Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data

Data Quality Metrics System Final Report

U.S. Food and Drug Administration

and the Sentinel Operations Center

December 31, 2019

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I. This project was funded by the FDA through HHS Mini-Sentinel contract number HHSF223200910006I. This work was supported by the Office of the Secretary PCORTF under Interagency Agreement #750016PE060001.

Data Quality Metrics System

Final Report

Table of Contents

I. EXECUTIVE SUMMARY

Growth in the availability and use of electronic health data for research has generated incredible opportunities to improve human health and delivery of health care, from identifying the right treatment for the right patient, to identifying influenza outbreaks, to monitoring the safety of medicines and vaccines. The availability of these real-world data (RWD) sources has also created confusion regarding the best way to find the right data source to answer the question and avoid mistakes by using an inappropriate source. The goal of the Data Quality Metrics (DQM) System project was to provide a harmonized data characterization toolkit to enable researchers to efficiently compare data sources to better contextualize data quality and fitness-for-purpose and to help with interpretation of findings – to find the right data to answer the question.

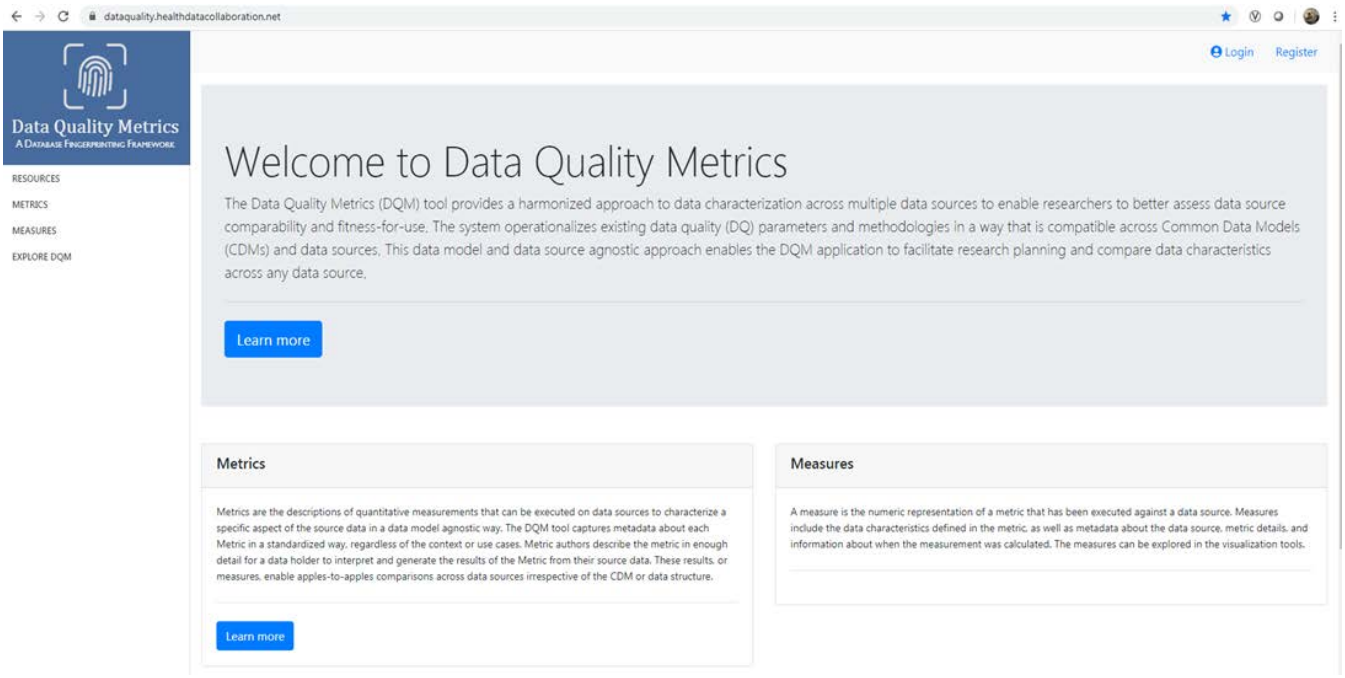
The proliferation of RWD sources such as electronic health records, health insurance claims data, and disease registries coupled with advances in data analytics, such as machine learning and artificial intelligence, is expected to generate substantial improvements in human health and health care delivery. The ability of new data sources and tools to generate new knowledge is unprecedented and growing rapidly. Research that previously took years can now be done in days or months. These advances heighten the importance of understanding data quality and comparing data characteristics across data sources to help researchers better match data sources to questions and to help decision makers better understand and interpret findings.

This project designed, tested, and released for open-source use a web-based data quality toolkit for exploring and describing the quality, completeness, and stability of data sources and visualization of data quality metrics from any data source. The DQM system enables flexible exploration of data source characteristics for multiple data sources at the same time. The flexible data quality metric data model embedded in the DQM system assists researchers and funding organizations in determining fitness-for-use of various data sources and research purposes

The following products were produced by the project and have been made publicly available for researchers and developers:

Documentation	DQM user and implementation guidance is available on the project GitHub repository: https://github.com/PopMedNet-Team/DataQualityMetrics Additional resources are provided on the DQM website (see below).
DQM source code	DQM website, software, and underlying data model were operationalized at the following link: https://dataquality.healthdatacollaboration.net/ The source code for the system is available in the project GitHub repository:

	https://github.com/PopMedNet-Team/DataQualityMetrics



Data Quality Metrics System Website Homepage (https://dataquality.healthdatacollaboration.net/)

II. OVERVIEW AND OBJECTIVES

The increasing availability of real-world data (RWD) sources has created confusion regarding the best way to find the right data source to answer the question and avoid mistakes by using an inappropriate source. The goal of the Data Quality Metrics (DQM) System project was to provide a harmonized data characterization toolkit to enable researchers to efficiently compare data sources to better contextualize data quality and fitness-for-purpose and help with interpretation of findings – to find the right data to answer the question. In this context we use "data quality" as a general term to describe various characteristics of a specific data source; these characteristics do not represent value judgements but rather agnostic measures for use by researchers to help assess a data source's fitness for use. The project adopted the Harmonized Data Quality Framework that defines data quality standards and metrics in a general and theoretical fashion and applied the framework to a variety of real-world data sources and research needs.[1] The framework aimed to address widespread variation in how individual

[1] Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Washington, DC).* 2016;4(1):1244.

institutions and networks of institutions assess data quality and describe data characteristics; a harmonized terminology and framework allows researchers and funders to approach data quality and characterization from a unified perspective. This project leveraged the framework to create a system that uses a shared vocabulary and standardized format for assessing and reporting on data. Operationalizing the framework (i.e., bringing it from theory into practice) and developing a tool for analyses allows researchers to evaluate data quality (DQ) consistently and effectively across data sources.

We created and implemented a data quality data model to contain a set of metadata standards and metrics describing: 1) Data quality and characteristics; 2) Data sources and institutional characteristics; and 3) Fitness-for-use. These standards were the basis for a web-based data quality toolkit to enable exploring and describing the quality, completeness, and stability of data sources and visualization of data quality metrics from any data source. The open-source web-based system (the DQM system) was designed to enable flexible exploration of DQ characteristics for multiple data sources at the same time. This work included the creation of a flexible data quality data model that is agnostic to the underlying data source, making it compatible across any Common Data Model (CDMs). The flexible data quality metric data model will assist researchers and funding organizations in determining fitness-for-use of various data sources and research purposes. Together, the information described provides a standardized data source "fingerprint" that can be expanded to provide additional granularity. The "fingerprint" of each unique data source is made up of various data characterizations and information/metadata and provides a consistent data description for each data source; the "fingerprint" is an agnostic characterization of the data that researchers can use to assess fitness for purpose. For example, a database "fingerprint" can provide the distribution of laboratory results available for a specific population but the researcher has to make the specific fitness for purpose assessment based on the specific question to be answered. Further, the "fingerprint" can describe the proportion of measures that fall outside an expected range, but only the researcher can assess whether the data are appropriate for use for the specific use case. Rather than executing data quality checks with binary results (i.e., pass/fail), the DQM system provides the information and data source metadata needed to allow context-specific evaluation.

The project had three distinct phases:

- **Discovery and Design**: evaluate existing data quality frameworks and processes and develop a data quality data model to enable exploration of data quality metrics in a way that is flexible and agnostic to CDM
- **Development and Testing**: develop web-based system and accompanying database in which to store data quality information; integrate feedback from key stakeholders
- **Implementation and Release**: publish technical and user documentation and the source code to a public GitHub repository

This final study report summarizes the problems addressed, the study methodology, findings, and lessons learned. The appendices include the other project reports and deliverables generated

throughout the course of the project, including detailed information on the technical design and implementation of the system; a guide for system end users; and feedback provided by stakeholders that ultimately informed design and implementation.

III. BACKGROUND - PROBLEMS ADDRESSED

The proliferation of RWD sources such as electronic health records, health insurance claims data, and disease registries coupled with advances in data analytics, such as machine learning and artificial intelligence, is expected to generate substantial improvements in human health and health care delivery. The ability of new data sources and tools to generate new knowledge is unprecedented and growing rapidly. Research that previously took years can now be done in days or months. These advances heighten the importance of understanding data quality and comparing data characteristics across data sources to help researchers better match data sources to questions and to help decision makers better understand and interpret findings.

Understanding data quality and comparing quality in a consistent "apples-to-apples" manner is a critical foundational need to support the growing use of RWD. Differences in how data are collected and represented in different data sources and distributed research networks makes it difficult for investigators to judge the fitness of a data source for a particular research project. The DQM system was developed as a step toward addressing that critical challenge by enabling consistent apples-to-apples comparisons through establishment of a flexible data quality metric standards that can be used across all types of data sources. Establishing standardized data quality metrics and implementing an open-source toolkit required in-depth systems design work coupled with real-world use cases and software development expertise.

The DQM system was designed to be flexible so it can accommodate the capture of data quality metric metadata, data source metadata, data quality output, and data quality output searching and visualizations. The initial set of metrics were intended as a starting point, with the system designed to be expanded by the community of users.

This project addresses critical strategic priorities for clinical research in the US generally, and for the Department of Health and Human Services (HHS) specifically, including the use of clinical data and publicly-funded data systems for research. Of particular interest to HHS is standards-based use of patient-contributed data (for which the system does not currently contain metrics and would be part of future work), electronic health record data, and health insurance data.

IV. METHODOLOGY

The DQM system was developed and tested in three sequential phases. The development approach was selected to maximize the flexibility of the system for future use while creating a final, open-source product that could be used and expanded by the stakeholder community. Each phase is described below.

A. PHASE 1: DISCOVERY AND DESIGN

Throughout the Discovery and Design phase, the project team evaluated existing DQ frameworks and processes, and developed a data quality data model to enable exploration of data quality metrics in a way that is flexible and agnostic to any specific Common Data Model (CDM). The foundation of this was the Harmonized Data Quality Framework developed by Kahn et al[1]; the project team operationalized the conceptual framework to inform the data quality data model underlying the web-based system. In essence, the project team's goal was to bring the theoretical data quality framework into practice. To do so, the project team created use cases based on data quality and characterizations found in various networks, such as Sentinel and PCORnet. Each of the use cases were then mapped to the relevant Data Quality Harmonized categories, thereby forming the basis of the data quality data model and system.

The project team leveraged the work of a prior APSE project – the Cross Network Directory Service (CNDS)[2] – that focused on the discovery of data sources and researchers appropriate for a specific study. DQM extends the work of the CNDS in two ways; first by leveraging many of the CNDS governance and access control capabilities[3], and second, by allowing investigators to take a deeper dive into the data sources by investigating the characteristics of the data sources and the quality of specific data elements and domains. This phase of the project included detailed work on use cases and data model design. As part of that investigation three key components of the DQM system were identified and designed for development and testing.

- **Metrics**: Metrics are the descriptions of quantitative measurements that can be executed on data sources to characterize a specific aspect of the source data in a data model agnostic way. Metric authors describe the metric in enough detail for a data holder to interpret and generate the results of the metric from their source data.
- **Measures**: A measure is the numeric representation of a metric that has been executed against a data source, i.e. the results to the metric. Measures include the data characteristics defined in the metric, as well as metadata about the data source, metric details, and information regarding when the measurement was calculated.
- **Exploration**: The DQM visualization tools overlay the metadata, metrics, and measures. Users can explore and evaluate data sources for specific characteristics, trends, and quality. DQM does not determine whether a data source passes or fails the execution of a metric, but rather provides a view of data characteristics that enable a user to determine if the data are fit for their purpose.

[2] Malenfant JM, Hochstadt J, Nolan B, Barrett K, Corriveau D, Dee D, Harris M, Herzig-Marx C, Nair VP, Wyner Z, Brown JS. Cross-Network Directory Service: Infrastructure to enable collaborations across distributed research networks. Learn Health Sys. 2019;3:e10187.https://doi.org/10.1002/lrh2.1018712.

[3] Davies M, Erickson K, Wyner Z, Malenfant JM, Rosen R, Brown JS. Software-enabled Distributed Network Governance: The PopMedNet™ Experience. EGEMS (Wash DC). 2016 Mar 30;4(2):1213. DOI: 10.13063/2327-9214.1213.

B. PHASE 2: DEVELOPMENT AND TESTING

The data quality data model designed in Phase 1 was implemented in Phase 2 as a beta-version of the DQM System web portal. The project team created a user-friendly web portal that allows users to author metrics describing data quality and characterization measures. The DQM system was populated with metrics developed from an initial list of use cases based on existing networks such as Sentinel and PCORnet. This ensured that the system was flexible and could handle various types of metrics that were agnostic to CDMs. The project team also tested how to upload measures. Through an iterative process the project team modified the system until it could address all use cases. Visualizations were developed using Qlik Sense, a commonly-used business intelligence visualization tool that enables development of custom applications. The beta-version of the system embedded custom Qlik apps directly into the web application, though the system architecture allows use of any visualization tool preferred by the user.

Once an operational beta-version of the software was developed we held four stakeholder sessions to elicit feedback from community members with interest in the theoretical work of data quality and in evaluation of fitness-for-use. The DQM software was updated based on the stakeholder feedback, including numerous changes to text, the metadata model, and visualization. Feedback that could not be incorporated into the final software release was documented for future work.

C. PHASE 3: IMPLEMENTATION AND RELEASE

The last phase of the project was to document and release the software for use by the open-source community and anyone interested. In addition to public posting of all project material, the project team presented the DQM system work to stakeholder audiences including the Data Quality Collaboratory Webinar and the FDA OSE Safety Seminar. The presentations, also available publicly, describe the project goals, objectives, and results.

The project outputs listed in the following section are available online in the GitHub repository and DQM system, and have been included in this report as appendices.

V. ACCOMPLISHMENTS AND OUTPUTS

Accomplishments throughout the project are noted below.

A. IMPLEMENTATION AND USER DOCUMENTATION

The open source code for the DQM system was posted on the DQM GitHub repository with accompanying technical and user documentation for public access. The web-based Data Quality Metrics system (i.e., the DQM website hosted and available to the public) was implemented and is available online here: https://dataquality.healthdatacollaboration.net/

- **Discovery and Design documentation:** Discovery and Design documentation (see Appendix A) describes the metadata standards and relevant use cases, technical specifications for implementing the standards, and a dictionary describing each metadata

element. The document also includes information about the data quality data model; it is intended for software developers and other technical stakeholders.

- **Technical Documentation:** The Technical Documentation (see Appendix B) provides technical information appropriate for software developers and other technical users to facilitate their use of the DQM system. It is available in the GitHub repository for reference with the system source code.
    - System visualization was implemented using Qlik Sense, although any other business intelligence or visualization tool (e.g., Tableau) could be used within the DQM system. Details on the specific Qlik visualizations can be found in the technical documentation (see Appendix B) and user documentation (see Appendix E).
- **User Documentation**: The User Documentation (see Appendix E) provides detailed user information related to the use of the web-based DQM system. The report is written to support researcher/investigator users of the system by describing all elements of the web-based system and providing instructional detail on use by an individual.

B. EXTERNAL REVIEW AND TESTING DOCUMENTATION

- **Project Requirements and Testing Table:** All requirements and design specifications were documented in JIRA. During the system testing in Phase 2, all bug reports and updates to the system were also recorded in JIRA. A table containing the list of JIRA issues for this project is available (see Appendix C) for technical stakeholders to view to gain a better understanding of the process of creating the DQM system.

  Stakeholder Summary: The Stakeholder Summary (see Appendix D) documents the stakeholder engagement activities, including documentation of stakeholder comments and disposition of comments. This feedback informed additional testing and updates to the system to ensure end user goals were addressed.
    - Once a functional beta version of the software was ready for external review, four stakeholder sessions were held to elicit feedback and inform a final proof-of-concept system. Over 25 participants from the US and Europe attended the stakeholder sessions. Visualizations for selected use cases were created and revised based on stakeholder feedback.

VI. LESSONS LEARNED AND CONSIDERATIONS FOR FUTURE WORK

A. LESSONS LEARNED

The project team has assessed lessons learned throughout the project period within two significant themes: governance and requirements for contributors. Through engagement with various stakeholder groups, common feedback arose around the future coordination of the DQM system and concerns regarding governance, and data confidentiality and sharing agreements. These conversations further informed lessons learned related to the role of a

coordinating center and the development of role-based access controls and business rules within the system.

1. Governance

a) Coordinating Center

Operationalizing the DQM system will require designated funding and a Coordinating Center to operate the system. Although the software is open source and freely available, operating a network requires resources to manage and update the website and to engage with system users. Activities include registration of users, updates to software, development of visualizations, and monitoring of metrics and submitted measures. Adherence to established data sharing and data use agreements is another critical role of the Coordinating Center.

b) Governance and Implementation

This project developed a beta-version of the system, for which all code and technical documentation was made publicly available on GitHub for individual developers interested in downloading and instantiating their own systems. A production-level version with web hosting would be required to enable any individual interested in the system to utilize it and would maintain a live version of the web portal. For a future, production-level version of the DQM system, additional discussions are required to address concerns about governance for the contribution and evaluation of data. Questions about who gets to see what data and for what purpose were of primary concern. Even though the data quality data model does not use person-level information, data sharing and use agreements have strict controls of access and use of such data. More work is needed in this space to convince data holders to make their aggregate data available of use within the system. Some of this work will require data use agreements, but another aspect of the work relates to trust and the security and access controls embedded in the system. There is risk that the most "conservative" data source (i.e., the source with the most restrictive access control model) will dictate the system specifications for the rest of the data holders; approaches to avoid this outcome are critical.

