Data Quality Monitoring, Transparency and Governance: Enterprise process for data quality stewardship and governance for real-world data

Parsa Mirhaji, Selvin Soby, Erin Henninger, Chandra Nelapatla, Manuel Wahle, Boudewijn Aasman, Eran Bellin

Background

Ascertaining and maintaining high quality of data is the key prerequisite for establishing a reliable and usable clinical data warehouse for research, discovery, and collaboration. It is also one of the most challenging and resource intensive processes to establish, maintain and scale especially in large enterprises where real-world observational health data is to be used by multiple clinical, operational, research, and collaborative stakeholders.[1,2] Supporting multitudes of analytic, data science, and operational use-cases throughout distributed teams of informaticians, biostatisticians, data scientists, research collaborators, and operational administrators requires a robust, transparent, and scalable data quality stewardship, governance process that would institutionalize data literacy and culture and give rise to data-driven research networks that can support learning healthcare systems.

In this presentation we will introduce an enterprise approach to characterize, assess, and ascertain data quality and to establish a transparent process for data stewardship, governance, and monitoring data quality for OHDSI/OMOP based clinical data warehouses.

Methods

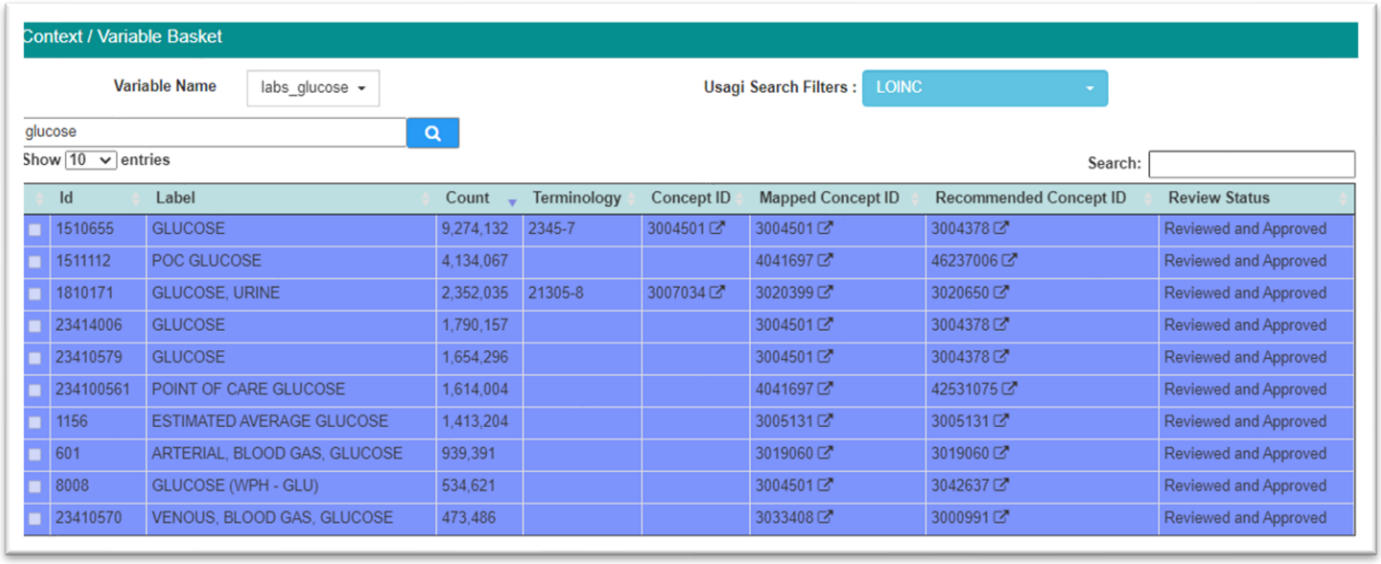
A: Metadata management: We have used principles of the Semantic Web and RDF/RDFS modeling to interrogate and represent the information and data model itself from the source as a machine understandable knowledge-graph suitable for querying, machine-reasoning, and linkage. This enables a loss-less and granular representation of the schema, relationships, metadata, data discovery, and optimized strategies for extraction and retrieval. The approach also facilitates automated data quality assessment contextualized to the type and semantics of the data and its relationships at the source.

