non-malignant disease researchers were confused by the focus on survival outcomes. We felt that both survival and non-survival outcomes were both classes under the section “Outcomes” and as such should be represented in one section. Finally, we created the section “Clinical Trial Enrollment” as multiple participants felt it did not belong to the set of classes in the “Patient” section. We added the following classes to the “Patient” section: Inpatient/Outpatient status (Current Service, Diet Status, Activity Status, and Primary Care Provider). Multiple participants described this as an integral class aiding cohort identification.

We added the following classes to the “Treatment” Section: Other Health Service Interaction (Anesthesia, Non-primary treatment care teams) based on an inference observed by participants 2, 3, and 7, in that many of the interventions in their studies are secondary treatments or care processes to a primary intervention the patient is receiving. This class was also of interested to participant 6, as this subject was concerned with what effect this may have on major outcomes of interest.

4 Discussion

4.1 Implications of Results

We posit the way medical researchers organize medical concepts may aid the efficient elicitation of data needs, and may provide an easier interface for informaticians to map CDE or EHR data elements to medical concepts described in the data needs. The Carpenter framework is representative for how researchers conceptually organize cancer research data needs. We hypothesized the Carpenter framework was a well-organized and comprehensive representation of concepts used in comparative effectiveness research (CER) for cancer and that it could be extended with new classes identified through real-world data to represent data needs for various medical domains. Our enrichment of the Carpenter framework utilizing three datasets provides some interesting findings. First, we confirmed that the Carpenter framework is a well-organized and comprehensive representation of medical concepts used in CER for cancer. This was observed through the high preservation of the original classes in our data-enriched schema. 79% of concepts were preserved in our data-enriched schema. Furthermore, 86% of the sections and 86% of the directed edges were preserved, suggesting the conceptual organization was persevered. Additionally, our data-enriched schema extends the breadth of classes represented for other medical domains and research approaches.

Finally, the evaluation of our data-enriched schema provided significant insight regarding the understandability of the schema. Specifically, the reorganization of the core sections in line with the directed edge representing a temporal sequence was a major adjustment intended to convey a focus on the sections across a timeline. Additionally, our intended use of the enriched schema as an aid for the specification of data needs showed initial promise. During the course of the evaluation, specifically the mapping component, the data-enriched schema stimulated many participants to describe addition medical concepts they required to complete their research. Many saw the enriched schema as a mechanism to help aid the specification of their needs, and others saw it as a tool to be used during a data needs negotiation with an informatician. We expand on this idea in the next section.

4.2 Intended Use Case

Our final schema presented in Figure 4 may serve as a bridge between the medical researcher and the informatician. Both stakeholders may use this schema to specify and elicit key medical concepts needed for a research project. We envision the employment of this schema in three scenarios. The first would be to refine a data request by providing a template through which the medical researcher could specify their data need initially. The representation may stimulate the researcher to define their data need with increased granularity and clarity. The second would provide a concept schema through which an informatician could orient themselves to the mental model of researchers, allowing them to better engage and elicit additional criteria related to the initial data request. The schema may facilitate the negotiation between the researcher and informatician by supplying a checklist through which the data need can be defined. The third would serve as a metadata schema for indexing and reusing data requests. The concept schema can provide a compact list of codes for annotating the data requests.

4.3 Limitations

Our study has several limitations. First, as the evaluation confirmed, the data enriched schema does contain ambiguity. The abstraction of granular medical concepts introduces ambiguity. However, the more positively reviewed aspect of the data enriched schema was its conceptual organization of medical concepts used in research. Second, each dataset we chose contains an inherent bias. Clinical Trials represent the current state of research as influenced by major health concerns, for example cardiovascular disease, metabolic disease, and cancer. As such, this dataset may overemphasize these medical domains affecting our ability to generalize our results to other domains. Similarly, the institutional data request logs are also a representation of the research priorities at Columbia University and as such may skew the results toward those domains. Thirdly, the EHR SQL query dataset is from one domain of medicine and hence may not cover variables outside Urology.

5 Conclusions

We used a data-driven approach to develop a conceptual schema for defining clinical research data needs. Our evaluation confirms the satisfactory concept class coverage of this schema and its generalizability across disease domains. This schema has the potential to facilitate communication between researchers and informaticians, or to serve as a metadata schema for indexing, organizing data requests thereby empowering knowledge reuse among researchers. Future studies are warranted to test these potentials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was sponsored by the U.S. National Library of Medicine grants **R01LM009886** (PI: Weng) and **T15LM007079** (PI: Hripcsak) and U.S. National Center for Advancing Translational Science grant **UL1 TR000040** (PI: Ginsberg).