2. Potential requirements for contributors

The project team has discussed incentives for and barriers to participation with multiple stakeholders throughout the project period. The value proposition is not clear for a system such as this until there is sufficient adoption; that is, it is hard to get the first set of users and much easier to get the next set. One approach is to have one or more networks support implementation of the DQM system and have their network data sources join to create the critical mass of users that will help convince others to participate. Another strategy is to implement a production version of the system with a Coordinating Center within FDA to enable the data quality of data sources in the Sentinel system, and more broadly, moving forward.

B. CONSIDERATIONS FOR FUTURE WORK

1. Currently the DQM system leverages the CNDS data source registration infrastructure. The DQM infrastructure was designed to be compatible, but independent from, PopMedNet. However, it is possible that coupling the CNDS, PopMedNet, and DQM will provide efficiencies and enable easier adoption since multiple networks use PopMedNet.
2. Additional features to enhance the management of the metrics. Currently, the DQM Site Administrators do not have the ability to alter the existing metric fields or the ability to designate them as required or optional. These changes are coded in by a developer. The system would be greatly enhanced by creating the ability for the Site Administrators to edit the metrics as the community provides additional feedback.
3. Currently, the permission schema underlying the DQM System creates four types of users:
    - Public: public users can view metrics and visualization, and comment on the metric discussion boards;
    - Authors: ability to author metrics;
    - Submitters: ability to submit measures;
    - System Administrators: ability to assign any of the permissions listed here to users, review submitted metrics and publish them, and suspend or delete submitted measures.

    A user in the DQM system can be one or more of the above user types. In order to implement this system, more granular governance will be necessary. Additional user types, and more restricted access to the visualizations will likely be needed in a future system.

    In addition to these system enhancements, the stakeholder engagement process generated many additional enhancements and features that could be implemented in future work. The list of those enhancements is available in Appendix D.

VII. GLOSSARY

- **Data quality (DQ)**: describes various characteristics of a specific data source; these characteristics do not represent value judgements, but rather agnostic measures for use by researchers to help assess a data source's fitness for use
- **Data Quality Metrics System**: web-based system with accompanying visualizations that provides a harmonized data characterization toolkit, based on the framework put forth by Kahn et al., to enable researchers to efficiently compare data sources to better contextualize data quality and fitness-for-purpose
- **Data quality Metric**: describes quantitative measurements that characterize a specific aspect of the source data in a data model agnostic way
- **Data quality Metric standard**: DQM system contains a flexible, reusable set of Metrics that are intended to characterize aspects of data in a manner that is consistent (standard) across sources
- **Data quality Metric data model**: underlying data model to the DQM system that enables the capture of information on a contributing data source; is compatible across any Common Data

Model; captures information related to the Metric of interest, Measure data, and metadata about the execution and source

- **Data quality Metric metadata**: information that describes a Metric and enables users to execute locally; includes information such as: description, expected results, results type, domain, DQ Harmonization Category, etc.
- **Data quality Measures**: a numeric representation of a metric that has been executed against a data source and metadata on the data source
- **Data quality output:** the numeric output generated by executing a Metric locally and uploading Measure data into the DQM system; enables exploration and characterization of a data source
- **Data source metadata:** information that describes the data source, such as the organization, data set date range, and technical environment, as well as details on when the metric was executed; submitted alongside the Measures
- **Data quality harmonization**: defining data quality standards and Metrics in a general and harmonized fashion

VIII. APPENDICES

A. DISCOVERY AND DESIGN DOCUMENTATION

The following documentation describes the metadata standards and relevant use cases, technical specifications for implementing the standards, and a dictionary describing each metadata element. The document also includes information about the data quality data model that underlies that DQM web system, found here:
https://dataquality.healthdatacollaboration.net/

Data Quality Metrics Project

Discovery and Design Documentation

Prepared by: Sentinel Coordinating Center

SUBMITTED: NOVERMBER 30, 2018

UPDATED: December 31, 2019

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I. This project was funded by the FDA through HHS Mini-Sentinel contract number HHSF223200910006I. This work was supported by the Office of the Secretary PCORTF under Interagency Agreement #750016PE060001.

Data Quality Metrics Project

Discovery and Design Documentation

I. EXECUTIVE SUMMARY

The goal of this project is to provide a harmonized approach to data characterization across multiple data sources to enable researchers to better understand candidate data sources before querying and analyzing them. This work includes the creation of a system that operationalizes existing data quality (DQ) parameters and methodologies in a way that is compatible across multiple Common Data Models (CDMs) to increase research planning efficiency and improve the interpretability of analytic results. We will create and implement a set of metadata standards and metrics describing: 1) Data quality and characteristics; 2) Data sources and institutional characteristics; and 3) Fitness-for-use. These standards will be the basis for a flexible data quality collation system that is able to incorporate data metrics from any data source. The system will be designed to enable flexible exploration of DQ characteristics for multiple data sources at the same time. Importantly, the project will provide an open-source, web-based platform for exploring and describing the quality, completeness, and stability of data sources.

The project is organized into three phases: (1) Discovery & Design, (2): Development & Testing, and (3) Implementation & Release. The key deliverables from this project include a generalizable set of metadata standards and technical specifications for implementation. Together, the information described will provide a standardized data source "fingerprint" that can be expanded to provide additional granularity. Additionally, a system to maintain and query the data model will be implemented and available as open source technology such that the system will provide approaches to access the data model and can use any business intelligence tool of choice to interact with the data. A stakeholder group drawn from communities of interest will provide guidance on how this project can take advantage of existing data quality frameworks and standards and offer periodic review of work to date.

This Discovery and Design report is intended for technical stakeholders who have expertise in electronic health data resources and/or software development processes.

II. BACKGROUND

The first set of deliverables for Phase 1 includes a document describing requirements and use cases, design for a proposed set of metadata standards, technical specifications for implementing the standards, and a data dictionary. **This document contains the project deliverables for Phase 1: Discovery & Design Objectives 1-3.**

To date we have articulated 78 **use cases** to support development of the data quality metric data model and open-source toolkit (the DQM system). The Phase 1 Report includes 22 items of interest (metadata) describing a source system and 12 items of metadata describing each Metric. After receiving feedback, the implemented DQM system captures 25 items of interest (metadata) describing the source data system and its measures, as well as 15 items of metadata describing each metric. This information will form the basis for a data dictionary and for the technical implementation specifications. Based on the

use cases and review of current data quality standards, we identified the following structures to contextualize the quality of data:

- Time component (e.g., number of encounters by clinical setting per year)
- Person-based construct (e.g., number of prescriptions ordered per person per year)
- External context (e.g., rates of asthma by age compared to expected population rates)

From a **design** perspective, we developed a system architecture and a data model, including a draft architecture of an ideal end state. The architecture describes the components of the system, the agents (investigators, data sources, administrators, etc.) that will engage with the system, and the transactional relationships among the components.

The second design component is a fully extensible data model that will hold data source metrics and related metadata. The proposed set of standard metrics are intended as a starting point, with the system designed to be expanded by the community of users. Accordingly, the data model can accommodate virtually any metric proposed by the community. Validation of the model through iterative prototyping has started and will continue throughout the system development and implementation phases.

Key challenges faced during Phase 1 include the following:

- Designing a system that can capture virtually any data metric imaginable for any data source is a significant design challenge. Our experience with creating a flexible metadata model in the Cross-Network Directory Services (CNDS) project has given us a solid background for undertaking this challenge, which we believe our data model meets [2].
- Developing a generic set of data quality metrics applicable to any data source also has been a challenge, requiring a technical approach that can meet current and future requirements.
- A project objective is to enable investigators to browse data quality metrics in a simple and intuitive fashion. We intend to meet this challenge by providing a data visualization application using a freely available and high-quality business intelligence tool. Designing this visualization application is one of the more novel aspects of our implementation.

Although there are important challenges to overcome, the potential benefits of standardizing how data sources are characterized and “fingerprinted” presents substantial potential value. Differences in how data are collected and represented in different distributed research networks makes it difficult for investigators to judge the fitness of a data source for a research project. The Data Quality Metrics (DQM) system will go a long way towards addressing that problem. For example, despite the high concordance between the structure of the Sentinel and PCORnet common data models, significant differences exist at the data level – different definitions of gender, race, and ethnicity; differences due to data originating from health insurer claims versus provider-based EHR systems; differences in frequency of update; differences in what an “encounter” means.

Noteworthy is the leverage created by the combination of the ASPE funded Cross-Network Directory Services (CNDS) project and this Data Quality Metrics project. While CNDS provides the ability to find data sources that might be of interest based upon data provenance (EHR, claims, registry, etc.), types of

information (demographics, encounters, lab results, etc.), and other factors identified by the research community, DQM enables drilling deeply into the characteristics of data sources.

We note the close alignment of the goals of this project with the Department of Health and Human Services (HHS) objectives at several levels [3]. The results of this project address two of HHS's strategic core research functionalities:

- Use of Clinical Data for Research
- Use of Enhanced Publicly-Funded Data Systems for Research

The project addresses two HHS developmental components [4]:

- Services: Resources that researchers can use to capture, store, link, analyze, or exchange data or evidence.
- Standards: Nationally accepted specifications that have been widely approved and adopted because of market forces, community consensus, or regulatory requirements.

The project applies to four data sources prioritized by HHS[4]:

- Patient Contributed Data
- Clinical Data
- Administrative claims data
- Other: metadata on administrative claims, clinical, and EHR data maintained by healthcare organizations

Finally, the project supports several HHS milestones[4]:

- Support the development of a set of research Common Data Elements (CDEs)in specific gap areas and support development of a governance structure for CDE harmonization.
- Support the development of repositories/portals for CDEs, standards for utilizing CDEs for research, and services to allow researchers to easily utilize standardized components.
- Support the development of a core set of standards for the collection and integration of prevalent use cases of PPI for PCOR, by leveraging existing standards and filling gaps.
- Develop standards that support secure, electronic query of structured data across clinical research and delivery systems, including standards for open-source access.
- Establish services and tools to support data access, querying, and use, including privacy-preserving analytics and queries. These services and tools would be leveraged nationally and are not likely to be developed by the private sector.
- Develop and test metadata standards that describe data quality.
- Develop support services and tools that can be leveraged nationally and are not likely to be developed by the private sector; these tools would test the quality of unstructured and structured data to answer PCOR questions.
- Support the further development of key federally-initiated data systems for research.
- Support the enhancement of strategic publicly-funded data systems (including CMS data) to facilitate their access and use, and ease retrieval of data for research purposes.

III. SUMMARY OF PROJECT OBJECTIVES

The project is organized into three phases: (1) Discovery & Design, (2): Development & Testing, and (3) Implementation & Release. The key deliverables for each project phase are described below and summarized in Table 1.

A. PHASE 1: DISCOVERY AND DESIGN

Phase 1 has five objectives.

OBJECTIVE 1: A document, in a form suitable as a stand-alone report, describing the proposed set of metadata standards and relevant use cases, technical specifications for implementing the standards, and a dictionary describing each metadata element

Identification and documentation of metadata standards and design and technical specifications:

This project will utilize the Harmonized Data Quality Framework put forth by Kahn et al.– which defines data quality standards and metrics in a general and harmonized fashion – and apply it to a variety of data sources and research needs[1]. Operationalizing that framework and developing a tool for analyses will allow researchers to evaluate data quality (DQ) at any life stage of a data source in a consistent manner, and to effectively compare data sources based on the same metrics.

A flexible data quality metric data model will assist researchers in determining fitness-for-use of various data sources and research purposes.

Deliverables for this objective enable access to a standardized data dictionary for participating organizations and researchers to write transforms that load their data into the DQ proposed model in a prescribed format.

Key deliverables include documentation of proposed generalizable data quality metrics and relevant use cases, technical specifications for implementation, a data dictionary, a data model, a list of key stakeholders, and results of stakeholder reviews.

OBJECTIVE 2: Develop a data model that will illustrate how the data and information will be managed once the system is implemented

The data model will enable independent development of tools to query and view the DQ output within and across contributing sites. The tools can be customized based on the internal standards of organizations that utilize them or evaluated against other data quality frameworks. The model is being developed under the assumption that the data holders will execute the DQ tests or measures per the data dictionary provided and those results will be transferred and stored in a DQ relational database management system (RDBMS) (e.g. SQLServer or PostgreSQL). These data will be used to populate secure, interactive web-based visualization dashboards (e.g. Qlik Sense) of participating data partners so that researchers can evaluate fitness-for-use of sources they are interested in utilizing.

Key deliverables include a logical data model that is designed to be portable to virtually any quality related checks or rules applied to an e-health data source.

OBJECTIVE 3: All project requirements and specifications will be captured in the system used to manage software development projects (JIRA)

Defining requirements will follow an agile approach and evolve during development iterations and user feedback. System documentation and related artifacts (e.g. implementation and validation details and release notes) will be made publicly available.

OBJECTIVE 4: List of key stakeholders who will vet the proposed metadata standards. Representatives from funders of major research networks will be included as stakeholders, including FDA, PCORI, and NIH, as well as others as possible, given the time constraints and resources

The project team will collaborate with internal and external stakeholders to ensure that metrics of interest are sufficiently captured and documented, which will inform final technical specifications.

OBJECTIVE 5: Summary of stakeholder comments and disposition of comments for the proposed metadata standards. Public domain dissemination via meeting presentations is an option

Stakeholder engagement

Key deliverables for this objective include providing a list of key stakeholders who will vet the proposed metadata standards. Representatives from funders of major research networks will be included as stakeholders, including FDA, PCORI, and NIH, as well as others as possible given the time constraints and resources.

Stakeholder engagement summary

Key deliverables for this objective include a report summarizing stakeholder comments and disposition of comments for the proposed metadata standards.

Table 1. Summary of Key Deliverables

Deliverable	Completion Date	Status
Project Initiation	1/1/2017	Complete
Phase 1: Discovery & Design		

Deliverable	Completion Date	Status
Objective 1: A document, in a form suitable as a stand-alone report, describing the proposed set of metadata standards and relevant use cases, technical specifications for implementing the standards, and a dictionary describing each metadata element.	11/30/2018	Complete The boundaries of the system have been defined. This is a dynamic report that will be continually updated throughout Phases 2 & 3
Objective 2: Develop a data model that will illustrate how the data and information will be managed once the system is implemented.	10/29/2018	Complete
Objective 3: All project requirements and specifications will be captured in the system used to manage software development projects (JIRA).	6/30/2018	Complete Initial design documents captured in JIRA. Defining requirements will follow an agile approach and evolve during development iterations
Objective 4: List of key stakeholders who will vet the proposed metadata standards. Representatives from funders of major research networks will be included as stakeholders, including FDA, PCORI, and NIH, as well as others as possible given the time constraints and resources.	12/03/2018	Complete
Objective 5: Summary of stakeholder comments and disposition of comments for the proposed metadata standards. Public domain dissemination via meeting presentations is an option.	10/29/2019	Complete Following approval of an extension, the project team presented to stakeholders during four sessions in September 2019.

B. PHASE 2: DEVELOPMENT & TESTING OF METADATA STANDARDS AND SYSTEM WITHIN AT LEAST TWO DISTRIBUTED RESEARCH NETWORKS

In Phase 2 we will demonstrate our "data fingerprinting" system using synthetic data sets that reflect those used by existing networks, such as PCORnet, ESP, and Sentinel, as well as consider how our system can be used by an open network where anyone can review, contribute to, and utilize the DQ data model and explore database fingerprints approved for public consumption—a priority interest for the NIH community and others [5-9].

Key deliverables include testing of the implemented system from at least 2 distributed research networks; a data model to consume metrics from data partners; demonstration of functionality through beta testing, quality assurance and user acceptance testing; and a summary report of testing results.

C. PHASE 3: IMPLEMENTATION & RELEASE CULMINATION OF PROJECT PHASES.

Following iterations of testing and any necessary changes to the functionality, all documentation and software will be made available to the open source community for review and implementation.

Key deliverables include publication of the open source software production release with companion technical and user documentation on a publicly accessible platform (i.e., HealthData.gov or other option based on consultation with FDA).

IV. PROJECT METHODOLOGY

System Requirements & Design:

Due to the overlapping nature of the 3 project phases, the team worked concurrently on activities related to all phases. Accomplishments to date from the Discovery and Design Phase pertain to requirements and design; we have undertaken prototyping activities as well. We are ensuring that we take enough time with system design so that the software can be placed in the public domain and configured to best use by any party and for any data source.

As part of the use case development and at FDA's request, we also summarized relevant publications since the publication by Kahn et al. that provides the conceptual framework for our model of data quality and characteristics metrics [1].

Regarding the implementation activities, the primary format for capturing requirements is the Use Case. A use case is a readable description of one specific metric of data quality or one specific characteristic of interest for a data source.

Prototyping:

To validate the data model design, we undertook two activities. Early on we built a scaled-down version of the data model so we could explore interactions among the various model elements (type of common

data model, type of metric, where each metric fits in the Kahn framework, units of measure for metrics, etc.). This activity was helpful preparation for engaging a data modeler.

We have engaged a data modeler who has recommended a logical model for capturing metrics, basing his work on the use cases that we developed. We believe the model is now complete and are validating the model by creating data metrics based on test data sources available to us. One such test data source adheres to the PCORnet common data model and another adheres to the Sentinel common data model. We also can create synthetic data sources in any model format we wish.

See the Project workflow: design-to-implementation section for details on our processes to meet the project objectives.

V. REQUIREMENTS

A. DATA QUALITY METRICS

We propose a pragmatic approach to developing consistent data quality metrics through development of an extensible data model based on a collection of data quality standards and metrics included in the Harmonized Data Quality framework put forth by Kahn et al. An extensible data quality data model must be flexible and independent of the source data model.