B: Knowledge and terminology management: We have translated and represented the vocabularies available in OMOP and ATHENA through Simple Knowledge Organization System (SKOS) which is a W3C standard for representation of thesauri, dictionaries, and vocabulary systems for linkage and semantic web.[3,4] We used it to enable semantic indexing, mapping, and linkage of all metadata, logic, ETL processes, and their logs to the underlying OMOP concepts. Our SKOS model also provides semantic linkage and mapping to the SKOS representation of source vocabularies from the UMLS.[5] We have further extended this SKOS model to also represent custom concepts that are not currently present in OHDSI/ATHENA

C: Semantic indexing: We have developed a robust vocabulary mapping framework that utilizes a combination of domain expertise provided by informatics analysts, mapping information provided by source systems, heuristics, and NLP for terminology mapping. For custom concepts that may not directly map to an OHDSI/ATHENA terminology system, we extend our SKOS based knowledge representation to provide a semantic model consistent and linked to OHDSI/ATHENA.

D: Natural language processing: We have built an integrated and scalable NLP process (using the Elastic Search and cTAKES Open-Source tools) based on semantic indexing of clinical and biomedical concepts using OHDSI/ATHENA vocabularies to provide automated suggestions during the concept

mapping process. Each variable in the source database is checked against



Id	Label	Count	Terminology	Concept ID	Mapped Concept ID	Recommended Concept ID	Review Status
1510655	GLUCOSE	9,274,132	2345-7	3004501	3004501	3004378	Reviewed and Approved
1511112	POC GLUCOSE	4,134,067			4041697	46237006	Reviewed and Approved
1810171	GLUCOSE, URINE	2,352,035	21305-8	3007034	3020399	3020650	Reviewed and Approved
23414006	GLUCOSE	1,790,157			3004501	3004378	Reviewed and Approved
23410579	GLUCOSE	1,654,296			3004501	3004378	Reviewed and Approved
234100561	POINT OF CARE GLUCOSE	1,614,004			4041697	42531075	Reviewed and Approved
1156	ESTIMATED AVERAGE GLUCOSE	1,413,204			3005131	3005131	Reviewed and Approved
601	ARTERIAL, BLOOD GAS, GLUCOSE	939,391			3019060	3019060	Reviewed and Approved
8008	GLUCOSE (WPH - GLU)	534,621			3004501	3042637	Reviewed and Approved
23410570	VENOUS, BLOOD GAS, GLUCOSE	473,486			3033408	3000991	Reviewed and Approved

Figure 1: Interrogator data variable search results to aid in mapping decision support

E: Quality Characterization: We have developed a data quality characterization and monitoring process (QualiPy) that uses machine learning, pattern recognition, and big-data analytics to characterize and identify anomalies in longitudinal data. The process uses signal detection by using both a generalized linear model and an isolation forest-based machine-learning algorithms. Rather than treating data quality information as snapshots from a moment in time, QualiPy measures and characterizes data quality over time. This includes different aspects of the quality of the underlying data such as conformance, completeness, integrity, variability, and overall trends.