References

1. Hruby GW, McKiernan J, Bakken S, Weng C. A centralized research data repository enhances retrospective outcomes research capacity: a case report. J Am Med Inform Assoc. 2013; 20(3):563–567. [PubMed: 23322812]
2. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc. 2006:1040. [PubMed: 17238659]
3. Zhang GQ, Siegler T, Saxman P, Sandberg N, Mueller R, Johnson N, Hunscher D, Arabandi S. VISAGE: A Query Interface for Clinical Research. AMIA Summits Transl Sci Proc. 2010; 2010:76–80. [PubMed: 21347154]
4. Hripcsak G, Duke J, Shah N, Reich C, Huser V, Schuemie M, Suchard M, Park R, Wong I, Rijnbeek P. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Studies in health technology and informatics. 2014; 216:574–578. [PubMed: 26262116]
5. Shenvi EC, Meeker D, Boxwala AA. Understanding data requirements of retrospective studies. International journal of medical informatics. 2014 Jan; 84(1):76–84. [PubMed: 25453276]
6. von Eschenbach AC, Buetow K. Cancer informatics vision: caBIG™. Cancer informatics. 2006; 2:22. [PubMed: 19458755]
7. Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, Gustafson S, Buetow KH. caCORE: a common infrastructure for cancer informatics. Bioinformatics. 2003; 19(18):2404–2412. [PubMed: 14668224]

8. Hruby, GW.; C, J.; Patel, VL.; Weng, C. AMIA Summit on Clinical Research Informatics. San Francisco, CA, USA: 2014. Toward a Cognitive Task Analysis for Biomedical Query Mediation; p. 218-222.

9. Hruby, GW.; B, M.; Cimino, JJ.; Gao, J.; Wilcox, AB.; Hirschberg, J.; Weng, C. AMIA Summits on Translational Science Proceedings. San Francisco: 2013. Characterization of the Biomedical Query Mediation Process; p. 89-93.

10. Vakkari P. Task complexity, problem structure and information actions: integrating studies on information seeking and retrieval. Information Processing & Management. 1999; 35(6):819–837.

11. 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. 2014 Nov 14.c2014:616–625. [PubMed: 25954367]

12. Vechtomova O, Zhang H. Articulating complex information needs using query templates. Journal of Information Science. 2009; 35(4):439–452.

13. 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(1):16. [PubMed: 17573961]

14. Rao P, Andrei A, Fried A, Gonzalez D, Shine D. Assessing quality and efficiency of discharge summaries. American Journal of Medical Quality. 2005; 20(6):337–343. [PubMed: 16280397]

15. Snowball R. Using the clinical question to teach search strategy: fostering transferable conceptual skills in user education by active learning. Health Libraries Review. 1997; 14(3):167–172.

16. Ely JW, Osheroff JA, Ebell MH, Chambliss ML, Vinson DC, Stevermer JJ, Pifer EA. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. BMJ: British Medical Journal. 2002; 324(7339):710. [PubMed: 11909789]

17. 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(1):4. [PubMed: 11716797]

18. 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(11):1150–1158. [PubMed: 23017633]

19. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. Journal of the American Medical Informatics Association. 2011; 18(Suppl 1):i116–i124. [PubMed: 21807647]

20. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. Applied statistics. 1979:100–108.

21. Lopetegui MA, Bai S, Yen PY, Lai A, Embi P, Payne PR. Inter-observer reliability assessments in time motion studies: the foundation for meaningful clinical workflow analysis. AMIA Annu Symp Proc. 2013; 2013:889–896. [PubMed: 24551381]

22. Barbour RS. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? British medical journal. 2001; 322(7294):1115. [PubMed: 11337448]

23. Mehmood, K.; Cherfi, SS-S. Advances in Conceptual Modeling-Challenging Perspectives. Springer; 2009. Evaluating the functionality of conceptual models; p. 222-231.

Highlights

- We validated and enriched an existing data needs framework using data-driven methods
- The new schema can generalize beyond cancer research
- The schema can serve as a template for specifying medical researchers' data needs
- The schema can facilitate the indexing of EHR data requests and modular data queries to improve EHR data query reuse

Summary Points

What is known about data needs of researchers	What this study adds to the knowledge?
• The Carpenter framework captures the data needs of researchers for cancer research • The PICO framework helps structure clinicians' information needs • There is no counterpart of PICO to clinical research data needs as the Carpenter framework is to cancer research data needs; therefore, we need something similar to PICO and based on the Carpenter framework that works for diseases other than cancer	• We validated and enriched the existing Carpenter data needs framework using data-driven methods • The new schema can generalize beyond cancer research • The schema can serve as a template for specifying medical researchers' data needs for reusing EHR data for clinical research • The schema can facilitate indexing of data requests and modular data query reuse

