A System and User Interface for Standardized Preparation of Analytic Data Sets

Daniella Meeker[1], Christopher Skeels[1], Laura Pearlman[2], Karl Czajkowski[2], Lucila Ohno-Machado[3]

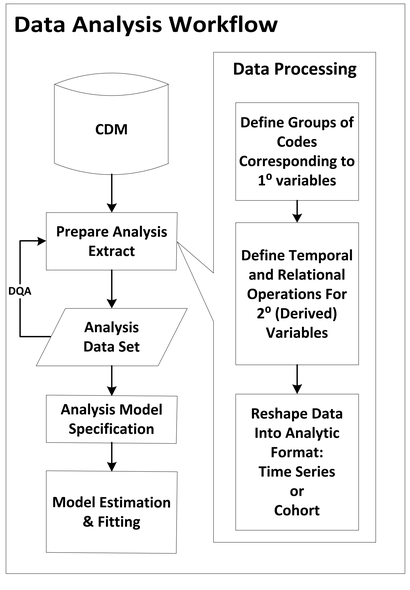
[1]RAND Corporation, [2]University of Southern California Information Sciences Institute, [3]University of California, San Diego

Abstract

Data processing is frequently more important to study design and inference in secondary analysis than analytic models or estimation methods. Investigators participating in multisite research are often faced with the tradeoff between depending on staff at partnering sites to prepare data or facing legal agreements and more complex IRB review. To support investigators in the SCAlable National Network for Comparative Effectiveness (SCANNER), three related needs were identified: (1) Investigators unfamiliar with database programming required a graphical user interface for developing well-defined specifications for data processing rules for new studies (2) Sites that have invested in data standardization should be provided with executable programs for data processing. (3) Rules should be reusable and discoverable across different studies. We used two studies in the SCANNER portfolio as test cases for determining whether procurement was an option to meet these needs. While several mature tools for building reports and distributed queries are available, our test cases were not fully supported. We identified the following gaps in in meaningful support of the process of "bringing the questions to the data": Tools with the most sophisticated capabilities for specifying data processing rules did not generate specifications that could be translated into computable queries. Tools for authoring distributed queries did not support the level of complexity necessary to produce analytic data sets for our test cases. Report generation software required direct linkage to underlying data systems and terminologies and did not generate human readable specifications that could be distributed independently or feasibly translated to executable code for other sources. We developed a system that integrates strengths of several existing tools and standards.

Background

Decoupling Data Set Preparation from Analytic Model Estimation. Arguably, the most important knowledge-based activity in a study are the applied during preparing data for analysis. Study design in secondary analysis hinges on preparation of data [7]. After data has been prepared, specification of a statistical model is quick work -- estimation algorithms are agnostic to the underlying context, semantics, and metadata. However, particularly in multisite studies where data are independently pre-processed, if specifications are not clear, many decisions are left to programmers that are unfamiliar with the study context. Efforts such as MiniSentinel and OMOP leverage a common data model to distribute programs that couple model fitting with the creation of analytic data sets [8]. The limitation of this coupled approach is that the analytic models are not decoupled from the data sets, limiting scalability of analysis given a data set. This restricts the ability of distributed research networks to take advantage of the wealth of existing open-source tools for analysis and data mining. Another advantage of decoupling data set preparation from modeling is the ability to generate safe-harbor data that can be published under HIPAA guidance, while retaining analytic utility. Typically causal inference in CER relies heavily upon temporal relationships between health events. HIPAA requires a Data Use Agreement between researchers and covered entities if dates or zip codes are included in shared data; often DUAs for sharing data can take months or years to finalize. The majority of analytic data sets only rely of temporal relationships, not specific dates, thus data can be processed to preserve temporal information and remove specific dates. This is particularly useful in distributed research networks where organizations are concerned about any architecture that enables incoming requests to systems that include Protected Health Information. Furthermore, by decoupling data preparation specifications and programs, meta-data can be made available that describes the security policies that apply to a given data set. The advantages of decoupling are clear, however there are few interfaces available that enable investigators unfamiliar with database programming to define standardized specifications that can be converted to executable programs.