The project team has determined, through iterative discussion and an exploration of current data quality standards, the following structures that are necessary to contextualize the quality of data:

- Time component (e.g., number of encounters by clinical setting per year)
- Person-based construct (e.g., number of prescriptions ordered per person per year)
- External context (e.g., patient distribution by race vis-à-vis racial distribution nationally; counts of patients with diabetes compared with HbA1C lab test result distribution across all patients in a database)

As illustrated in Figure 1, this project will utilize the Kahn framework, which describes and defines data quality standards and metrics in a general and harmonized fashion and will apply it to a variety of data sources and research needs. Operationalizing that framework and developing a tool for analyses will allow researchers to evaluate data quality at any life stage of a data source in a consistent manner, and to effectively compare data sources based on the same metrics. A standard data quality metric data model will assist researchers in determining fitness-for-use of various data sources and research purposes. We are developing a DQ data model with flexible and extensible framework to allow data sources to utilize analytic tools irrespective of the CDM the data source adheres to in its local environment. We will demonstrate our “data fingerprinting” system using synthetic data sets that reflect those used by existing networks, such as PCORnet and Sentinel, as well as consider how our system can be used by an open network where anyone can review, contribute to, and utilize the DQ data model and explore database fingerprints approved for public consumption— a priority interest for the NIH community and others[5-9]. We will continue to collaborate with DQ stakeholders and share our work and experience throughout the project.

A consensus on priority data quality metrics and, more importantly, a data model design for DQ that allows effective comparison of data sources will allow researchers and organizations to better understand their own data quality and establish the fitness-for-use of data sources based on the same DQ processes, as opposed to the comparison of study-specific data characterization and quality assessments. We aim to design and conduct a reference implementation that demonstrates a novel pragmatic approach to data quality that can be broadly used across nearly any data source and industry and that can be used either locally or in a distributed network.

Although several groups and researchers have done thorough evaluations of DQ metrics for specific data sources (e.g., birth defect surveillance systems, primary care data, medical registries), to our knowledge there is not currently a data model in place for generic quality measures that can be tailored to specific data sources [10-15]. While study-specific data characterization work provides a framework to evaluate data, it lacks a focus on extensibility and generalizability. Our model will enable users to add any data quality metric of value from their work, thus expanding the initial DQ metrics included in this reference implementation.

Our data model will accommodate a variety of DQ metric types and patterns that can be applied to established common data models (e.g. PCORnet, Sentinel, other health databases), and represent the DQ framework categories and types of metrics described by Kahn et al. We will use a sample set of metrics in this project, as described in the Use Cases Section.

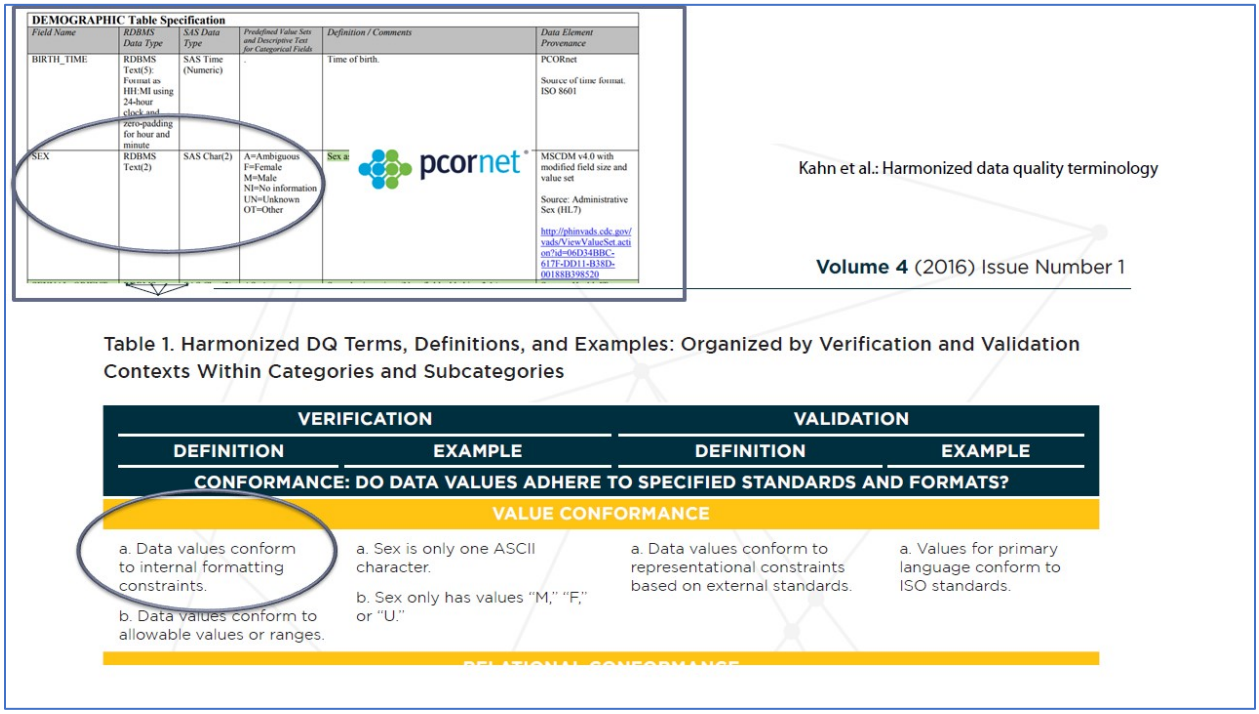


Figure 1. Example of a PCORnet data element and related Kahn DQ term

We are designing the system components (e.g. DQ metrics, data model, web portal, fingerprinting / visualization tools) using the Kahn framework and a template to describe the metrics using a parsable format that can be re-used and is data source independent. An example is described below.

> **Use case:** User needs to examine how PCORnet data partners adhere to the CDM specifications for the Sex field.
> One metric executed would check that the data respect the field length max of 2 characters.
> **Kahn Metric Type:** Value Conformance for Internal Formatting
> **Metric Name:** Field Length Metric
> **Metric details:** In a [PCORnet] **data source** review the [SEX] **column** in the [DEMOGRAPHIC] **table** and report the [Percent] **calculation** of records that do not adhere to the [2 character] **field length rule**
> This metric is designed so that anything in brackets would be modular and could be adapted for any data source that has field length rules.

The model is organized around a central table that captures measurements (counts of patients, maximum or minimum values, frequency of values, etc.) and surrounded by tables that identify, for each measurement, the source system, context (patient, member, encounter, claim, etc.), any relevant stratifications (age ranges), and other important qualifiers.

Data Model Features

WHO?

- Researchers can describe themselves, their organization, the network, and types/rules of data they have.

WHAT?

- They will also know what kinds of metrics they'd like to run and the concepts that are important (according to Kahn's framework).
- The question is described, not to the level of executable SQL code but enough so that anyone working with the database could develop the query to generate the metrics for their data source.
- Then, someone runs a metric and sends the results, i.e. measures, back. The information we want to know about the data/results is the response - who ran the metric and when plus the actual answer to the query (the counts).

WHY?

- We get all the measures and then we have an ability to do calculations/analyses on the counts to answer more specific research questions.
- Our model also covers the why piece - we describe our metrics/reasons for them and have documentation of metadata about the metric.

As illustrated in Figure 2 below, there is a need to explore and understand various characteristics about electronic health data sets.

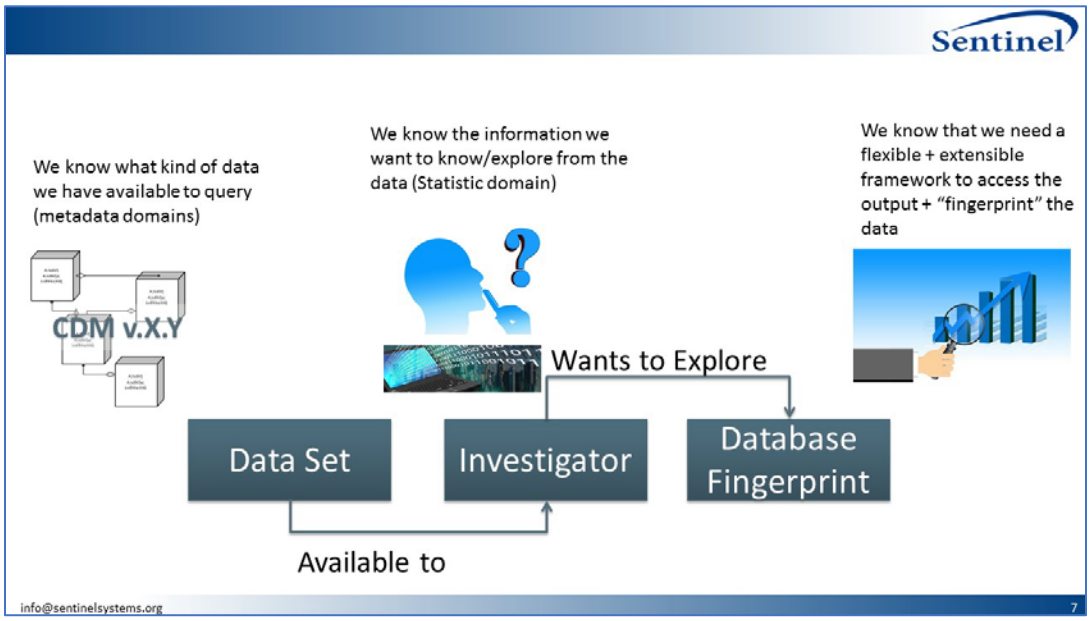


Figure 2. Concept Model

In Figure 3. The swim lanes represent the three key business processes: (1) catalog the DQ metrics; (2) execute metrics; (3) characterize the database(s)

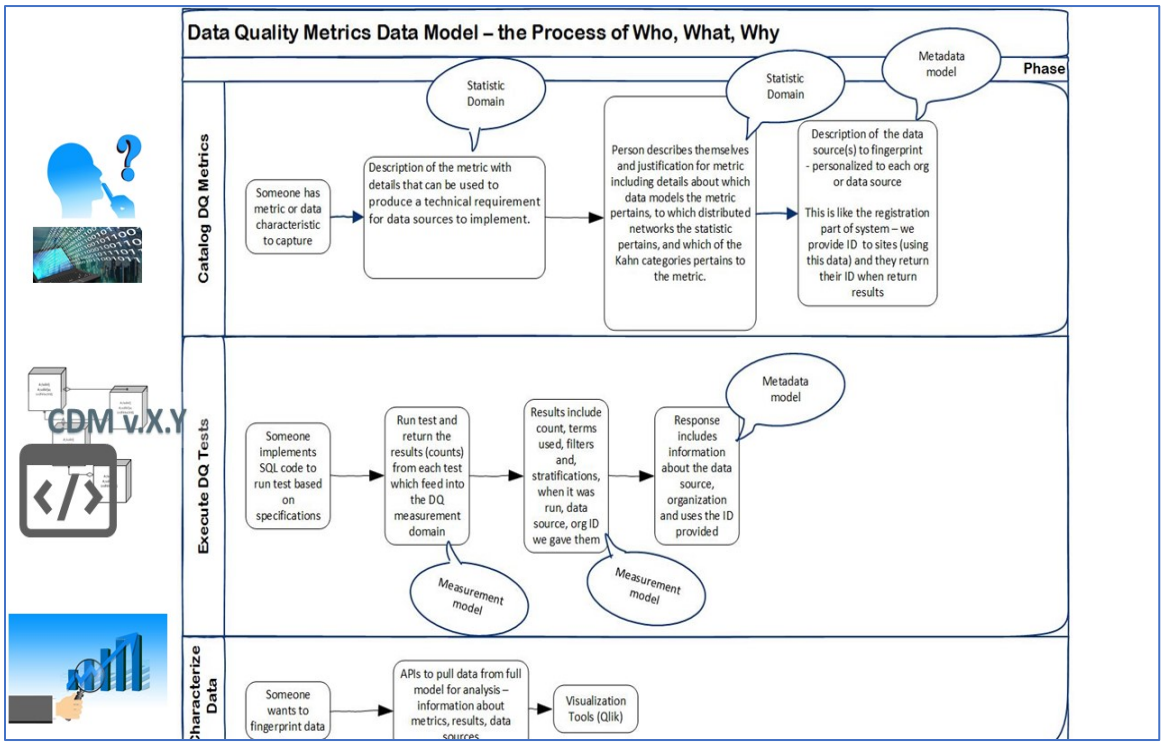


Figure 3. Workflow

Data Model Domains:

The data model domains are described in Figure 4.

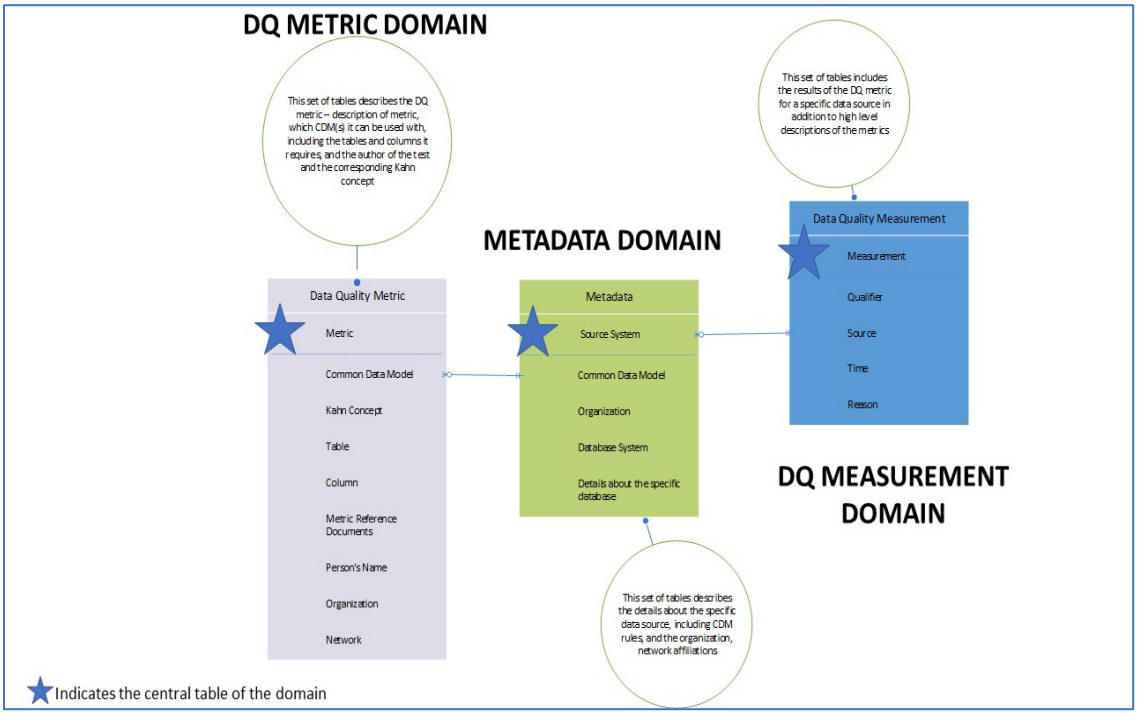


Figure 4. Data Model Domains

Metric Domain:

- The catalog that provides the opportunity for users to define some performance metrics or data checks that they think is worthwhile to capture and document.
- There could be several versions of the technical requirements since one metric could be applied in several ways to different contexts, e.g.:
    - If the metric/characteristic relates to the demographic profile of patients/members, we could have views from race, ethnicity, age, gender, and other perspectives.
    - If the metric/characteristic relates to cyclical variations in medical encounters, we could have views from the perspective of healthcare setting (inpatient, outpatient, OR) and date/time (month of the year, time of day).

Measurement Domain:

- Measurement is the fact table of our use cases and stores results in the form of counts.

Metadata Domain:

- Information about the entities captured in the DQ catalog (e.g. organizations, data sources, networks, CDMs)

VI. IMPLEMENTED DATA MODEL

Following iterative design discussions, a final data quality data model (Figure 5) was implemented as the underlying structure of the system.

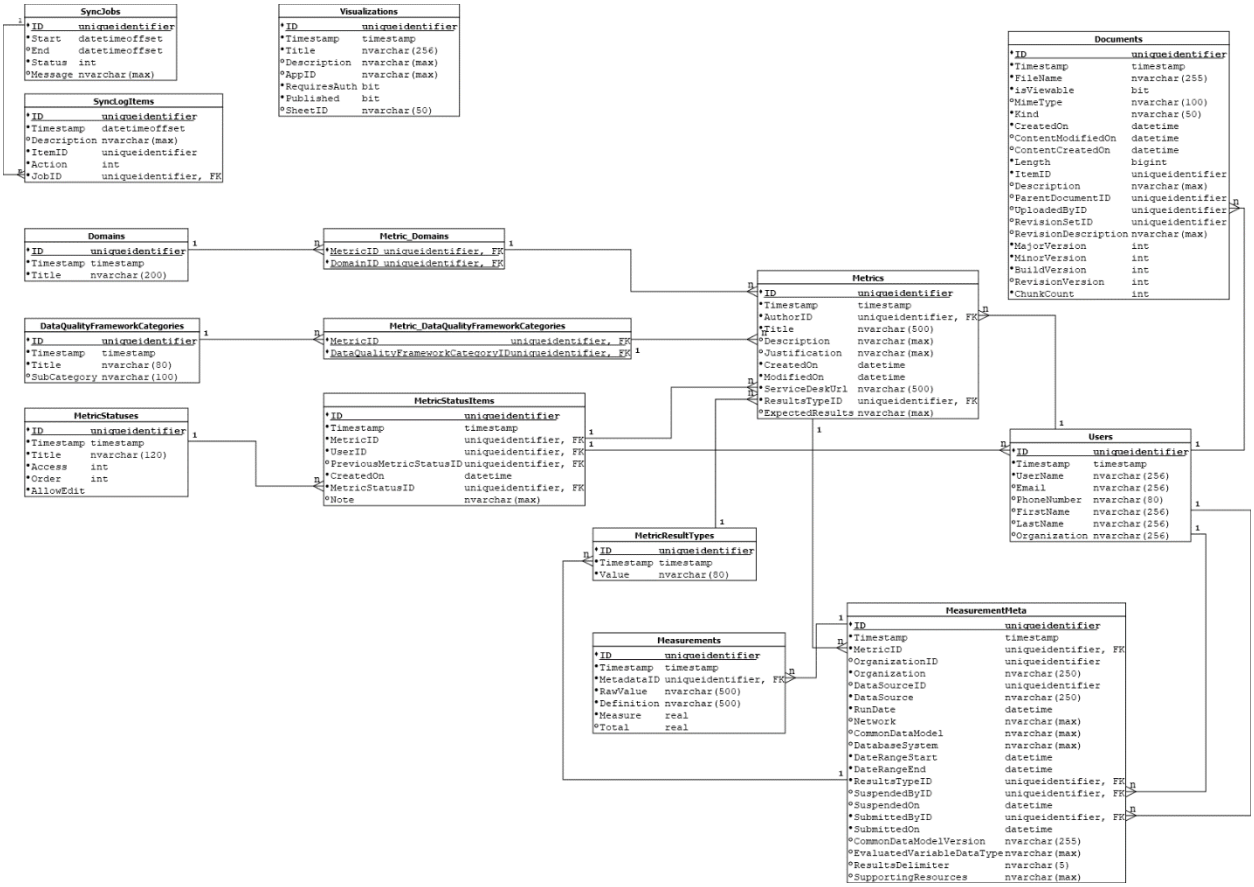


Figure 5. Data Quality Data Model

- Solid dots indicate non-nullable fields.
- Underlined fields indicate Primary Keys.
- Relations are indicated by the connecting lines and their connectors.
- All non-collection tables have a primary key that is named ID.
- A non-nullable timestamp field is included on all tables that require optimistic concurrency for Entity Framework.