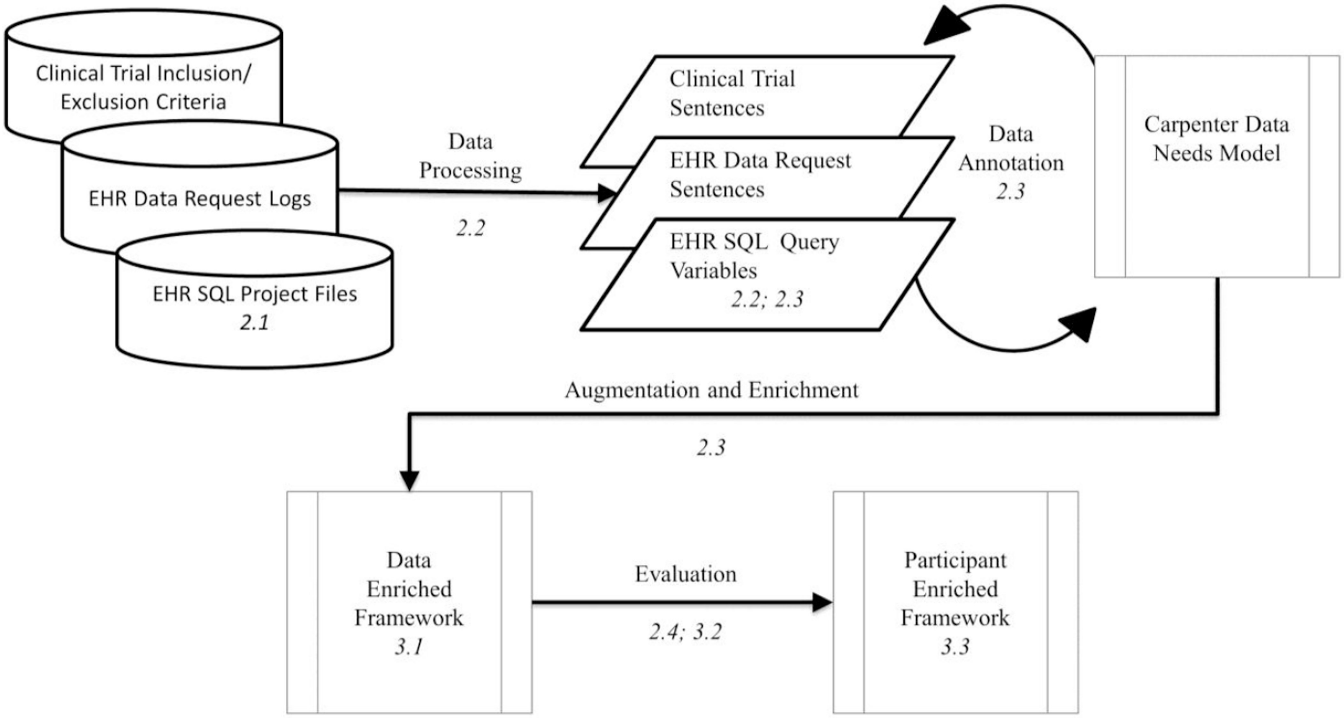

Figure 1.

Research Design. The corresponding section from both the Methods and Results sections are noted with an *italicized number*.