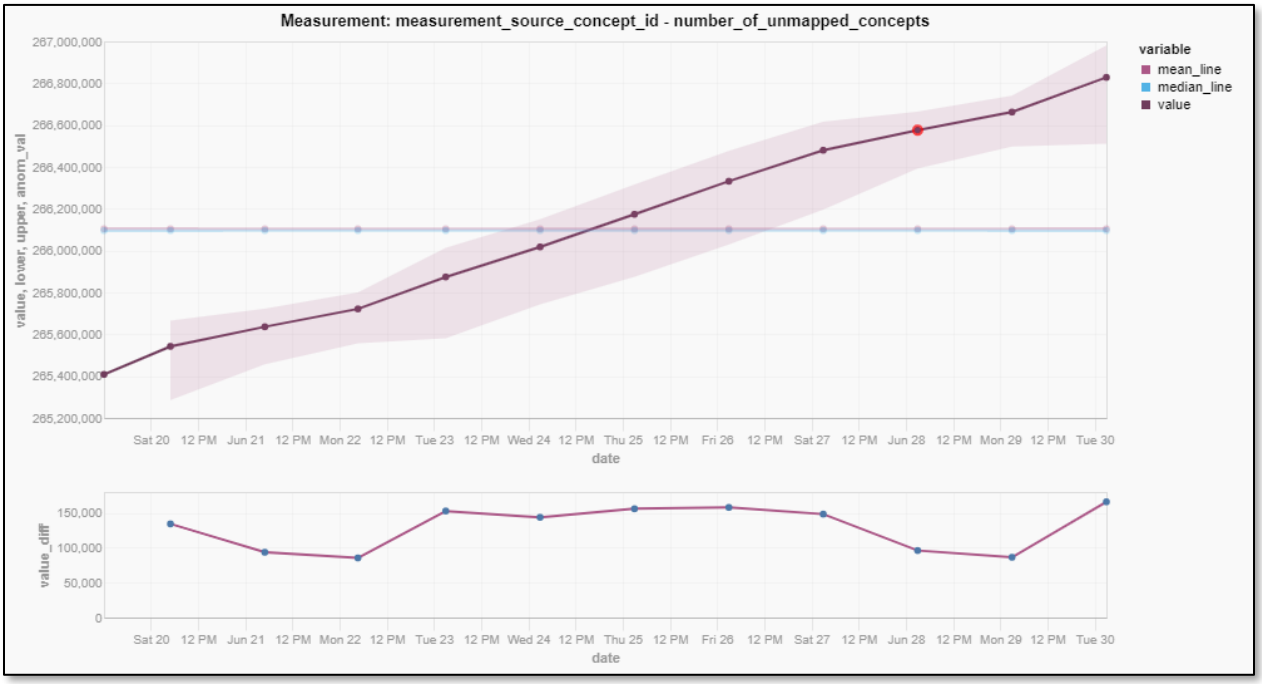


Figure 2: Numerical anomaly detection over time

F: Information visualization: We have developed a series of visualizations and interactive web-based user-interfaces to expose the underlying metadata, mappings, and a data quality characterization report that comparatively illustrates source and destination data, and their prospective change over-time.