The root entities are Metric and Measure Metadata; all other entities support defining attribute of those entities. Entity relationships are depicted below in Figures 6 and 7 and detailed in Table 2.

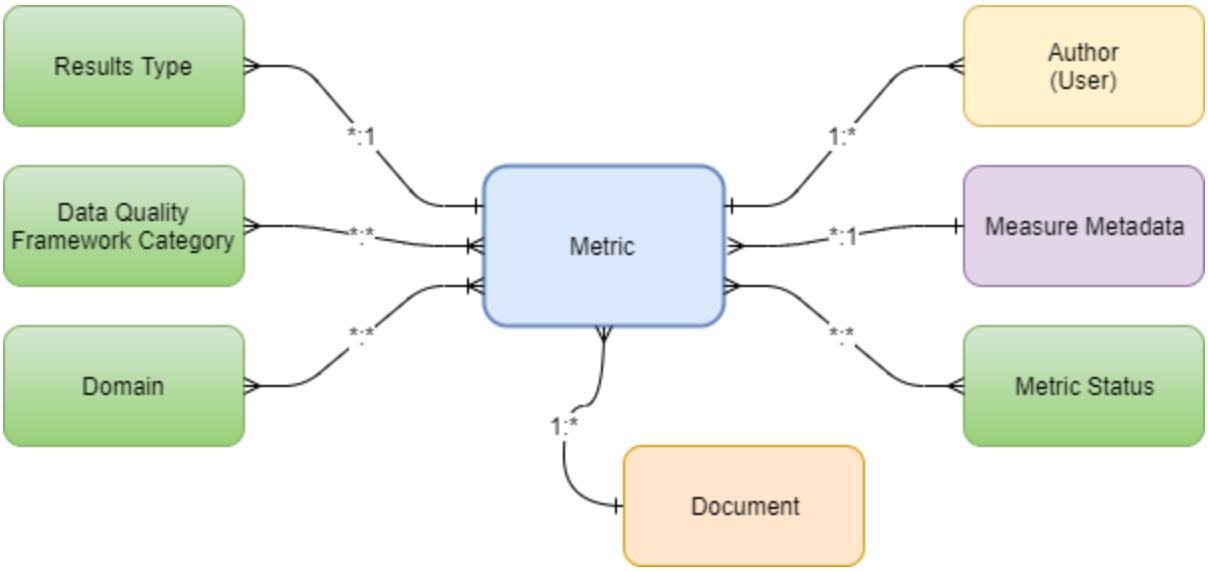


Figure 6. Metric Entity

- A User can author zero or more Metrics. A metric must have an author.
- A Metric has a collection of statuses, each status item is immutable. A new status item is created for each status change, the most current item is the current status of the Metric. A metric status item contains the date the status changed, the status, the User that changed the status, a reference to the previous status item, and an optional note regarding the status change.
- A Metric has a single Results Type association. A Results Type can be associated to more than one Metric.
- A Metric has one or more Data Quality Framework Category associations. A Data Quality Framework Category can be associated to more than one Metric.
- A Metric has one or more Domain associations. A Domain can be associated to more than one Metric.
- A Metric has zero or more Measure Metadata associations. A Measure Metadata must be associated to a Metric.
- A Metric has zero or more Document associations. A Document must be associated to an entity.

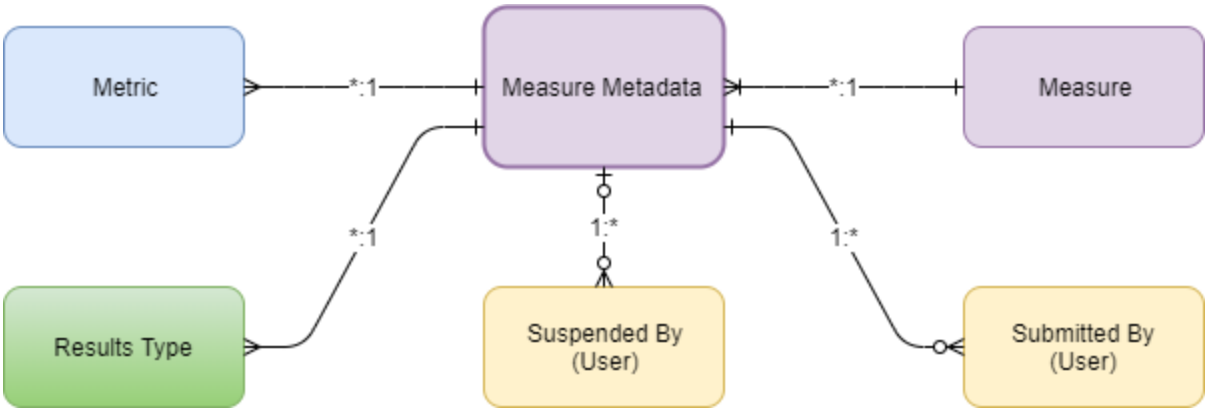


Figure 7. Measure Metadata Entity

- Measure metadata represents the metadata about a collection of measures.
- Measure metadata must be associated to a single Metric. A Metric can be associated to more than one Measure metadata.
- Measure metadata must be associated to a single Results Type. A Results Type can be associated to more than one Measure metadata.
- Measure metadata is associated to more than one Measure. A Measure must be associated to a single Measure metadata.
- Measure metadata must be associated to a single User representing who submitted the measure data. A User can be associated to more than one Measure metadata as the submitter.
- Measure metadata may have an association to a single User representing who suspended the measure data. A User can be associated to more than one Measure metadata as the suspender.

Entity	Details
User	• Represents a "person" • Requires a User Name. Optionally: a first and last name, email address, phone number, and associated organization name
Results type	• Indicates the Results Type of a Metric, and/or Measure • Comprised of a display title • Can be associated with many Metrics
Data Quality Framework Category	• Indicates the category a Metric could be classified as • The category classifications are based on definitions defined by the Khan framework. • Comprised of a Title and optional Sub-category • Can be associated with many Metrics
Domain	• Indicates the domain a Metric belongs to • A domain is comprised of a title. • Can be associated with many Metrics

Metric status	• The definition of a status a Metric can be assigned
	  • Comprised of a title, an access level, a logical order value, and if editing of the Metric is allowed while in the status
	• The access levels define which users have access to a Metric, and are comprised of the following values:
	  • None = no access level specified
	  • Author = only the author of the Metric has access
	  • System Administrator = only Users with the System Administrator claim can access the Metric
	  • Authenticated Users = only Users who have been authenticated can access the Metric
	  • Public = any User can access the Metric

Metric status item	• The instance of a status for a Metric • Comprised of the Metric, User, Metric Status, the previous Metric Status, Creation date, and a note

Entity	Details
	• A Metric will have one or more status items; the one with the most recent creation date is the current status.

Metric	• The definition of a Metric is comprised of:
	  • Title, Description, Justification, Expected Results, Created On and Modified On dates, Service Desk URL
	  • An Author - the User creating the Metric
	  • Results Type
	  • One or more Data Quality Framework Categories
	  • One or more Domains
	  • One or more Metric Status Items
	  • Zero or more Measures (Measure Metadata)

Measure Metadata	• Represents the metadata about a collection of Measures
	• The definition of a Measure Metadata is comprised of:
	  • A Metric; Measures are the quantitative result of a query based on a Metric definition
	  • Organization name, and optionally it's ID
	  • Data Source name, and optionally it's ID
	  • A run date for when the data was collected
	  • The network the Data Source belongs to
	  • The Common Data Model the data may belong to
	  • The Database System the data was stored in
	  • Date Range Start is the earliest date of the data set
	  • Date Range End is the latest date of the data set
	  • Results Type ID, the ID of the Results Type associated to the Measures. Must match the Results Type defined on the associated Metric.
	  • Suspended By, the User who suspended the Measures excluding it from available queries
	  • Submitted By, the User who uploaded the Measures to DQM
	  • Common Data Model Version, the version number of the CDM the Measure data may belong to
	  • Results Delimiter, the delimiter used if the values of the Measures are compounded and the result of more than one value.
	  • Supporting Resources, a URL to a location providing resources (application, scripts, documentation, etc.) used to obtain the measures.
	  • A collection of one or more Measures

Measure	• Represents the instance of a Measure
	• Comprised of:
	  • Raw Value represents the unmodified value of the stratifier that the measure is for
	  • Definition represents a display value for the Raw Value: i.e. Raw Value = 'M' and the Definition = 'Male'
	  • Measure is the numerical quantity of the result. Depending on the Results Type defined by the Metric it could be a count, percentage, range, or vector.

Entity	Details
	• Total is the optional value representing the total of all the Measure values. It could be greater than the sum of the Measure values included.

Table 2. Entity details

VII. USE CASES AND METaDATA

Over 100 data checks were identified, which include metrics of interest for PCORnet, Sentinel, and other electronic health data sources which the system must accommodate. The implemented DQM system includes an additional 27 items for capturing metadata related to the data source system, the measures, and the metrics. More information on the implemented model can be found in the Technical Documentation. For the purpose of this implementation, we will select 3 representative metrics for implementation and testing from the list of metrics below.

1. Number of patients by birth year?
2. Number of patients with an age less than zero?
3. Number of patients with an age greater than 120 years?
4. Number of patients with an age greater than 85?
5. Total number of encounters?
6. Total number of encounters by year and month-year?
7. Number of inpatient encounters per year and month-year?
8. Number of emergency department encounters per year and month-year?
9. Number of outpatient encounters per year and month-year?
10. Number of all encounters per facility location?
11. Number of inpatient encounters per person per year?
12. Number of outpatient encounters per person per year per?
13. Number of emergency department encounters per person per year.
14. Number of medications dispensed per year?
15. Number of medications dispensed per patient?
16. Number of medications dispensed per patient per year?
17. Number of medications dispensed by the patient age group
18. Number of prescriptions written per year and month-year?
19. Number of encounters with a diabetes diagnosis by year and month-year?
20. Number of patients with diabetes diagnosis by year and month-year?
21. Number of inpatient encounters with a diabetes diagnosis by year and month-year?
22. Number of patients with a diabetes diagnosis in inpatient setting by year and month-year?
23. Number of outpatient encounters with a diabetes diagnosis by year and month-year?
24. Number of patients with a diabetes diagnosis in outpatient setting by year and month-year??
25. Number of records for the Race field?
26. What are the observed values for race?
27. Number of race values = null?
28. Number of race values = White?
29. Number of race values = Asian?

30. Number of race values = Black in refresh 1?
31. Number of race values = Black in refresh 2?
32. Number of Race = White with a diagnosis of diabetes?
33. Number of Race = Black with a diagnosis of diabetes?
34. Number of Race= Black with a diagnosis of diabetes by age group?
35. Number of Race= White with a diagnosis of diabetes?
36. Number of Race= White with a diagnosis of diabetes by age group?
37. What are the values for sex?
38. Number of patients with null sex?
39. Frequency of values for lab tests? (all possible lab tests recorded)
40. Distribution of HbA1c lab test results by HbA1c group?
41. Number of patients with a diagnosis of diabetes and also a HbA1c lab test result?
42. Frequency of diagnosis code types overall and by year?
43. Frequency of procedure code types overall and by year?
44. Count of encounters by diagnosis code.
45. Count of patients by diagnosis code.
46. Number of patients with an encounter with an ICD-9 diagnosis code that starts with 001 – 139?
47. Number of patients with an encounter with an ICD-9 diagnosis code that starts with 140-239?
48. Number of patients with an encounter with an ICD-9 diagnosis code that starts with 240 -279?
49. Number of patients with no value for birth date?
50. Number of patients with no value for race?
51. Number of patients with no value for sex?
52. Number of encounters with no admit date by encounter type?
53. Number of encounters with no discharge date by encounter type?
54. Number of patients who have an encounter, but no enrollment?
55. Number of encounters that do not have a value for code type by encounter type?
56. Number of encounters have a discharge date before an admit date by year and month-year?
57. Number of patients with an encounter after their death date?
58. Number of patients with a birth date after their death date?
59. Number of patients with enrollment start date after their death date?
60. Number of encounters with encounter dates in the future?
61. What system is used to store the source data? E.g. Oracle, SQL Server, etc.
62. Number of patients with no PATID?
63. Number of duplicate values for their PATID?
64. How many patients with a non-conforming value for PATID?
65. Number of patients with recorded blood pressure?
66. Frequency of Discharge Disposition for inpatient encounters.
67. Number of non-inpatient encounters with Discharge Disposition populated?
68. Distribution of length of stay (discharge date – admission date +1) for inpatient encounters?
69. Number of medications dispensed with a days supply of 0?
70. Number of medications dispensed with a days supply of less than 0?
71. Number of medications dispensed with a days supply of between 0 and less than 1?
72. Number of medications dispensed with missing days supply?

73. Number of medications dispensed grouped by days supply (1-30, 31-60, 61-90, 90-100, 100-999, 1000+)?
74. Number of medications dispensed with amount dispensed of 0?
75. Number of medications dispensed with amount dispensed of less than 0?
76. Number of medications dispensed with amount dispensed between 0 and less than 1?
77. Number of medications dispensed with missing amount dispensed?
78. Number of medications dispensed grouped by days supply (1-10, 11-30, 30-60, 61-90, 90-100, 100-999, 1000+)?

Metadata about the source system:

79. Network Affiliation
80. Common Data Model
81. Common Data Model Version number
82. Type of RDBMS where source data are stored
83. RDBMS version number
84. ETL Version
85. Source data as-of date
86. Organization ID
87. Min date by CDM table
88. Max date by CDM table
89. Count of total rows by CDM table
    a. Name of person who submitted the data for this ETL
    b. Email of person who submitted the data for this ETL

Metadata about each metric:

90. Unique identifier of the metric captured in our data model
91. Network the metric is associated with
92. CDM the metric is associated with
93. CDM version the metric is associated with
94. Date the metric was created in our data model
95. Person who authored the metric metadata
96. Organization that authored the metric metadata
97. Number of results we have for the metric
98. List of the tables associated with the metric
99. List of the fields associated with the metric
100. Free text describing the metric
101. Word or PDF document file for the metric

VIII. PROJECT WORKFLOW: DESIGN-TO-IMPLEMENTATION

The following diagrams represent:

- Major processes: Green rectangles representing action-oriented objectives being completed as a small step toward a milestone. Each green rectangle is a unique process to follow.
- Milestones: Yellow diamonds illustrating significant markers of progress, tied to project deliverables; represents a major decision point for next steps. Successful completion of

milestone leads to a next process or completed task; failure requires return to the process for modification.

- Deliverables: Blue ovals representing a quantifiable achievement

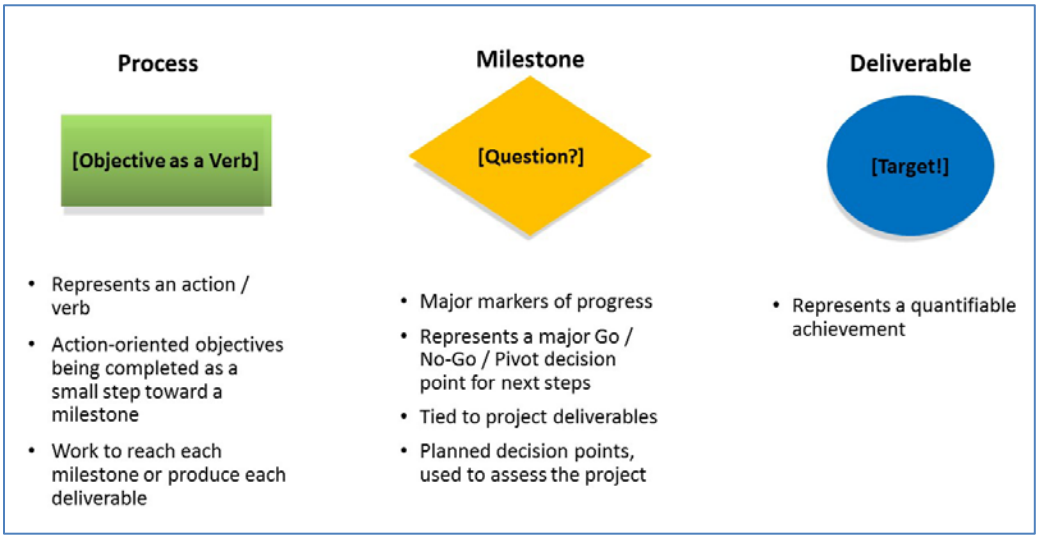


Figure 8. Project Workflow Legend

These artifacts are all captured and managed in the JIRA tracking application for this project and development work is coordinated using the tool JIRA is a market-leading commercial application purpose-built for software development teams. JIRA provides access control capabilities so that only authorized personnel are able to view, create, or modify JIRA items.