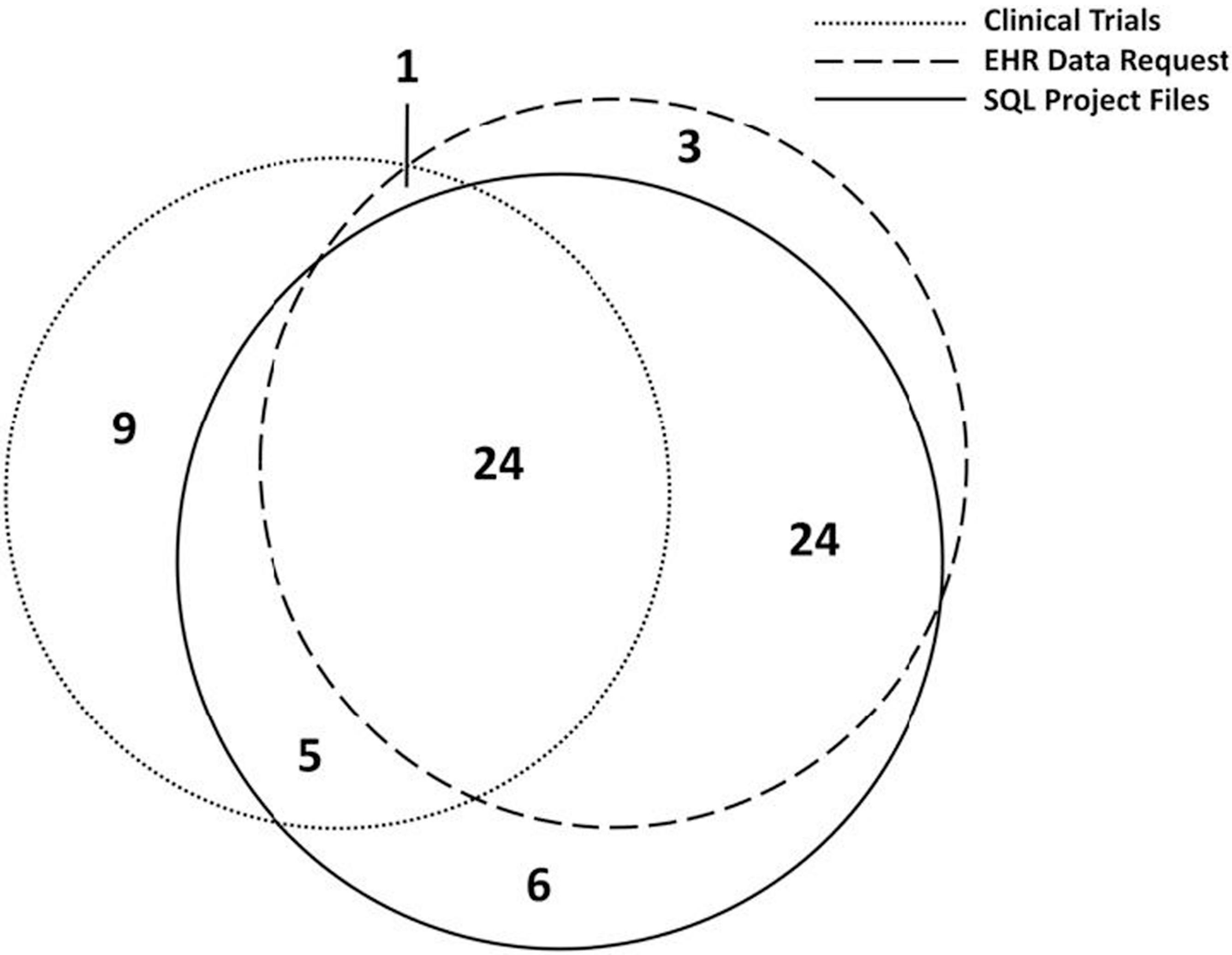

Figure 2.

This Venn diagram displays how the concepts from the three respective data sources mapped to the classes within the data enriched schema. The three datasets share coverage for 33% (24/72) classes represented in the data enriched schema.