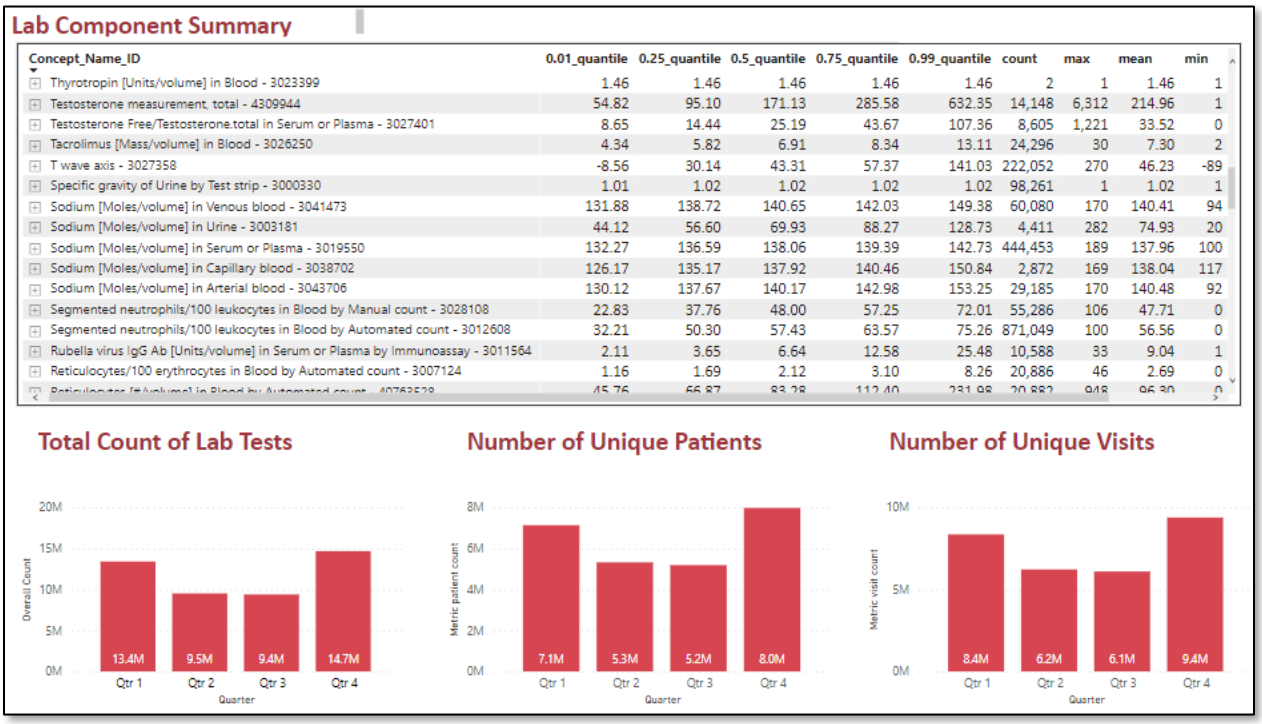


Figure 3: Web-based, interactive data quality dashboard

G: Validation and review: All mappings, data quality reports (as characterized at the source and within the OMOP-CDM) are reviewed by two domain experts by a informaticist and is either approved for deployment or rejected in order to further analyze and make corrections. Audit histories of all reviews and validation steps are logged and tracked to help maintain data governance.

H: Governance: All metadata, mappings, quality reports (including changes over time), validation, and review processes are tracked, logged, and transparently available to all users and administrators or the system (including but not limited to informatics support teams, information technology support teams, and enterprise data governance teams).

I: Systems Overview

Our software platform (Interrogator) was developed to implement an end-to-end process to keep OMOP/OHDSI based clinical data warehouses updated with heterogeneous real-world observational health data. This platform enables informatics analysts with advanced knowledge management, semantic indexing, and quality assessment tools. The platform simultaneously enables researchers with transparent and granular information about quality of underlying data for their use-cases. History of all knowledge and data management interactions are recorded and made available to our stewards for data governance.

Results

OHDSI/OMOP has become the primary clinical data warehouse technology for enterprise use across Montefiore Health System and is currently supporting greater than 300 users with hundreds of unique cohorts, and 6 multi-center research consortiums. The pursuit of transparent data quality, and governance and scalable data management has given rise to adoption of OHDSI/OMOP as our strategic resource that supports data science, informatics research, research informatics operations, centers of excellence for patient recruitment and clinical trial mapping, construction of disease specific registries, internal and external collaborations between care teams and investigators, and has inspired innovations and collaborations that were not previously possible. To achieve the full potential of Real-World-Data (RWD) analysis for science, we have implemented and extended our instance of ATLAS to democratize access to high quality clinical data, while preserving confidentiality and security of personal health information.

Conclusion

We have developed an end-to-end process to support and transparently represent a modern, scalable quality monitoring, data stewardship, and governance process for OMOP/OHDSI based clinical data warehouses to support enterprise use-cases for RWD research, care coordination operations, and collaboration.

References

1. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, Schilling LM, Weiskopf NG, Williams AE, Zozus MN. Transparent reporting of data quality in distributed data networks. EGEMS (Wash DC). 2015 Mar 23;3(1):1052.
2. Kunapareddy N, Mirhaji P, Richards D, Casscells SW. Information integration from heterogeneous data sources: a Semantic Web approach. AMIA Annu Symp Proc. 2006;2006:992.
3. ATHENA Standardized Vocabularies – OHDSI n.d. https://www.ohdsi.org/analytic314 tools/athena-standardized-vocabularies/
4. Achard F. XML, Bioinformatics and Data integration. Bioinformatics. 2001;17:115–125.
5. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc 1998; 5 (1): 1–11.