Figure 8. Illustrates the initial activities involved for the implementation phase of the project. The work includes defining functional specifications and designs using narratives, process diagrams, wireframes, mock-ups, user stories, providing test data, etc. Specifically, the requirements include describing what the system should do and how the system should do it. The major achievements are set-up of the back-end infrastructure and demonstrating that the selected DQ metrics can be captured per the specifications in the database via the DQM website. Development adheres to the agile development methodology. The essence of the agile approach is to keep development cycles very short, enabling nearly continuous adjustment as requirements and priorities change.

Agile methods convey three key benefits:

1. Because development sprints are short, there is frequent feedback to know whether the project is on track or not and can respond more quickly if corrective action is needed.
2. Because each sprint delivers a working component of the overall system, stakeholders see a regular stream of results, can react to design decisions, and generally feel a better sense of connection to the project.
3. The project can react swiftly to shifts of priority and emphasis that normally occur during the course of a major software implementation.

The high-level agile process for this project is illustrated in Figure 9. Once development items are defined and scoped, the JIRA issues are scheduled into two-week development cycles or sprints, followed by quality assurance test and user acceptance test (UAT) cycles of varying durations. As new functionality is made available, we can demonstrate any high value features to stakeholders and integrate feedback into subsequent development iterations. The major accomplishments from this implementation phase is to showcase the system to stakeholders throughout the project and deliver a functional system at the conclusion.

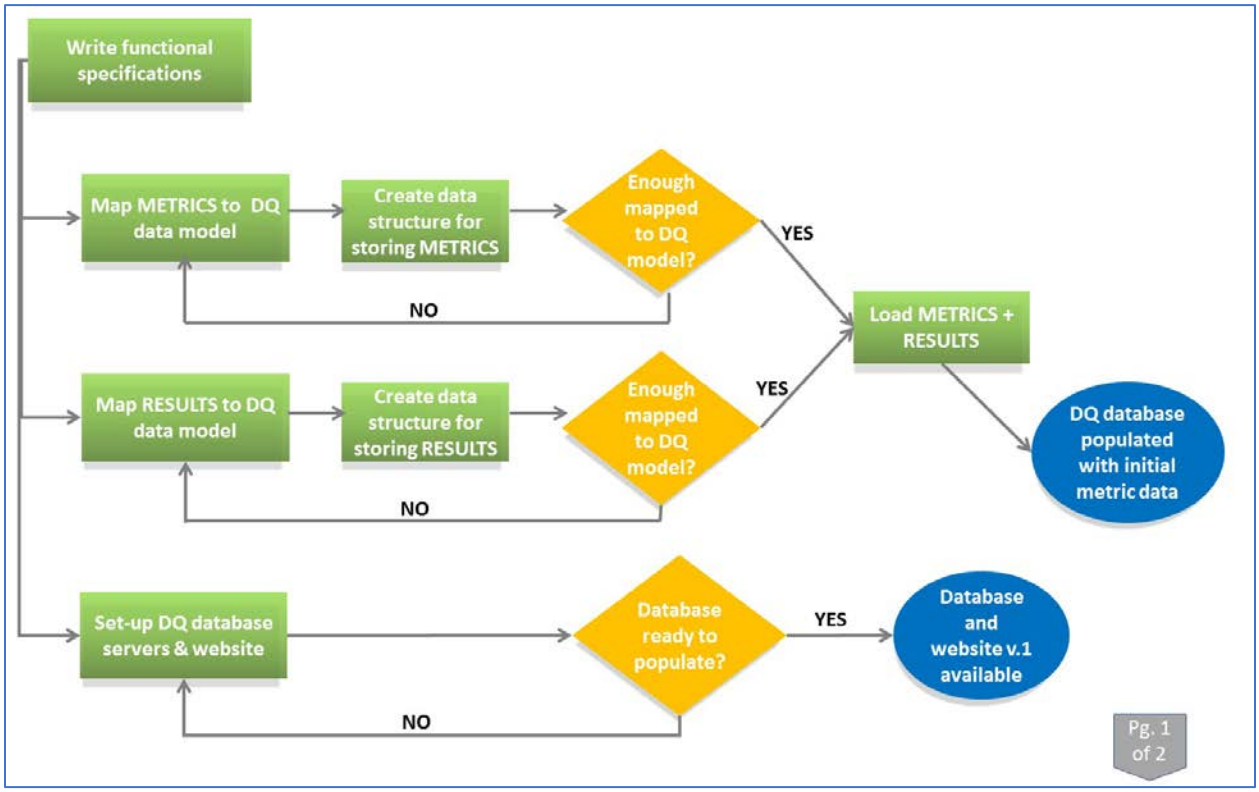


Figure 9. Project Workflow Diagram: initial activities to set-up DQM system environment

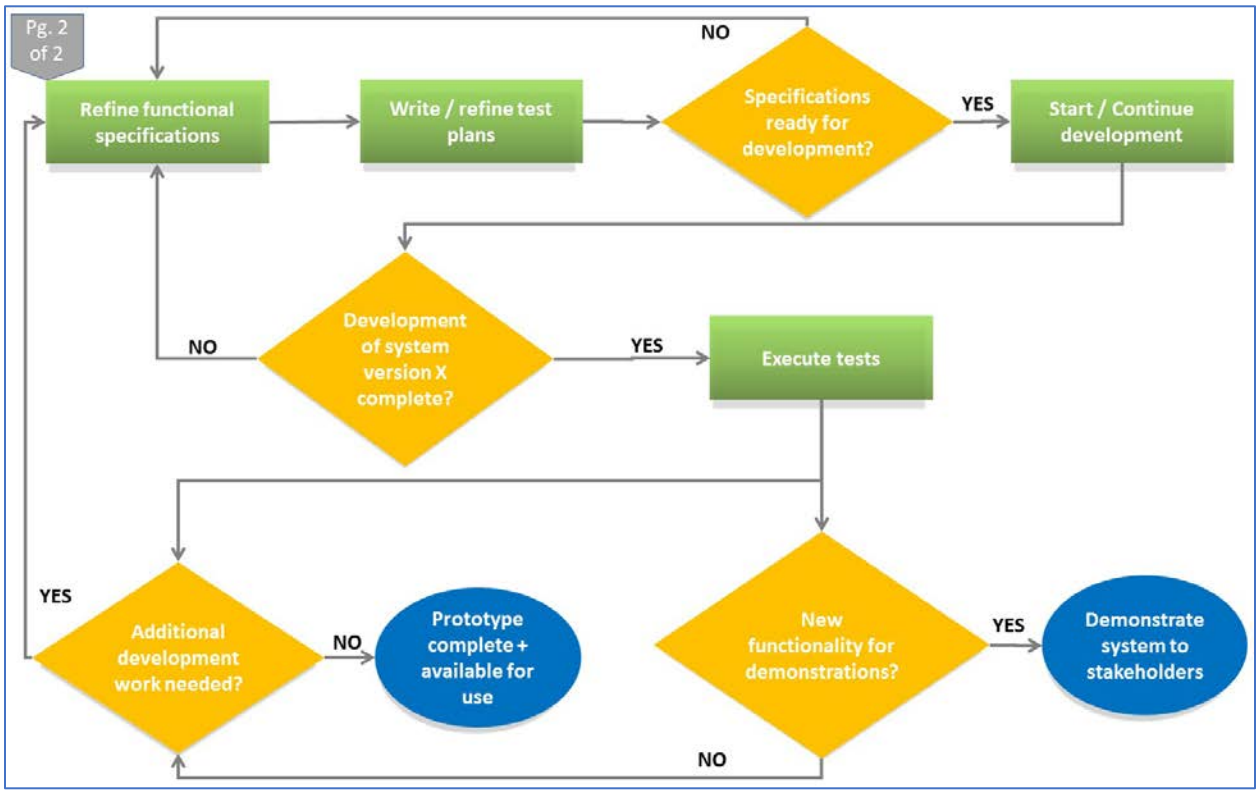


Figure 9. Project Workflow Diagram: initial activities to set-up DQM system environment

Technical specifications were developed throughout the design phase in collaboration with software developers and a data modeler. The team engaged in an iterative review and discussion of multiple resources, including the Kahn framework and existing data quality tools. The development of use cases framed the creation of the data model. In addition, the team met with subject matter experts on data storage and transport standards for e-health data.

IX. ARCHITECTURE OVERVIEW

The DQM system will use current web technologies to provide users with ability to import, store, and explore the output of DQ metrics produced from distinct data sets. Additionally, the web-based application will enable the creation, curation, and review of DQ metrics. Details on the key architecture components are described in the following sections.

At a high level, the infrastructure planned to support the application and information architecture includes the following:

1. Platform: We built the application web user interface and web APIs using the open source framework, .NET Core.
2. Database: Microsoft SQL Server will be used to store the DQ metrics and measure results in a database repository, the DQM Server, using DQ data model schema described above.
3. Communications: DQM measure results will be delivered to the DQM server via dedicated web service endpoints. As APIs will be used, the system will have the ability to

accept DQ metrics and / or DQ measure results directly via the API if desired. The web application interface will also provide a mechanism to manually curate DQ metrics and import measure results. Web services will also be implemented to transform and present the DQ metrics to end-users via the visualization tools.

4. The data structure for DQ metrics and measure results, later referred to as 'payload', will be codified to a common format that is not data model specific and allows for application portability and interoperability. JSON was selected as the language to express the metrics and results, though XML, BSON, or the next new flavor of structured data formatting would have been other options. Additionally, we are investigating the potential of leveraging parts of the data structure defined by the Fast Healthcare Interoperability Resources (FHIR) standards (https://www.hl7.org/fhir/overview.html). The FHIR standards are utilized for the transfer of electronic healthcare information based on existing logical models and can be extended for specific purposes. While we will not formally use FHIR services, there may be opportunities to structure the DQ payload in ways that align with current FHIR data structures.
5. Visualization: Qlik Sense was selected as the visualization tool for users to explore the characteristics of data sources. Qlik can connect to data sources using standard APIs, and the assumption is that other analytic tools able to load data via an API (e.g. Tableau) could be used in place of Qlik.

The following sections provide more information about the architecture of these components.

A. IN SCOPE

Activities considered in scope and related assumptions for this project include:

- We assume that the data quality metrics from each data source are received in a format that we defined and can consume.
- DQ metrics and results will be stored in a secure central repository
- We currently identified over 100 use cases to test the system. For the purpose of this implementation, we will select 3 representative metrics for implementation and testing.
- Metadata about data owners will be captured during the registration process via the related Cross Network Directory Service (CNDS) application and made available to this project via APIs
- For testing purposes, we will use Sentinel and PCORnet sample data.
- Investigate the potential for the use of Fast Healthcare Interoperability Resources (FHIR) Standards (https://www.hl7.org/fhir/overview.html)for this project. The FHIR standards are utilized for the transfer of electronic healthcare information based on existing logical models that can be extended for additional purposes aligned with lessons learned from previous HL7 implementations.
- The visualization / analytic tool used will be Qlik Sense
- Create a web-based system
    - Users with login credentials can access site and explore Qlik visualizations
    - Web portal will also contain spaces for:

        - Management of metadata and registration processes
        - Proposal of new metrics
- Types of users
    - Passive – view visualizations; submit feedback and propose DQ metrics
    - Administrative – organizational; administrates data
    - System Administrator – approval of metrics and potential management of user credentials

B. OUT OF SCOPE

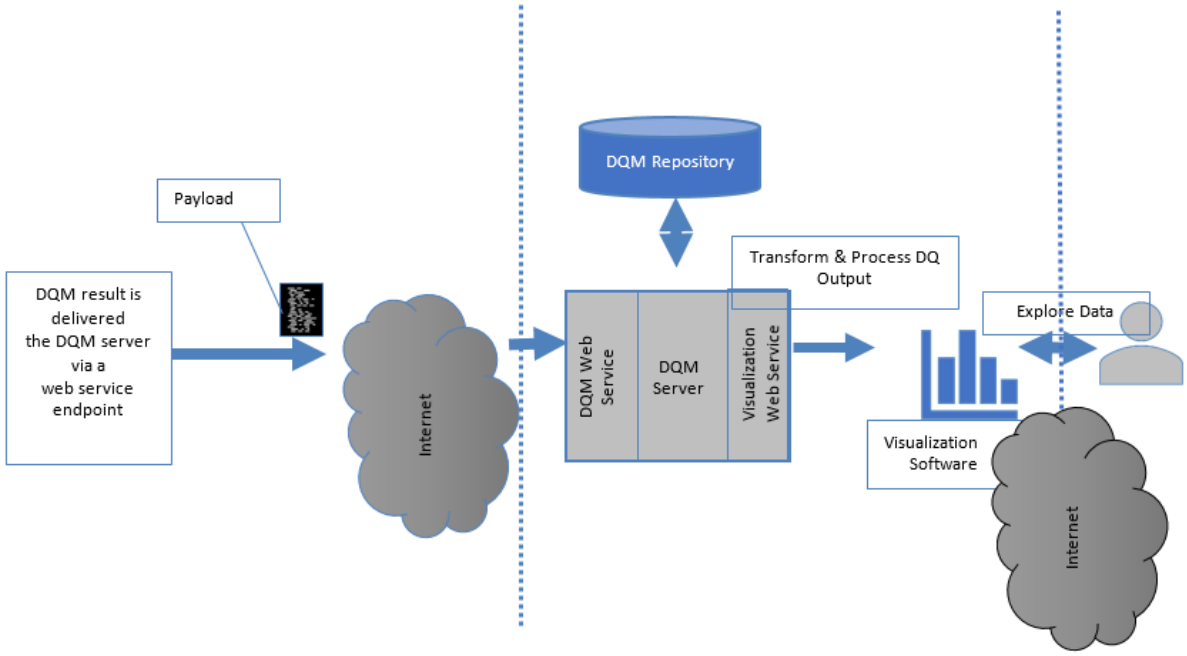
For the purpose of the reference implementation, out of scope items, as those typically captured in an "ideal end-state" document, include:

- Implementing and distributing technology that performs the execution of the DQ metrics and any automation related to receiving and responding to the DQ metrics. That is, we will not define how the source data owner queries for the data from their data sources. However, we will provide the ability to define a DQ metric with enough detail for a data holder to implement.
- Governance related to data access will be discussed and documented during stakeholder meetings, however it may not be addressed in the implementation given potential complexities and costs.

C. ARCHITECTURE: DQM SERVER

Given the assumptions in the last section, the workflow illustrated below has been designed to deliver data quality metric (DQM) results to the Visualization software. For demonstrating the DQM system, we have chosen to use the Visualization software Qlik.

- DQM result is provided as a payload according to specified schema.
- Payload is sent to the DQM web service.
- Payload is parsed and stored in the DQM data repository.
- Visualization software communicates with the Visualization web service to obtain raw or processed data for visualization, further processing and analysis.



The DQM server is a web application hosting the DQM and Visualization web services.

- DQM Web Service – REST endpoint for uploading payload
  - It will support FHIR's MeasureReport resource and accepts the MeasureReport JSON object. Please note that for purpose of this project, the DQM server is not a full FHIR server.
- Visualization Web Service – REST endpoint supporting for supplying raw or processed data for Qlik.
- DQM Server – is responsible for storing the payload into a data repository and for retrieving it for the Visualization web service.

D. ARCHITECTURE: DQM WEB SERVICE

The DQM web service will be a REST API supporting the standard POST action, which is also supported by the FHIR Measure and MeasureReport resources. FHIR has defined a large set of resources using the REST protocol. For the purpose of this project, only the POST and PUT operations on the MeasureReport and/or Measure resource are of interest. This design will enable us to demonstrate how our system can be FHIR-compatible.

- POST: https://[base]/MeasureReport
- PUT: https://[base]/MeasureReport/[id]

The POST operation will return 201 and 200 for create and update success respectively. Errors will return 400, 404, 422 error codes for Bad Request, Resource Not Found and Unprocessable Entity respectively.

E. ARCHITECTURE: PAYLOAD JSON SCHEMA

To capture the DQ response data (or invoke the FHIR resource action), we need a JSON payload. We will use the MeasureReport object as a reference, extending it as needed to express what we need in this project.

The JSON schema for the MeasureReport object is fairly extensive. The exact schema can be found in [16,17]but the fields of interest may be:

- MeasureReport.measure – reference to the measure evaluated to produce this report
- MeasureReport.group.stratifier.stratum.value – the value (count?) of a single stratum within a stratifier; for example, when stratifying by gender, there will be one stratum per gender value
- MeasureReport.group.stratifier.stratum.population

F. ARCHITECTURE: DATA REPOSITORY

The realization of the Logical Data Model described above will be a DQM repository (i.e., data store) that is capable of storing the DQM payload. This model is a key component in designing the Payload JSON Schema. This physical data model will be instantiated in a secure SQL Server database.

G. ARCHITECTURE: VISUALIZATION WEB SERVICE

The Visualization web service is a REST endpoint for supplying any visualization software (e.g., Qlik) with either raw data counts for the DQM repository or computed data.

The API for this web service is evolving, as described in the Project Workflow above, and will be implemented as needed to service the visualization software.

Data characterization (i.e. database fingerprinting) may be done at the visualization API or in the visualization software, if it has programming capabilities, depending on the structure and type of DQ metric. This feature is key to providing an open-source platform by which anyone would be able to use a visualization / analytic tool of choice to connect to the underlying DQ data model.

X. VISUALIZATION SOFTWARE

Qlik Sense will be the visualization software used for the reference implementation, where some DQM processing may be done in the visualization software. Qlik Sense was selected since we already use the software at HPHCI and the tool can connect to standard APIs to import data.

XI. DQM Metric Definition UI/Database

In addition to the ability to gather DQM data as described previously, a database of DQM metrics will also be kept. For the purpose of this scope of work, this is purely for cataloging DQM metrics and relevant metadata.

A web portal for adding and viewing new metrics will be created. A DQM payload can have its measure associated to an ID generated here. This website will be integrated with the data model; it will dynamically change based on metadata/DQ metrics management changes (e.g. a new data quality metric is added to describe the distribution of a specific ethnicity value and the change is immediately available to end users).

XII. PERMISSIONS

Access to the DQM Metrics or the DQM catalog will be controlled by role-based permissions:

- System Administrator = only Users with the System Administrator claim can access the Metric. System Administrators can review submitted, but unpublished metrics.
- Authenticated Users = only Users who have been authenticated can access the specified published Metric
- Public = any User can access the Metric

XIII. DATA DICTIONARY

Details on the entities contained within the model can be found in Table 2 in reference to the Implemented Data Model.