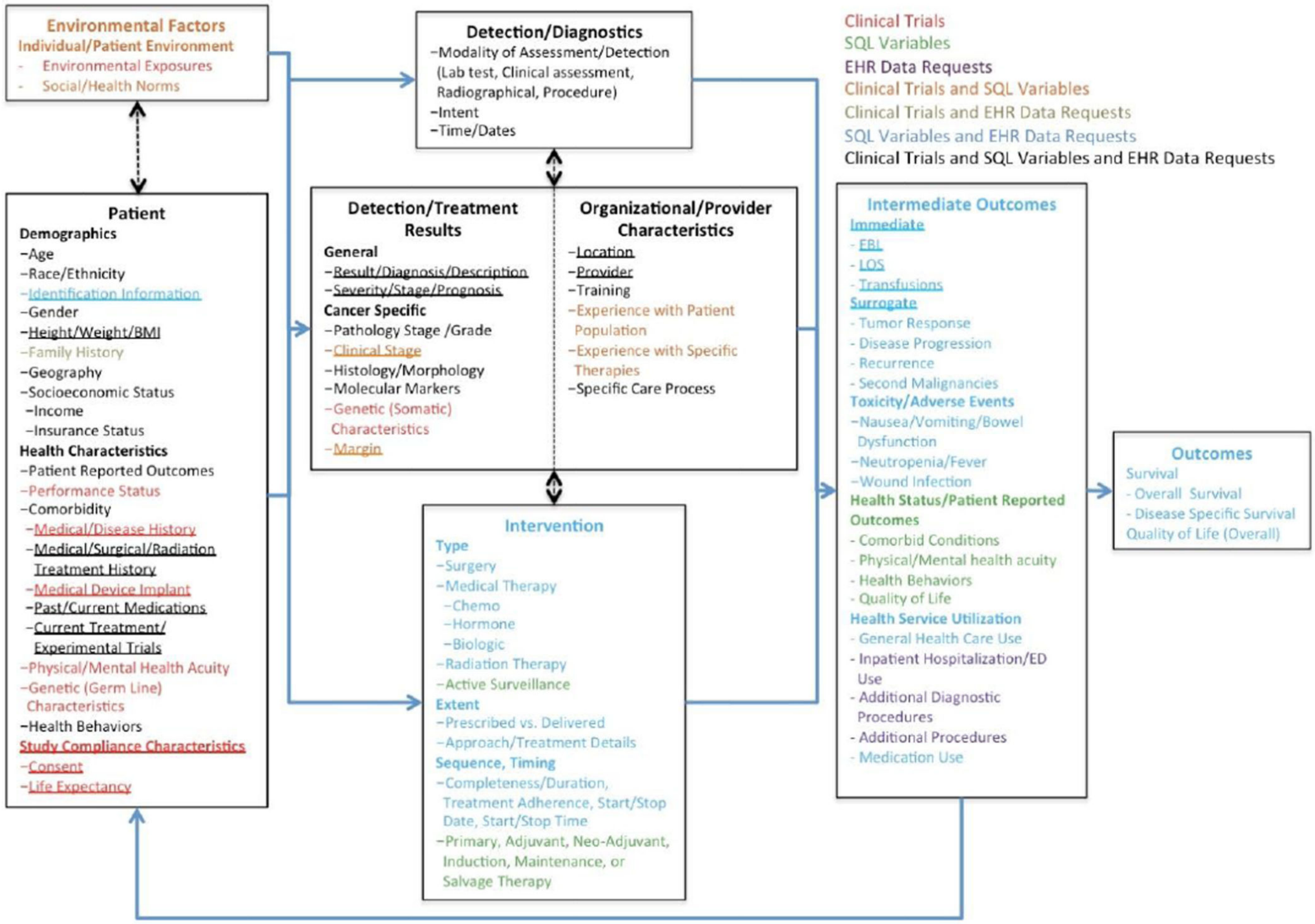
Environmental Factors
Individual/Patient Environment
- Environmental Exposures
- Social/Health Norms
Detection/Diagnostics
–Modality of Assessment/Detection (Lab test, Clinical assessment, Radiographical, Procedure)
–Intent
–Time/Dates
Clinical Trials
SQL Variables
EHR Data Requests
Clinical Trials and SQL Variables
Clinical Trials and EHR Data Requests
SQL Variables and EHR Data Requests
Clinical Trials and SQL Variables and EHR Data Requests
Patient
Demographics
–Age
–Race/Ethnicity
–Identification Information
–Gender
–Height/Weight/BMI
–Family History
–Geography
–Socioeconomic Status
–Income
–Insurance Status
Health Characteristics
–Patient Reported Outcomes
–Performance Status
–Comorbidity
–Medical/Disease History
–Medical/Surgical/Radiation Treatment History
–Medical Device Implant
–Past/Current Medications
–Current Treatment/ Experimental Trials
–Physical/Mental Health Acuity
–Genetic (Germ Line) Characteristics
–Health Behaviors
Study Compliance Characteristics
–Consent
–Life Expectancy
Detection/Treatment Results
General
–Result/Diagnosis/Description
–Severity/Stage/Prognosis
Cancer Specific
–Pathology Stage /Grade
–Clinical Stage
–Histology/Morphology
–Molecular Markers
–Genetic (Somatic) Characteristics
–Margin
Organizational/Provider Characteristics
–Location
–Provider
–Training
–Experience with Patient Population
–Experience with Specific Therapies
–Specific Care Process
Intermediate Outcomes
Immediate
- EBL
- LOS
- Transfusions
Surrogate
- Tumor Response
- Disease Progression
- Recurrence
- Second Malignancies
Toxicity/Adverse Events
–Nausea/Vomiting/Bowel Dysfunction
–Neutropenia/Fever
–Wound Infection
Health Status/Patient Reported Outcomes
- Comorbid Conditions
- Physical/Mental health acuity
- Health Behaviors
- Quality of Life
Health Service Utilization
- General Health Care Use
- Inpatient Hospitalization/ED Use
- Additional Diagnostic Procedures
- Additional Procedures
- Medication Use
Outcomes
Survival
- Overall Survival
- Disease Specific Survival
Quality of Life (Overall)
Intervention
Type
–Surgery
–Medical Therapy
–Chemo
–Hormone
–Biologic
–Radiation Therapy
–Active Surveillance
Extent
–Prescribed vs. Delivered
–Approach/Treatment Details
Sequence, Timing
–Completeness/Duration, Treatment Adherence, Start/Stop Date, Start/Stop Time
–Primary, Adjuvant, Neo-Adjuvant, Induction, Maintenance, or Salvage Therapy

Figure 3.