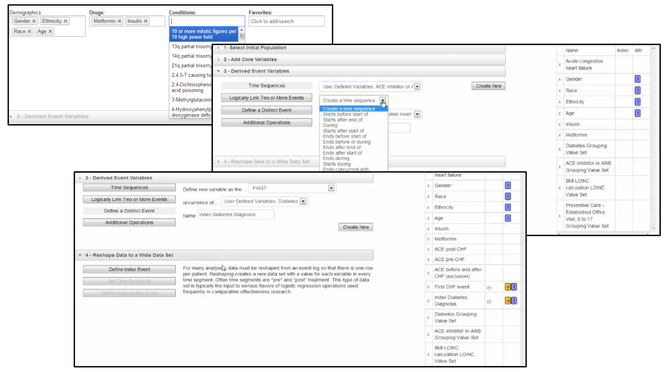


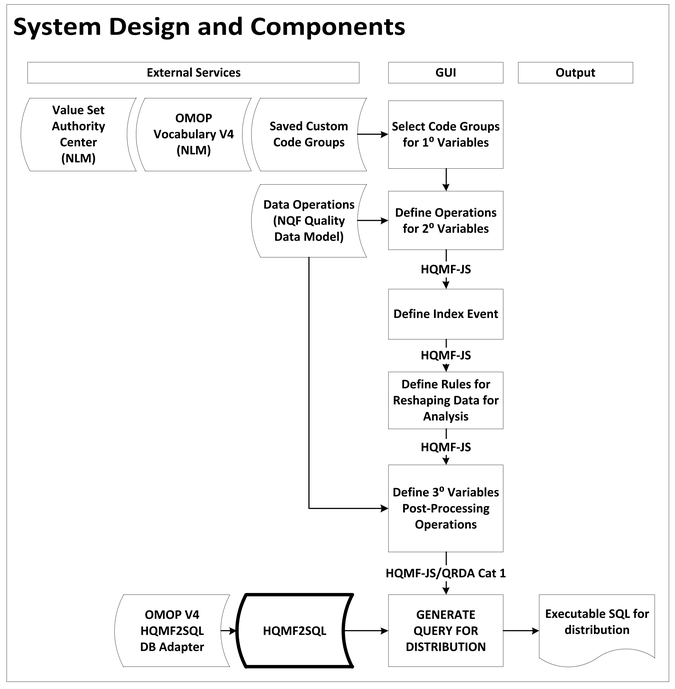
Differentiation form prior work. Systems for generating count queries and basic summary statistics based on standardized terminologies are broadly available, both commercially and as part of academic systems [9]. Report generating software also allows users to define and store complex rules. Few of these systems have been customized for use in creation of clinical analysis data sets. Related projects have applied quality measure standards for "phenotyping" purposes and composition of simple queries to summarize patient data, but to our knowledge, they have not yet fully implemented the complex data operations necessary to define analysis data sets and fully "bring the questions to the data" [10, 11]. The National Quality Forum's Measure Authoring Tool, a user interface created for CMS is used by measure developers to specify human-readable, complex clinical data processing specifications for eMeasures supports quality data model operations, but does not also generate executable programs [12]. Our system was designed to fill the gaps between these other projects, and can be incorporated into most query distribution platforms.

Requirements

- User interface should be support clinical researchers in creating specifications for general purpose analytic data sets without requiring knowledge of underlying data systems
- User interface must support data operations with sufficient complexity to meet analysis designs for two use-cases in comparative effectiveness studies
- Specifications documents must be transformed into executable programs that may be distributed to create data sets from source data systems
- Employ standard platforms and frameworks for programming, terminology, and specification syntax.
- Service oriented architecture integrated with SCANNER web services for data and policy management

Design





Component	Standard	Originator
Data Set Specifications	HQMF [1]	HL7
Data Set Specifications	HQMF-JS [2]	MITRE
Data Set Specifications	QRDA Category 1 [3, 4]	HL7
Standard Code Groupings	Value Set Authority Center	NLM/ONC
Standard Code Groupings	NLM via OMOP vocabulary	NLM/OMOP
Processing Operations	Quality Data Model [5]	National Quality Forum

Opportunities

- Extensions of HQMF2SQL module
- Adapters and plug-ins for additional data systems (e.g. MiniSentinel)
- Integration with OMOP Vocabulary *ConceptExplorer* ™
- Implementation with a simulated data "sandbox" so that investigators can view simulated results prior to distribution

References

1. HL7 Structured Documents Work Group. *Project Summary for Health Quality Measure Format*. 2012; Available from: http://www.hl7.org/special/committees/projman/searchableprojectindex.cfm?action=edit&ProjectNumber=756.
2. Pophealth Initiative. *Library to convert HQMF to JavaScript*. 2013; Available from: https://github.com/pophealth/hqmf2js.
3. Alschuler, L., C. Bennett, and K. Kallem, *Quality Reporting Document 81. Architecture (QRDA) Initiative Phase I, Final Report*. 2007, December.
4. Velamuri, S., *QRDA--technology overview and lessons learned.* Journal of healthcare information management: JHIM, 2010. **24**(3): p. 41.
5. National Quality Forum, *Quality Data Model, Technical Specification*. 2012.
6. McGraw, D. and A. Leiter, *A Policy AND TECHNOLOGY FRAMEWORK FOR USING CLINICAL DATA TO IMPROVE QUALITY.* 2012.
7. Carey, T.S., et al., *Taxonomy for Study Designs.* 2012.
8. Toh, S., et al., *Rapid Assessment of Cardiovascular Risk Among Users of Smoking Cessation Drugs Within the US Food and Drug Administration's Mini-Sentinel Program.* JAMA internal medicine, 2013. **173**(9): p. 817-819.
9. Horvath, M.M., et al., *The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement.* Journal of biomedical informatics, 2011. **44**(2): p. 266-276.
10. Thompson, W.K., et al. *An Evaluation of the NQF Quality Data Model for Representing Electronic Health Record Driven Phenotyping Algorithms*. in *American Medical Informatics Association (AMIA) Annual Symposium*. 2012.
11. Li, D., et al., *Modeling and Executing Electronic Health Records Driven Phenotyping Algorithms using the NQF Quality Data Model and JBoss® Drools Engine*, in *American Medical Informatics Association*. 2012: Chicago, IL.
12. National Quality Forum. *NQF: Measure Authoring Tool*. 2012 [cited 2012 October 12, 2012]; Available from: http://www.qualityforum.org/MAT/.

Acknowledgements

Supported by Agency for Healthcare Research and Quality through the American Recovery & Reinvestment Act of 2009 - R01 HS19913-01 and the National Institute on Aging – 1RC4AG039115-01