XIV. JIRA – PROJECT REQUIREMENTS AND SPECIFICATIONS

Project objectives and software development are being tracked through a project plan and documented in our JIRA project tracking software.

XV. LIST OF STAKEHOLDERS

A list of stakeholders was submitted in December of 2018 and revised according to feedback from FDA. This was the basis for invitations to the stakeholder sessions held in September 2019.

A. STAKEHOLDER REVIEW

The project team demonstrated functionality during four stakeholder sessions in September 2019. Feedback from various stakeholders has been implemented into the system as part of the iterative development and testing cycles; recordings of all four sessions can be found on the DQM website.

XVI. APPENDIX

1. Definitions

- **DQM Request** – request from the Analysis Center to the source data owner to execute a DQM query and deliver a DQM response; the request may be captured in a variety of formats
- **DQM Result** – results or counts produced from a DQM query
- **Payload** – DQM result in a specified format that can be transported electronically
- **DQM Server** - web server that hosts DQM and Visualization web services
- **DQM Web Service** – web-based software that consumes the Payload and stores it in the DQM repository
- **Visualization Web Service** – web-based software that provides processed or raw data from the DQM repository to Visualization software
- **DQM Repository** – a realization of the Logical Data Model (i.e., a relational database management system (RDBMS), NoSQL database, etc.)
- **Visualization Software** – software that can communicate or otherwise process the information from the Visualization web service; enables visualization, processing and analysis of the DQM data (e.g., Qlik Sense - https://www.qlik.com/us/products/qlik-sense)
- **Logical Data Model** – a data model that can store the definitions of the metrics, metadata about data sources, organizations, as well as the result payload
- **Harmonization** – process of unifying equivalent terms
- **JIRA** (https://www.atlassian.com/software)- issue tracking product developed by Atlassian which allows bug tracking and agile project management.
- **FHIR** - Fast Healthcare Interoperability Resources, pronounced "fire", is a draft standard describing data formats and elements (known as "resources") and an application programming interface (API) for exchanging electronic health records. The standard was created by the Health Level Seven International (HL7) health-care standards organization. FHIR was designed to be consistent, simple to use and understand, and have defined ways to extend for specific purposes. The standard uses coded data types and terminologies (e.g. SNOMDED, ICD-10, etc.) in addition to human readable text. FHIR Profiles are used to customize FHIR to your needs with descriptions of how an existing FHIR data model (i.e. Resource) was modified and, because FHIR is an open standard, Profiles are published in a repository for others to use. In addition to data structure, FHIR also uses standard transport mechanisms commonly used in healthcare and other industries, such as an Application Programming Interface (API) and JSON. There are several publicly available FHIR servers and sandboxes for testing new development efforts. [18]
- **API** (https://en.wikipedia.org/wiki/Application_programming_interface)– Application Programming Interface, set of defined communication protocols, data structures, and tools for software applications; specifies how software components interact [3]
- **CRUD (**https://www.bmc.com/blogs/rest-vs-crud-whats-the-difference/) - Create, Read, Update, and Delete, the standardized use of HTTP Action Verbs. CRUD principles are mapped to REST commands
- **REST** (https://www.bmc.com/blogs/rest-vs-crud-whats-the-difference/)- Representational State Transfer, an architectural style designed for APIs, It uses HTTP protocols like GET, PUT, POST to link resources to actions within a client-server relationship [4]
- **Qlik** – "Qlik Sense® (https://www.qlik.com/us/-/media/files/resource-library/global-us/direct/datasheets/ds-qlik-sense-datasheet-en.pdf) is a next-generation platform for modern, self-service oriented analytics, driving discovery and data literacy for all types of users across an organization"
- **JSON** - JavaScript Object Notation, open-standard file format commonly used to support application portability and interoperability

- **GUI and UI** (https://en.wikipedia.org/wiki/User_interface)– graphical user interface, a tactile and visual interface that humans use to interact with computers
- **CNDS** – Cross Network Directory Service [2,19]

XVII. References

References

1. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC).* 2016;4(1):1244.
2. Cross-Network Directory Service. https://www.sentinelinitiative.org/sentinel/methods/cross-network-directory-service. Accessed 9/28/2018, 2018.
3. Services OoHPASfPaEUSDoHaH. *2017 Annual Report of HHS Projects to Build Data Capacity for Patient-Centered Outcomes Research.* 12/22/2017 2017.
4. Jacqueline Amoozegar BB, Stephen Brown, Alexa Ortiz, Jeanette Renaud,Joshua Richardson,Suzanne West. *Building Data Capacity for Patient-Centered Outcomes Research in HHS: A Formative Evaluation of 2012-2016 Projects.* Division of Healthcare Quality and Outcomes Office of Health Policy/ASPE/HHS; 12/22/2017 2017.
5. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. *Health affairs (Project Hope).* 2014;33(7):1178-1186.
6. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association : JAMIA.* 2014;21(4):578-582.
7. Vogel J, Brown JS, Land T, Platt R, Klompas M. MDPHnet: secure, distributed sharing of electronic health record data for public health surveillance, evaluation, and planning. *Am J Public Health.* 2014;104(12):2265-2270.
8. Electronic medical record Support for Public health. 2018; https://www.esphealth.org/. Accessed 11/05/2018, 2018.
9. *National Institutes of Health (NIH) Strategic Plan for Data Science.* 2016-2020.
10. Callahan TJ, Bauck AE, Bertoch D, et al. A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. *EGEMS (Washington, DC).* 2017;5(1):8.

11. Huser V, Kahn MG, Brown JS, Gouripeddi R. Methods for examining data quality in healthcare integrated data repositories. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing.* 2018;23:628-633.

12. Huser V, DeFalco FJ, Schuemie M, et al. Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets. *EGEMS (Washington, DC).* 2016;4(1):1239.

13. Liaw ST, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. *International journal of medical informatics.* 2013;82(1):10-24.

14. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association : JAMIA.* 2002;9(6):600-611.

15. Anderka M, Mai CT, Romitti PA, et al. Development and implementation of the first national data quality standards for population-based birth defects surveillance programs in the United States. *BMC public health.* 2015;15:925-925.

16. Resource Measure - Content. 2018; http://build.fhir.org/measure.html. Accessed 11/02/2018, 2018.

17. HL7. FHIR - Resource Index. 2018; http://build.fhir.org/resourcelist.html. Accessed 12/04/2018, 2018.

18. HL7. FHIR Overview. 2018; https://www.hl7.org/fhir/overview.html. Accessed 12/06/2018, 2018.

19. Center SO. *Cross Network Directory Service Project: Design and Technical Documentation.* 1/31/2018 2018.

20. Curtis LH, Brown J, Platt R. Four Health Data Networks Illustrate The Potential For A Shared National Multipurpose Big-Data Network. *Health affairs (Project Hope).* 2014;33(7):1178-1186.

21. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association.* 2014;21(4):578-582.

22. Vogel J, Brown JS, Land T, Platt R, Klompas M. MDPHnet: Secure, Distributed Sharing of Electronic Health Record Data for Public Health Surveillance, Evaluation, and Planning. 2014;104(12):2265-2270.

23. Electronic medical record Support for Public Health 2018. 2018.

24. National Intitutes of Health (NIH) Strategic Plan for Data Science. 2016-2020.

B. TECHNICAL DOCUMENTATION

The following document provides technical information appropriate for software developers and other technical users to facilitate their use of the DQM system and the Qlik visualizations; it can be found in the DQM GitHub repository: https://github.com/PopMedNet-Team/DataQualityMetrics.

Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data Project

Technical Documentation

Prepared by: Sentinel Coordinating Center

December 31, 2019

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I. This project was funded by the FDA through HHS Mini-Sentinel contract number HHSF223200910006I. This work was supported by the Office of the Secretary PCORTF under Interagency Agreement #750016PE060001.

DATA QUALITY METRICS

Technical Documentation

I. BACKGROUND

The goal of the Data Quality Metrics project and system was to provide a harmonized approach to data characterization across multiple data sources to enable researchers to better understand candidate data sources before querying and analyzing them. This work included the creation of a system that operationalizes existing data quality (DQ) parameters and methodologies in a way that is compatible across multiple Common Data Models (CDMs) to increase research planning efficiency and improve the interpretability of analytic results.

We created and implemented a data quality data model to contain a set of metadata standards and metrics describing: 1) Data quality and characteristics; 2) Data sources and institutional characteristics; and 3) Fitness-for-use. These standards were the basis for a flexible data quality collation system that is able to incorporate data metrics from any data source. The system was designed to enable flexible exploration of DQ characteristics for multiple data sources at the same time.

Together, the information contained in the data model provides a standardized data source "fingerprint" that can be expanded to provide additional granularity. Additionally, the DQM system was enabled to maintain and query the data model and is available as open source web-based technology such that the system provides approaches to access the data model and can use any business intelligence tool of choice to interact with the data and explore and describe the quality, completeness, and stability of data sources.

This Technical Documentation report is intended for technical stakeholders who have expertise in electronic health data resources and/or software development processes.

A. SYSTEM OVERVIEW

We proposed a pragmatic approach to developing consistent data quality metrics through development of an extensible data model based on a collection of data quality standards and metrics included in the Harmonized Data Quality framework put forth by Kahn et al[1]. An extensible data quality data model must be flexible and independent of the source data model. The Kahn framework describes and defines data quality standards and metrics in a general and harmonized fashion and this system applies it to a variety of data sources and research needs. Operationalizing that framework and developing a tool for analyses allows researchers to evaluate data quality at any life stage of a data source in a consistent manner, and to effectively compare data sources based on the same metrics. A standard data quality metric data model will assist researchers in determining fitness-for-use of various data sources and research purposes.

We have demonstrated our "data fingerprinting" system using synthetic data sets that reflect those used by existing networks, such as PCORnet and Sentinel, with consideration as to how our system can be used by an open network where anyone can review, contribute to, and utilize the DQ data model and explore database fingerprints approved for public consumption— a priority interest for the NIH community and others[20-24].

Although several groups and researchers have done thorough evaluations of DQ metrics for

specific data sources (e.g., birth defect surveillance systems, primary care data, medical registries), to our knowledge there is not currently a data model in place for generic quality measures that can be tailored to specific data sources [10-15]. While study-specific data characterization work provides a framework to evaluate data, it lacks a focus on extensibility and generalizability. Our model will enable users to add any data quality metric of value from their work, thus expanding the initial DQ metrics included in this reference implementation.

We have articulated 78 **use cases**, and the implemented version of the data model captures 25 items of interest (metadata) describing the source system and its measures, as well as 15 items of metadata describing each metric. This information informed the development of the data quality data model and design of the DQM system. Based on the use cases and review of current data quality standards, we identified the following structures to contextualize the quality of data:

- Time component (e.g., number of encounters by clinical setting per year)
- Person-based construct (e.g., number of prescriptions ordered per person per year)
- External context (e.g., rates of asthma by age compared to expected population rates)

B. KEY FUNCTIONAL COMPONENTS

- **Metrics**

Metrics are the descriptions of quantitative measurements that can be executed on data sources to characterize a specific aspect of the source data in a data model agnostic way. The DQM tool captures metadata about each Metric in a standardized way, regardless of the context or use cases. Metric authors describe the metric in enough detail for a data holder to interpret and generate the results of the Metric from their source data. These results, or measures, enable apples-to-apples comparisons across data sources irrespective of the CDM or data structure.

- **Measures**

A Measure is the numeric representation of a metric that has been executed against a data source. Measures include the data characteristics defined in the metric, as well as metadata about the data source, metric details, and information regarding when the measurement was calculated. The Measures can be explored in the visualization tools found in Explore DQM.

- **Explore DQM**

The DQM visualization tools overlay the metadata, metrics, and measures. Users can explore and evaluate data sources for specific characteristics, trends, and quality. DQM does not determine whether a data source passes or fails the executing of a metric, but rather provides a view of data characteristics that enable a user to determine if the data are fit for their purpose.

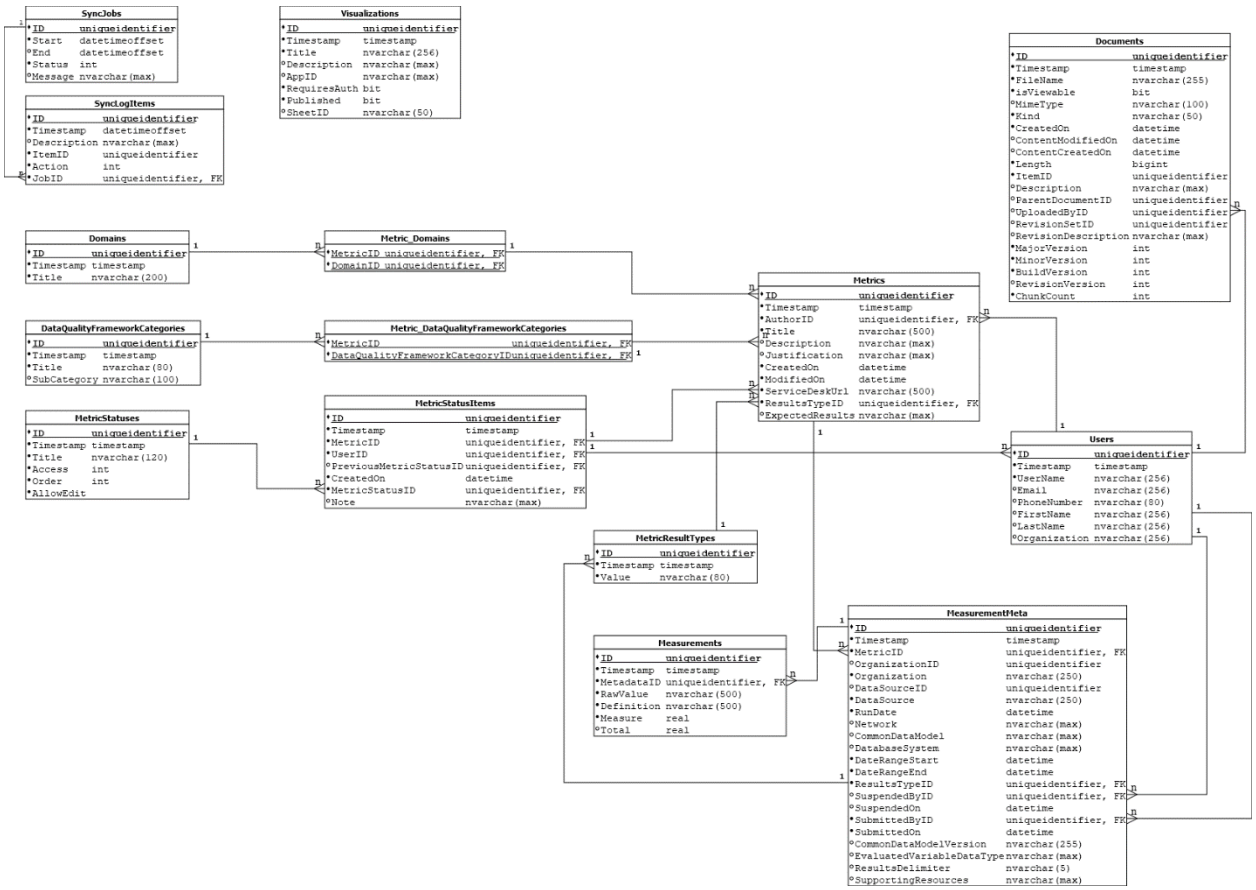
Additional details on implementation of the visualization tools can be found in documentation developed by Analytics8 – a data and analytics consulting firm that engaged in the work – in the

appendix.

II. DATA MODEL AND ENTITIES

Following iterative design discussions, a final data quality data model was implemented as the underlying structure of the system.

A. DATA QUALITY DATA MODEL



- Solid dots indicate non-nullable fields.
- Underlined fields indicate Primary Keys.
- Relations are indicated by the connecting lines and their connectors.
- All non-collection tables have a primary key that is named ID.
- A non-nullable timestamp field is included on all tables that require optimistic concurrency for Entity Framework.

The root entities are Metric and Measure Metadata; all other entities support defining attribute of those entities. Entity relationships are depicted below in figures 1 and 2 and further detailed in Section D.

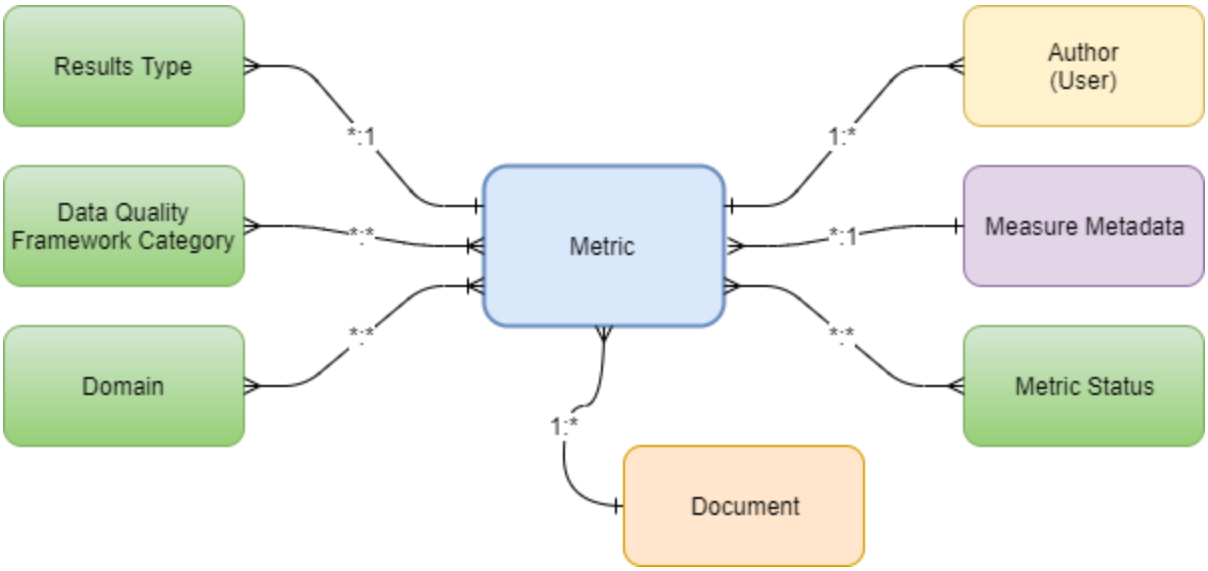


Figure 1. Metric Entity

- A User can author zero or more Metrics. A metric must have an author.
- A Metric has a collection of statuses, each status item is immutable.
    - A new status item is created for each status change, and the most current item is the current status of the Metric.
    - A metric status item contains the date the status changed, the status, the User that changed the status, a reference to the previous status item, and an optional note regarding the status change.
- A Metric has a single Results Type association. A Results Type can be associated to more than one Metric.
- A Metric has one or more Data Quality Framework Category associations. A Data Quality Framework Category can be associated to more than one Metric.
- A Metric has one or more Domain associations. A domain can be associated to more than one Metric.
- A Metric has zero or more Measure Metadata associations. Measure Metadata must be associated to a Metric.
- A Metric has zero or more Document associations. A document must be associated to an entity.