The data-enriched schema with 72 classes. The blue directed edges represent the temporal order as the patient moves through the care continuum. The cyclical nature of this graph implies the patient can re-enter the care cycle. The bi-directional edges indicate an association between the sections. New additions to the schema are underlined, and color-coded classes correspond to the dataset that contains the class.

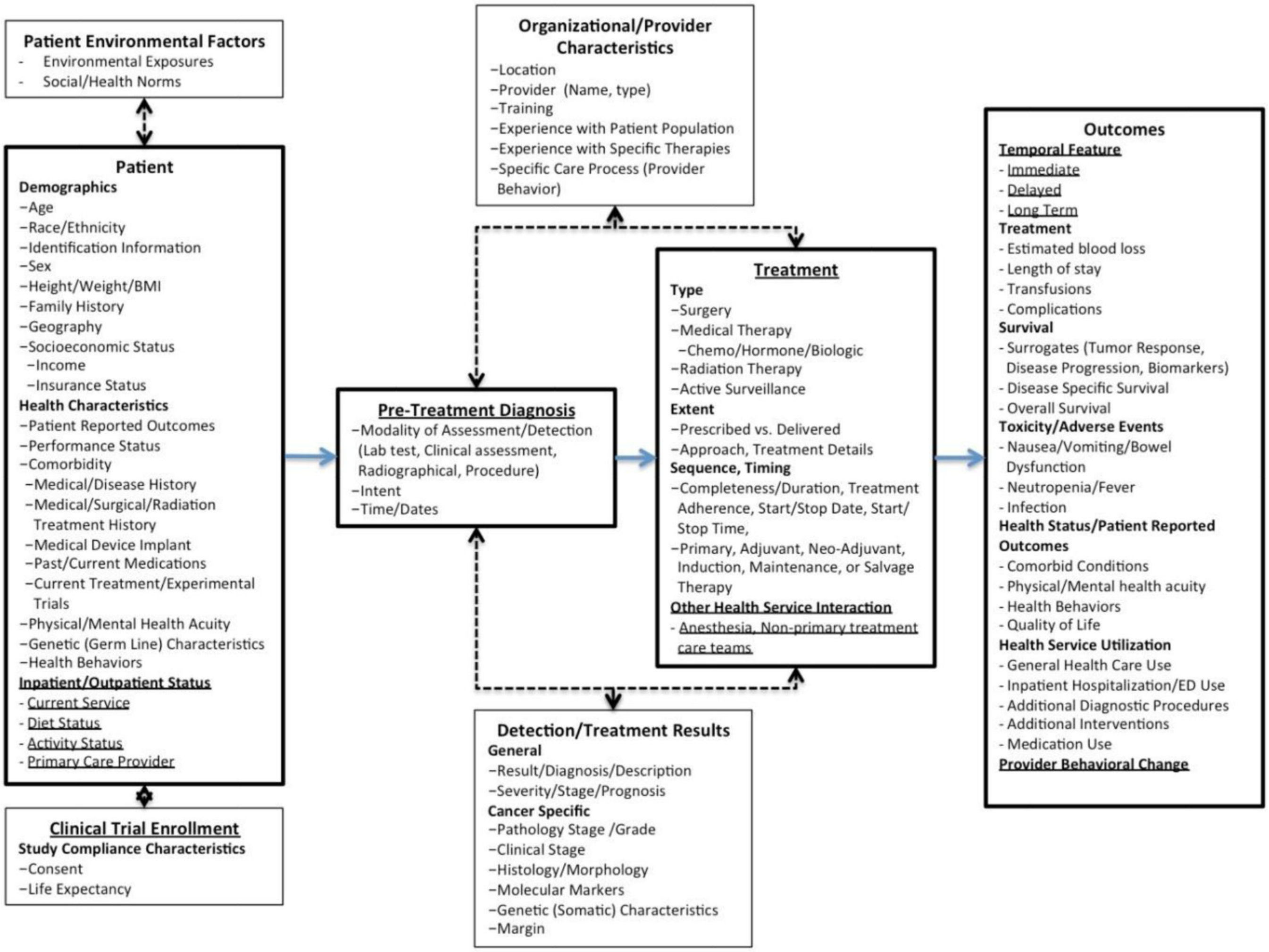
Patient Environmental Factors
- Environmental Exposures
- Social/Health Norms
Patient
Demographics
–Age
–Race/Ethnicity
–Identification Information
–Sex
–Height/Weight/BMI
–Family History
–Geography
–Socioeconomic Status
–Income
–Insurance Status
Health Characteristics
–Patient Reported Outcomes
–Performance Status
–Comorbidity
–Medical/Disease History
–Medical/Surgical/Radiation Treatment History
–Medical Device Implant
–Past/Current Medications
–Current Treatment/Experimental Trials
–Physical/Mental Health Acuity
–Genetic (Germ Line) Characteristics
–Health Behaviors
Inpatient/Outpatient Status
- Current Service
- Diet Status
- Activity Status
- Primary Care Provider
Clinical Trial Enrollment
Study Compliance Characteristics
–Consent
–Life Expectancy
Organizational/Provider Characteristics
–Location
–Provider (Name, type)
–Training
–Experience with Patient Population
–Experience with Specific Therapies
–Specific Care Process (Provider Behavior)
Pre-Treatment Diagnosis
–Modality of Assessment/Detection (Lab test, Clinical assessment, Radiographical, Procedure)
–Intent
–Time/Dates
Treatment
Type
–Surgery
–Medical Therapy
–Chemo/Hormone/Biologic
–Radiation Therapy
–Active Surveillance
Extent
–Prescribed vs. Delivered
–Approach, Treatment Details
Sequence, Timing
–Completeness/Duration, Treatment Adherence, Start/Stop Date, Start/Stop Time,
–Primary, Adjuvant, Neo-Adjuvant, Induction, Maintenance, or Salvage Therapy
Other Health Service Interaction
- Anesthesia, Non-primary treatment care teams
Detection/Treatment Results
General
–Result/Diagnosis/Description
–Severity/Stage/Prognosis
Cancer Specific
–Pathology Stage /Grade
–Clinical Stage
–Histology/Morphology
–Molecular Markers
–Genetic (Somatic) Characteristics
–Margin
Outcomes
Temporal Feature
- Immediate
- Delayed
- Long Term
Treatment
- Estimated blood loss
- Length of stay
- Transfusions
- Complications
Survival
- Surrogates (Tumor Response, Disease Progression, Biomarkers)
- Disease Specific Survival
- Overall Survival
Toxicity/Adverse Events
- Nausea/Vomiting/Bowel Dysfunction
- Neutropenia/Fever
- Infection
Health Status/Patient Reported Outcomes
- Comorbid Conditions
- Physical/Mental health acuity
- Health Behaviors
- Quality of Life
Health Service Utilization
- General Health Care Use
- Inpatient Hospitalization/ED Use
- Additional Diagnostic Procedures
- Additional Interventions
- Medication Use
Provider Behavioral Change