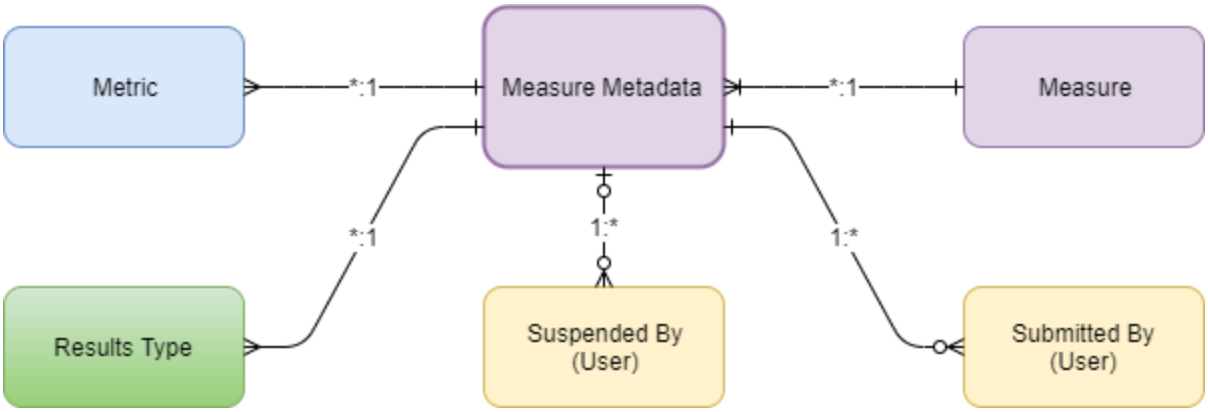


Figure 2. Measure Metadata Entity

- Measure metadata represents the metadata about a collection of measures.
- Measure metadata must be associated to a single Metric. A Metric can be associated to more than one Measure metadata.
- Measure metadata must be associated to a single Results Type. A Results Type can be associated to more than one Measure metadata.
- Measure metadata is associated to more than one Measure. A Measure must be associated to a single Measure metadata.
- Measure metadata must be associated to a single User representing who submitted the measure data. A User can be associated to more than one Measure metadata as the submitter.
- Measure metadata may have an association to a single User representing who suspended the measure data. A User can be associated to more than one Measure metadata as the suspender.

B. ENTITY DETAILS

Entity	Details
User	• Represents a "person" • Requires a UserName, and optionally: a first and last name, email address, phone number, and associated organization name
Results type	• Indicates the Results Type of a Metric, and/or Measure • Comprised of a display title • Can be associated with many Metrics
Data Quality Framework Category	• Indicates the category a Metric could be classified as • The category classifications are based on definitions defined by the Khan framework • Comprised of a Title and optional Sub-category • Can be associated with many Metrics
Domain	• Indicates the domain a Metric belongs to • A domain is comprised of a title • Can be associated with many Metrics
Metric status	• The definition of a status a Metric can be assigned • Comprised of a title, an access level, a logical order value, and if editing of the Metric is allowed while in the status

Entity	Details

	• The access levels define which users have access to a Metric, and are comprised of the following values:
	  • None = no access level specified
	  • Author = only the author of the Metric has access
	  • System Administrator = only Users with the System Administrator claim can access the Metric
	  • Authenticated Users = only Users who have been authenticated can access the Metric
	  • Public = any User can access the Metric

Metric status item	• The instance of a status for a Metric • Comprised of the Metric, User, Metric Status, the previous Metric Status, Creation date, and a note • A Metric will have one or more status items; the one with the most recent creation date is the current status.

Metric	• The definition of a Metric is comprised of:
	  • Title, Description, Justification, Expected Results, Created On and Modified On dates, Service Desk URL
	  • An Author - the User creating the Metric
	  • ResultsType
	  • One or more Data Quality Framework Categories
	  • One or more Domains
	  • One or more Metric Status Items
	  • Zero or more Measures (Measure Metadata)

Measure Metadata	• Represents the metadata about a collection of Measures
	• The definition of a Measure Metadata is comprised of:
	  • A Metric; Measures are the quantitative result of a query based on a Metric definition
	  • Organization name, and optionally it's ID
	  • DataSource name, and optionally it's ID
	  • A run date for when the data was collected
	  • The network the Data Source belongs to
	  • The Common Data Model the data may belong to
	  • The Database System the data was stored in
	  • Date Range Start is the earliest date of the data set
	  • Date Range End is the latest date of the data set
	  • Results Type ID, the ID of the Results Type associated to the Measures. Must match the Results Type defined on the associated Metric.
	  • Suspended By, the User who suspended the Measures excluding it from available queries
	  • Submitted By, the User who uploaded the Measures to DQM
	  • Common Data Model Version, the version number of the CDM the Measure data may belong to
	  • Results Delimiter, the delimiter used if the values of the Measures are compounded and the result of more than one value.
	  • Supporting Resources, a URL to a location providing resources

Entity	Details
	(application, scripts, documentation, etc.) used to obtain the measures. • A collection of one or more Measures

Measure	• Represents the instance of a Measure
	• Comprised of:
	  • Raw Value represents the unmodified value of the stratifier that the measure is for
	  • Definition represents a display value for the Raw Value: i.e. Raw Value = 'M' and the Definition = 'Male'
	  • Measure is the numerical quantity of the result. Depending on the Results Type defined by the Metric it could be a count, percentage, range, or vector.
	  • Total is the optional value representing the total of all the Measure values, it could be greater than the sum of the Measure values included.

III. WEBSITE CONFIGURATION SETTINGS

The website application and web jobs use the standard ASP.Net Core configuration framework to manage and access application configuration settings. The default base configuration file (appsettings.json) contains the default configuration values; the local developer base configuration settings are located in appsettings. Development.json with local configuration values overridden via the Debug environment variables found in the project properties. The local settings are stored in the launchSettings.json file for the specific launch profile, each developer should create their own launch profile.

Settings for the Azure deployed application are specified as environment variables within the Azure App Service configuration.

The configuration files are specified using json in a hierarchical structure. The hierarchical path of a specific setting can be stated by delimiting the path using a colon.

Example default configuration found in appsettings.json.

```
{
  "Logging": {
    "LogLevel": {
      "Default": "Warning"
    }
  },
  "AllowedHosts": "*",
```

```
  "PMNApiUrl": "",
  "PMNPortal": "",
  "PMNoAuthKey": "",
  "PMNoAuthHash": "",
  "QlikServer": "",
  "QlikServerQPSPort": "4243",
  "QlikQPSPrefix": "",
  "QlikUserDirectory": "",
  "QlikUserID": "",
  "QlikQPSCertThumbprint": "",
  "QlikCertLocation": "",
  "Files": {
    "Type": "ASPE.DQM.Files.LocalStorageFileService, ASPE.DQM.Files",
    "UploadDirectory": "",
    "StorageConnectionString": "",
    "FileStorageShare": "",
    "DataLakeStorageAccountName": "",
    "DataLakeStorageClientID": "",
    "DataLakeStorageClientSecret": "",
    "DataLakeStorageTenantID": "",
    "DataLakeStorageDirectory": ""
  },
  "ConnectionStrings": {
    "IdentityContextConnection": ""
  },
  "Sync": {
    "ServiceKey": ""
  }
}
```

A. CONFIGURATION SETTINGS

Setting Key	Description
Logging:LogLevel:Default	Specifies the logging level by default for system logging.
Logging:LogLevel:{namespace[classname]}_Requirements,_design,_and	Specifies the logging level for a specific namespace within the source. Examples include: "System", and "Microsoft"
Serilog:*	The configuration settings for Serilog. Refer to https://github.com/serilog/serilog-settings-configuration for documentation.
AllowedHosts	See: https://docs.microsoft.com/en-us/aspnet/core/fundamentals/servers/kestrel?view=aspnetcore-2.2#host-filtering-1
PMNApiUrl	The url to the API for the CNDS PMN instance.
PMNPortal	The url to the SSO login endpoint of the CNDS PMN portal instance.
PMNoAuthKey	The oauth authentication key for interacting with the PMN single sign-on.
PMNoAuthHash	The security hash for interacting with the PMN single sign-on.
QlikServer	The root domain of the Qlik server. Does not include the http scheme or trailing slash.
QlikServerQPSPort	The port of the QPS for the Qlik installation.
QlikQPSPrefix	The url prefix of the Qlik proxy
QlikUserDirectory	The user directory for Qlik authentication.
QlikUserID	The ID of the Qlik user DQM will use for impersonation.
QlikQPSCertThumbprint	The thumbprint of the certificate used to validate the connection to the Qlik server
QlikCertLocation	The certificate installation location, default is

Setting Key	Description
	"LocalMachine"
ConnectionStrings:IdentityContextConnection	The SQL Server connection string to the DQM database.
Sync:ServiceKey	The authentication key used for the CNDS/DQM user synchronization service.
Files	Configuration settings for file storage. Required settings depend upon the type of file storage.
Files:type	The type of file storage provider to use. Default is local file storage. The provider type is specified as "class name, assembly name".
Files:UploadDirectory	The path to the folder files should be saved. Required for LocalStorageFileService.
Files:StorageConnectionString	The connection string to the Azure storage account. Required for AzureBlobStorageFileService, and AzureFileStorageFileService.
Files:FileStorageShare	The Azure storage share key. Required for AzureBlobStorageFilesService, and AzureFileStorageFileService.
Files:DataLakeStorageAccountName	The Azure Data Lake storage account name. Required for AzureDataLakeFileService.
Files:DataLakeStorageClientID	The Azure Data Lake storage account client ID. Required for AzureDataLakeFileService.
Files:DataLakeStorageClientSecret	The Azure Data Lake storage account client secret. Required for AzureDataLakeFileService.
Files:DataLakeStorageTenantID	The Azure Data Lake storage account tenant ID. Required for AzureDataLakeFileService.
Files:DataLakeStorageDirectory	The Azure Data Lake storage account directory name. Required for AzureDataLakeFileService.

* Settings that have their key in **bold** are required.

IV. DEVELOPER SET UP

A. DQM APPLICATION REQUIREMENTS

- Windows 10
- Microsoft Visual Studio 2017+, all editions supported
- Microsoft SQL Server 2014 or greater
- WebPack Test Runner for Visual Studio by Mads Kristensen, not required but makes running WebPack builds much easier.
- NodeJS
- Typescript SDK
- .NET Core 2.2 SDK

B. DQM APPLICATION INSTRUCTIONS

1. Install Visual Studio, and apply any updates.
    a. Confirm the ASP.NET and web development option has been selected
    b. Confirm that .NET Core 2.2 is selected if available.
2. Install .NET Core SDK if not installed via Visual Studio.
3. Install SQL Server, and apply any updates. Make sure the current Windows User is authorized for the database, and Integrated Security is enabled.
4. Install Typescript SDK found at https://www.typescriptlang.org/#download-links
5. Install NodeJS found at https://nodejs.org/en/download/
6. Install the WebPack Test Runner from the Visual Studio Extensions gallery.
7. If support for Qlik applications is required, install the Qlik certificate into the Local Computer store
    a. Certificate and instructions are found in ~/QlikCert folder of the source
8. After installing the software dependencies and obtaining the source code for the application, the ASPE.DQM.sln can be opened using Visual Studio. Perform a build only of the solution and confirm all projects compiled successfully. Open the Task Runner Explorer panel from the "View => Other Windows" menu, under the webpack.config.js item expand "Run" and double click the "Run-Development" option. This will initiate the WebPack build which will compile the typescript, placing the output into the wwwroot/scripts folder of the web application.
9. If an existing copy of the DQM database is available, restore the database to SQL Server with the name "ASPE_DQM".
10. If starting without a copy of the DQM database, it can be created by running the migrations via the Package Manager Console in Visual Studio.
11. At this point the DQM web application can be launched using IIS Express via Visual Studio.

C. POPMEDNET APPLICATION REQUIREMENTS

For the DQM project, the final CNDS version was used. Any version greater than 6.2 of PopMedNet is compatible.

- Windows 10
- Microsoft Visual Studio 2017+. All editions supported

- Microsoft SQL Server 2014 or greater
- Typescript SDK version 3.2
- ASP.Net MVC 4 if the PopMedNet version is less than 6.12.0.0
- RazorGenerator extension for Visual Studio (https://github.com/RazorGenerator/RazorGenerator). Only required if making changes to .cshtml files
- Less compiler; only required if making changes to .less files
- .NET SDK 4.7.2

D. POPMEDNET APPLICATION INSTRUCTIONS

A. PopMedNet is used by DQM to manage User registration, and user permissions. No development is required for the usage and integration of PMN with DQM. The PMN instance can either be run via IIS Express using Visual Studio, or it can be compiled and deployed to an IIS instance.
B. After installing the software dependencies, and obtaining the source for the application, the PMN websites are ready to be built and optionally deployed.
C. Restore a compatible version of the PMN and CNDS databases to SQL Server, update the connection strings in the ConnectionStrings.config files found in the Lpp.Dns.Api, Lpp.Dns.Portal, and Lpp.CNDS.Api project folders. The ConnectionStrings.config can be created by making a copy of the ConnectionStrings-template.config file, and should not be added to source control.
D. Open the Lpp.Dns.Api solution with Visual Studio and build the entire solution. Using the Package Manager Console in Visual Studio confirm the PMN database is up to date by executing any pending migrations.
E. Open the DistributedNetworkSolution solution with Visual Studio and build the entire solution.
F. Open the Lpp.CNDS solution with Visual Studio and build the entire solution. Using the Package Manager Console in Visual Studio confirm the CNDS database is up to date by executing any pending migrations.
G. The CNDS website is only required if CNDS integration is part of the PMN instance being used. DQM does not have a dependency on CNDS, only PMN.
H. After confirming the solutions compile without errors, the websites can be run using IIS via Visual Studio or by publishing to a local folder and configuring websites in an IIS instance.
I. Depending upon how it is desired to run PMN; confirm that the correct URLs are configured in the DQM appsettings.Development.json file. The PMNApiUrl value should be the root URL of the PMN API website (i.e. http://localhost:24592), and the PMNPortal value should be the URL to the SSO login action for the PMN Portal website (i.e. http://localhost:60344/ssologin).
J. Confirm that the PMNoAuthKey and PMNoAuthHash values in the DQM configuration settings match the values specified in the Lpp.Dns.Portal/web.config for the settings SsoKey and SsoHash. The PMN SSO site does not need to be used, however DQM uses the SSO infrastructure in the PMN Portal site to enable cross-application authentication.

V. DQM VISUALIZATION SET-UP

A. DQ METRICS & DQ MEASURES LOAD SCRIPT DETAILS

Harvard Pilgrim Health Care

DQMetrics & DQMeasures Load Script Details

Written by: Chris Domain

REVISION HISTORY

Date	Version	Description	Author
10/16/2019	1.0	Initial Document Creation	Chris Domain

DOCUMENT OVERVIEW

This document provides details on all the load scripts used in the DQMetrics Final and DQMeasures final applications. For each app I will be describing what each script is being used for and how it affects the final application.