Figure 4.

Participant enriched schema. The major sections are aligned and highlight as boxes with thicker borders. The sections are connected in series with blue directed edges to simplify the implications of a temporal flow. The associated sections are connected with dashed, undirected edges. Our participants added 9 additional classes to the enriched schema. These are underlined within the sections. These classes were not found in the original framework. Additionally, section names that were changed are also underlined.

Table 1

The datasets used in this study and their characteristics

Data Source	Source Quantity	Annotation Quantity	Medical Domain Representativeness	Use of Data
Clinical Research Inclusion/Exclusion Criteria	181,356 Studies	1000 Sentences	No domain selection	Cohort identification
EHR Data Request Logs	432 Requests	897 Sentences	No domain selection	Cohort identification and dataset generation
EHR SQL Queries	204 Projects	1,445 Variables	Urology domain	Dataset generation for retrospective studies

Table 2

Evaluation metrics and their definitions

Metric	Definition
Class coverage	The percent of concept classes representing clinical research data needs included
Schema generalizability	The median percentage of class coverage across disease domains of our evaluators
Class preservation	The percent of classes from the original framework included in the enriched schema
Understandability	Evaluator's assessment of the clarity of the classes within the enriched schema
Structural correctness	The validity of the semantic relations and hierarchical relations among classes

Table 3

Characteristics of the Participants in the Evaluation Study

Participant	Department	Title	Research Expertise	Years of Research Experience	Number of data requests submitted/year
1	Hematology Oncology	Fellow	Quality Improvement	3	3
2	Emergency Department	Emergency Medicine Director	EHR health practice research	10+	5
3	Pediatrics; Infectious Disease	Professor of Clinical Pediatrics	Observational Epidemiology	28	5
4	Medicine; Behavioral Cardiovascular Health	Assistant Professor of Medicine	Prospective and Retrospective Studies	4	3
5	Medicine; Digestive and Liver Diseases	Assistant Professor of Medicine	Retrospective, Epidemiology	16	5–10
6	Urology Department	Professor and Chair	Prospective and Retrospective	15	52+
7	Medicine; Naomi Berrie Diabetes Center	Professor of Clinical Diabetes, Medicine and Pediatrics	Clinical Trials and Retrospective	16	20+
8	Division of Colorectal Surgery	Chief	Retrospective Outcomes Research	14	52

Table 4

The number of sections, classes and edges preserved from the original framework to the data enriched schema. We calculate the degree of preservation as the ratio of preserved entities over the total number of entities from the data enriched schema. Both major elements of the Carpenter framework, sections and the directed edges were maintained. However, the enriched schema deviated from the granular details of the original framework.

Elements	Carpenter Framework	Data-enriched Schema	Preserved Elements	Degree of Preservation
Sections	8	7	6	86%
Classes	63	72	57	79%
Directed Edges	8	7	6	86%
Bi-directional edges	4	3	1	33%
Total	**83**	**89**	**70**	**79%**

Table 5

Participant breakdown of the evaluation attributes of generalizability and class coverage

Participant	**Department**	**Concepts Identified**	**Concepts Mapped**	**Class Coverage**	**Participant comments for concepts not mapped to classes within the data-enriched schema**
1	Hematology Oncology	10	9	90%	*No class described the cost associated with tests*
2	Emergency Department	11	8	72%	*No class covered “Diet Status” for patients; The one concept existed as classes, but the participant didn’t map the concept (“Provider Behavior”; the last concept was a complex concept assessing if an order was part of a larger set of orders.*
3	Pediatrics; Infectious Disease	10	6	60%	*This participant provided concepts from a study assessing secondary preventative options for a primary treatment (e.g. The success of peri-op prophylaxis for patients undergoing cardiac treatment). The schema did not provide a class that described other health service interactions on a patient treatment regimen. This case highlights a theme of studies our schema would be unable to adequately represent.*
4	Medicine; Behavioral Cardiovascular Health	9	9	100%	*NA*
5	Medicine; Digestive and Liver Diseases	11	11	100%	*NA*
6	Urology Department	11	10	90%	*The concept listed was a set of lab tests, pre and post-operative treatment Creatinine values, that do not represent a disease status, but a health status measuring collateral damage of a primary treatment choice. While this potentially could be mapped to some classes within our schema, the association could be considered vague.*
7	Medicine; Naomi Berrie Diabetes Center	10	10	100%	*NA*
8	Division of Colorectal Surgery	10	10	100%	*NA*
	Generalizability			95%	

Table 6

The subjective metrics of understandability and structure. Within each metric, we ordered themes based on occurrence in our interview transcripts. Additionally, we provide a definition and contextualized quote for each.

Metric	Dimension	Definition	Sentiment	Sentiment Freq.	Example Quote
Understandability	Ambiguity	Difficulty in differentiating classes from different sections	Oppose	21	*"My main question is, I feel like the middle part represents what you are studying such as like a diagnostic test, that's fine, but there is some overlap conceptually between what is the test you are studying versus the test result and I think that is what informs the eligibility."*
	Precision	Applicability of the concept schema to data needs	Oppose	16	*"Interventions as two different ways, a risk factor or as a management, like the way people were randomly assigned. I think this is great, but very specific to cancer"*
			Support	3	*"Prescribed vs Delivered, well that's what I was getting at, ordered vs. Delivered So prescribed is the provider order and delivered is the administration record"*
	Organization	Alignment of the concept organization with user conceptualization	Support	11	*Organizational/Provider Characteristics, oh, location is there, I found it."*
			Oppose	3	*"So, first I was a little confused as to where to look first, cause the first thing to hit my eye was the 'Environmental Factors', and I was looking for the patient stuff, but I found it."*
	Generalizability	Generalizability of the concept schema to real experience	Support	4	*"I think it's great by the way, congratulations, I think that everything I could do could go into these buckets, but I think it makes a lot of sense, there is nothing loco here. "*
Structural Correctness	Temporality	The temporal relationships among concept classes	Support	5	*"The overarching flow, is what you would predict as we are all time orientated, I started as a patient and now I am dead"*
			Oppose	3	*"Well at first glance I have no idea, there is directionality of the arrows…yeah not clear"*
	Association	The bi-directional relationships among concept classes	Support	5	*"So the dotted lines seem like they're more of an interaction between the blocks, it is not so systematic in that it must flow in one direction…"*
			Oppose	2	*"I don't really get why organization/provider characteristics are here, paired with results as opposed to anywhere else, 'Organizational/Provider' could be paired with patient, the intervention, I don't really see why it has to be attached to 'detection/treatment.'"*
	Subsumption	The hierarchical relationships among the classes and the sections were they are located.	Support	3	*"The rest of the parent child-relationships seem fine."*
			Oppose	3	*"So the 'Study Compliance Characteristics' Yes you have consent, but 'Life Expectancy' how do you, I just don't understand, how are you getting that… how is that grouped with consent, I think that is the only one that doesn't really make any sense."*