DQ METRICS APPLICATION

API/REST Connections:

The DQMetrics Application pulls data from five separate API's using five rest connectors. Below I've listed the names of the rest connectors as well as the API URL's that they are connected to:

REST_METRICS:https://dataquality.healthdatacollaboration.net/api/qlik-export/metrics

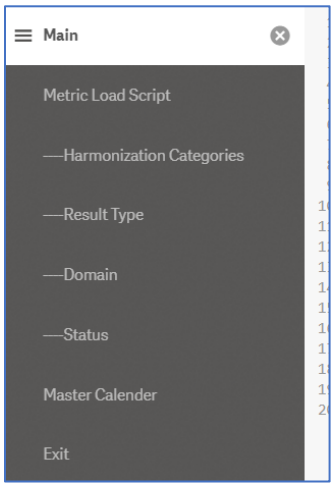
REST_HARMONIZATION_CATEGORIES:https://dataquality.healthdatacollaboration.net/api/qlik-export/data-quality-harmonization-categories

REST_RESULTS_TYPES:https://dataquality.healthdatacollaboration.net/api/qlik-export/results-types

REST_DOMAINS:https://dataquality.healthdatacollaboration.net/api/qlik-export/domains
REST_STATUSES: https://dataquality.healthdatacollaboration.net/api/qlik-export/metric-statuses

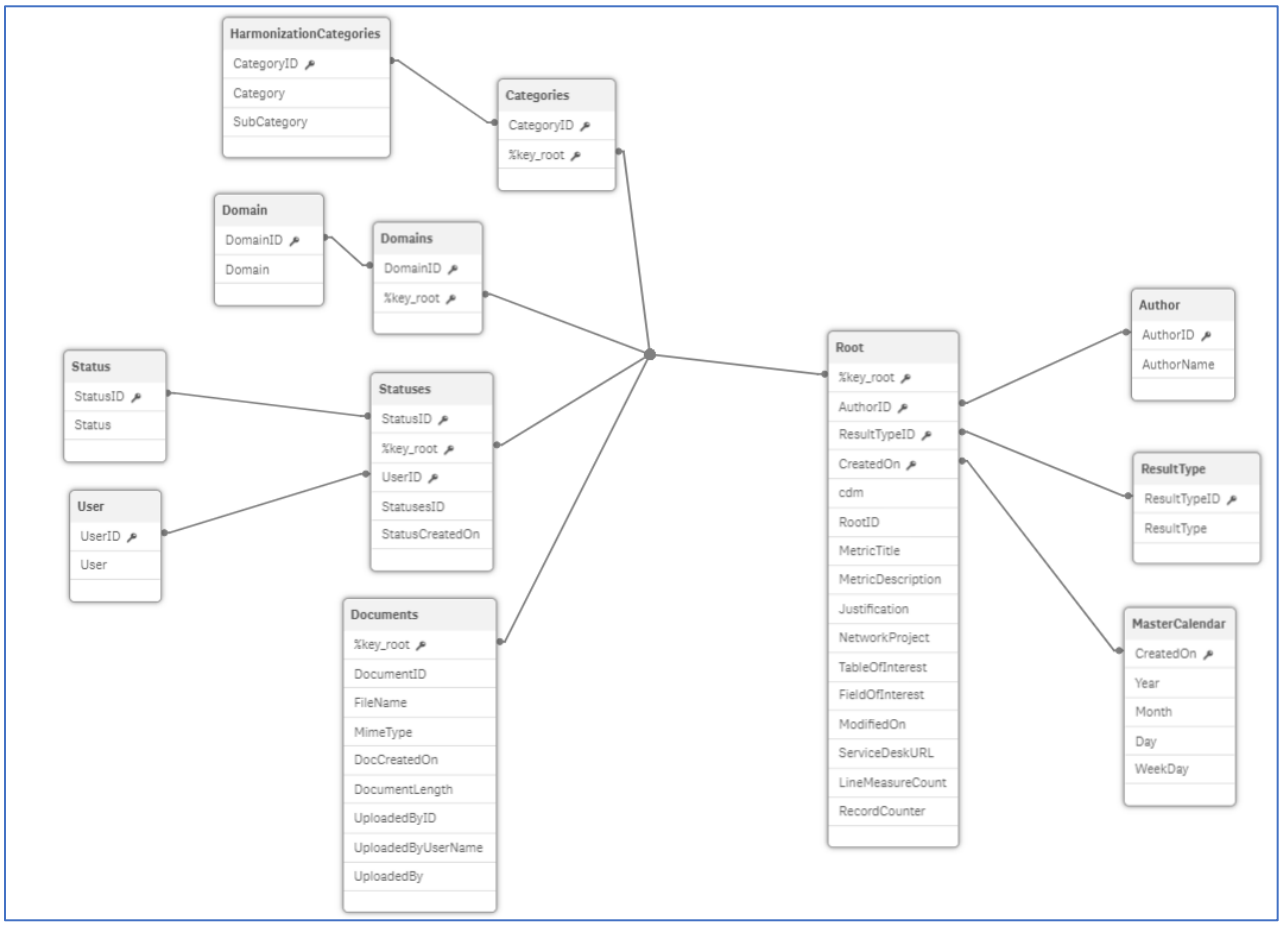
From here on out I will be referring to the connections by their rest connector names.

Load Script Overview:



The DQMetrics script is broken up into eight sections. **Main** initializes the Qlik Sense settings. **Metric Load Script** connects using the REST_METRIC connector, it pulls in seven tables: Root, Author, User, Statuses, Categories, Domains, and Documents. **----Harmonization Categories** connects using the REST_HARMONIZATION_CATEGORIES connector, it pulls in one table: HarmonizationCategories. **----Result Type** uses the REST_RESULT_TYPES connector, it pulls in one table: ResultType. **----Domain** uses the REST_DOMAINS connector, it pulls in one table: Domain. **----Status** uses the REST_STATUSES connector, it pulls in one table: Status. **Master Calendar** creates and additional table MasterCalender used for date visualizations. Finally **Exit** just contains the Exit Script to stop the script.

The final data model looks like this:



Metric Load Script:

In this section all the data from REST_METRICS is pulled into a temporary table named RestConnectorMasterTable, the seven final tables are created using resident loads from the master table. Once the seven final tables are created the RestConnectorMasterTable is dropped. In this section the only editing done is mostly by renaming fields. In the Root, Categories, Domains, Statuses, and Documents tables I have renamed their key values to %key_values, this is how the supporting tables are linked to the Root table. At the bottom of the Root table you will see “1 AS RecordCounter”, the one measure in this application (# Metrics) sums this field to get the count of Metrics. Summing is more efficient than counting in Qlik.

```
Author:
LOAD DISTINCT
    [authorID] As AuthorID,
    [author] AS AuthorName
RESIDENT RestConnectorMasterTable
WHERE NOT IsNull([__KEY_root]);


User:
LOAD DISTINCT
    [userID] AS UserID,
    [user] AS User
RESIDENT RestConnectorMasterTable
WHERE NOT IsNull([__FK_status]);
```

The only other additions to this section is the creation of the Author and User tables. I created these tables as a distinct load so all the Authors/Users can be visualized in one place. Knowing this information helps with visualization creation as well as for filtering, the tables look like this:

Author	
Rows	6
Fields	2
Keys	1
Tags	$key $ascii $text

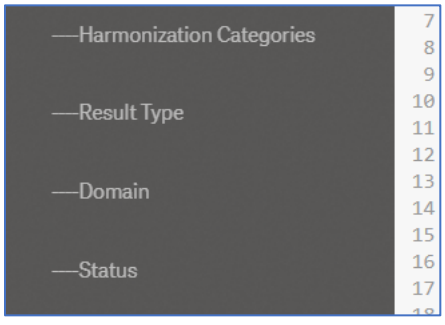
Preview of data

AuthorID	AuthorName
8bf95fbb-75c6-4f5d-894b-a575009cfcb9	Dee, Daniel
d79b3d3b-60ec-4109-a653-a7860105a2a1	nannapaneni, lakshmi
4bbae1ec-6ba7-42f4-a125-a77600d78242	Malenfant, Jessica
41c7442a-5822-4ff0-a316-aacb011b756c	User. Test
3631e528-3662-4ff1-a5d1-a57200c2d4bb	Nolan, Bridget
a563861d-22f3-4cea-beaf-a6c600a4f117	Barrett, Kimberly

This section alone is responsible for this portion of the data model:

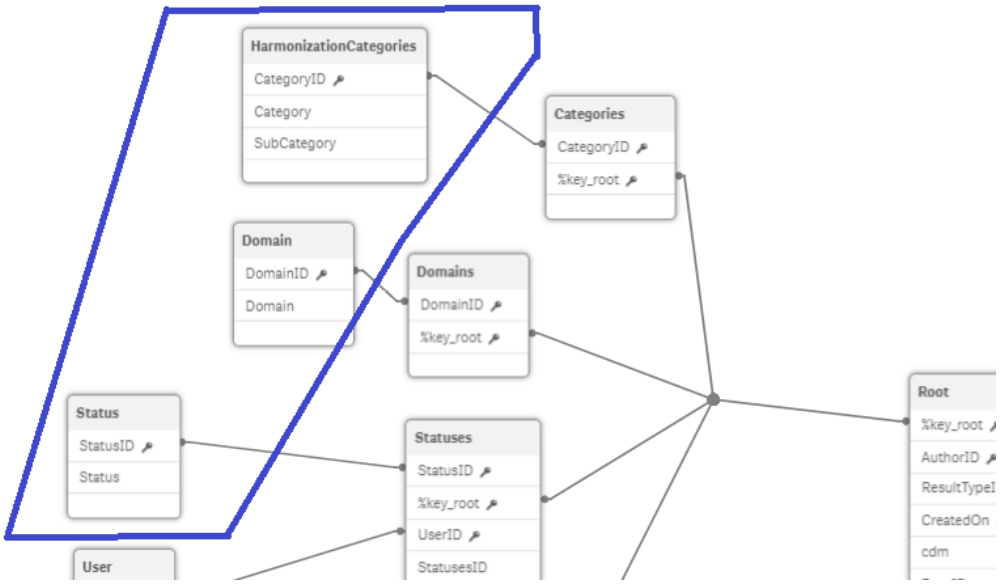
The rest of the sections are supporting tables used for filtering and visualization enhancement.

The '----' Indented Sections:



The four sections above pull data from the other four rest connectors: REST_HARMONIZATION_CATEGORIES, REST_RESULTS_TYPES, REST_DOMAINS, REST_STATUSES. Each one comes with two fields, [id] and [title] (harmonization categories has an extra for subcategory). The [title]s are renamed to what their value represents and the [id]s are renamed to match the ID in the

metric tables: DomainID, StatusID, etc.



As seen above in the data model, these tables sit on the peripheral of the schema connected by their renamed ID's. They exist to show all the possible values a given category, domain, or status can be regardless if those values exist in the metrics data. These tables allow us to have a complete view of possible values and is important when it comes to filtering and creating visuals later on.

Master Calendar:

The master calendar is the last section in the metric script. It's connected to the Root table by the CreatedOn date field. The way it works is by finding the minimum and maximum date in the CreatedOn dataset. It then fills in a table with every single day between the min and max date to create a full date dataset.

```
MinMaxTemp:
LOAD
    Date(Min(FieldValue('CreatedOn', RecNo()))) as MinDate,
    Date(Max(FieldValue('CreatedOn', RecNo()))) as MaxDate
AUTOGENERATE FieldValueCount('CreatedOn');

LET vToday = NUM(PEEK('MaxDate',0,'MinMaxTemp'));

//**** Create the OrderDate field ****
MasterCalendar_Temp:
LOAD
    Date(MinDate + IterNo() - 1) as CreatedOn //Create the CreatedOn field
RESIDENT MinMaxTemp
WHILE MinDate + IterNo() - 1 <= MaxDate;

DROP TABLE MinMaxTemp;
```

The rest of the script is just for formatting. The reason we use a master calendar is so we have all the date values in a given time regardless of whether or not data was gathered on that day. In the application we use the master calendar CreatedOn value in our visualizations instead of the one from the root table. It allows line charts or any other chart of date vs value to be distributed properly across a time span instead of clumping the dates together.

```
MasterCalendar:
LOAD
    CreatedOn,
    Year(CreatedOn) as Year,
    Month(CreatedOn) as Month,
    Day(CreatedOn) as Day,
    WeekDay(CreatedOn) as WeekDay;
 //**** Generate a temp table of dates ****
LOAD
    Date(MinDate + IterNo() - 1) as CreatedOn
WHILE MinDate + IterNo() - 1 <= MaxDate;
//**** Retrieve Min and Max dates from OrderDate field ****
LOAD
    Min(FieldValue('CreatedOn', RecNo())) as MinDate,
    Max(FieldValue('CreatedOn', RecNo())) as MaxDate
    AUTOGENERATE FieldValueCount('CreatedOn');

DROP TABLE MasterCalendar_Temp;
```

In Analysis:

All the fields used to create visualizations in this application have been made as master dimensions and measures. When editing a sheet in Qlik Sense you can go to the left side of the screen and click on master items below the fields tab. There are seven master dimensions and one master measure. They are the only fields I used to create every visual in this app.

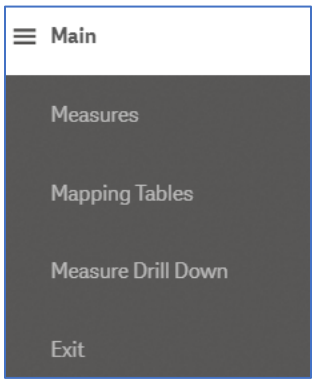
DQ MEASURES APPLICATION

API/REST Connection:

The DQMeasures Application pulls data from one API. The rest connector name and API URL are listed below:

REST_MEASURES_BY_METRICS:https://dataquality.healthdatacollaboration.net/api/qlik-export/measures-by-metric

Load Script Overview:



The DQMeasures script is broken up into five sections. **Main** initializes the Qlik Sense settings. **Measures** connects using the REST_MEASURES_BY_METRIC connector, it pulls in three tables: Root, Metadata, and MeasuresMaster. **Mapping Tables** contains all the mapping tables used in Measure Drill Down. **Measure Drill Down** are resident tables based off the MeasuresMaster table, and are used to create visualizations focused on a particular type of measure. Finally **Exit** just contains the Exit Script to stop the script.

Measures Section:

In this section the three main tables are pulled into Qlik, most of the editing here is just renaming fields but there are a few important things to note.

1.) The Root table which contains all the measure types is connected to the Metadata table by a field I have named %key_root, and the Metadata table is connected to all the rest of the tables including the MeasuresMaster table by a field I have named %key_measures.
2.) To be able to divide up the MeasuresMaster table I needed to map the %key_root value to the MeasuresMaster table, and I have renamed that field to RootValue. This is what the Root_map

table it for, it is not seen in the final data model.

```
Root_map:
MAPPING LOAD
    [__KEY_measurements] AS %key_measures,
    [__FK_measurements] AS %key_root
RESIDENT RestConnectorMasterTable
WHERE NOT IsNull([__FK_measurements]);
```

3.) Likewise there is a Suspended_Map table which maps the [suspendedOn] date value from the Metadata table to the MeasuresMaster table. This allowed me to write a condition at the bottom of the Metadata and MeasuresMaster tables that states only records which have not been suspend are pulled into Qlik. If someone suspends a record in the website then when the app is refreshed that record will no longer appear in the app. Allows junk data to be cleaned by the end user.

```
Suspended_Map:
MAPPING LOAD
    %key_measures,
    [suspendedOn]
RESIDENT MetaData;

MeasuresMaster:
LOAD
    [rawValue] As RawValue,
    [definition] As Definition,
    [measure] AS Measure,
    [total] As Total,
    APPLYMAP('Root_map',[__FK_measures],-1) AS RootValue,
//      APPLYMAP('Suspended_Map',[__FK_measures],1) AS Suspend,
    [__FK_measures] AS %key_measures,
    1 AS RowCounter
RESIDENT RestConnectorMasterTable
WHERE NOT IsNull([__FK_measures]) AND IsNull(APPLYMAP('Suspended_Map',[__FK_measures],1));
```

4.) The final notable thing in this section is that I added counter values in the Metadata table and MeasureMaster table. In the analysis these values are summed to create the # Rows and # Submissions master measures.

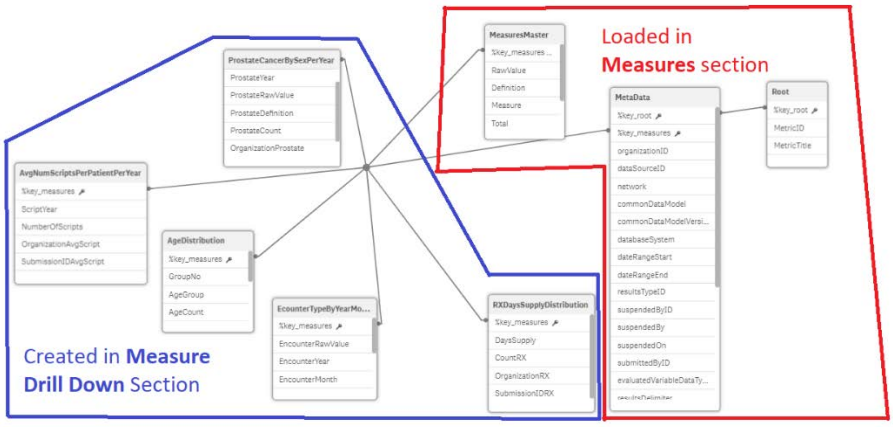
Mapping Tables:

This section contains five additional mapping tables that are used by the tables in the Measure Drill Down section. This application has a lot of filters that are based in the Metadata table. When creating visualizations we can easily pull the fields directly from the Metadata table to filter our data but it presents a slight problem. For example if we wanted to filter by Organization for RX Counts Per Year data and we pulled the Organization field for the filter directly from the Metadata table it will show every single Organization for any type of submissions as options. Even if those Organizations have no submissions for RX Counts Per Year. When the organization is mapped to RX Counts Per year then used as a filter then only Organizations that have submissions for RX Count will appear in the filter.

Measure Drill Down:

This section pulls data from the MeasuresMasters by filtering on the RootValue mentioned early. Five tables are created here for five focus areas: Age Distribution, Average Number of Scripts Per Patient Per Year, Prostate Cancer By Sex Per Year, Encounter Type By Year Month, and RX Days Supply Distribution. These tables utilize the maps from the previous sections for filter values. Whenever there was multiple data in a single column I split it using the subfield() function.

The resulting data model looks like this:



Harvard Pilgrim Health Care

Registering a sheet in the DQM site

Written by: Chris Domain

REVISION HISTORY

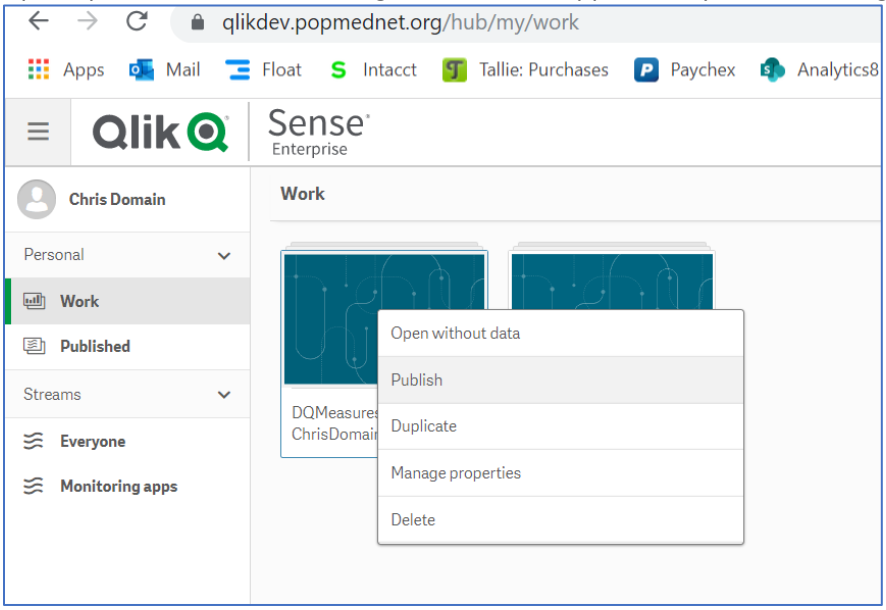
Date	Version	Description	Author
10/16/2019	1.0	Initial Document Creation	Chris Domain

DOCUMENT OVERVIEW

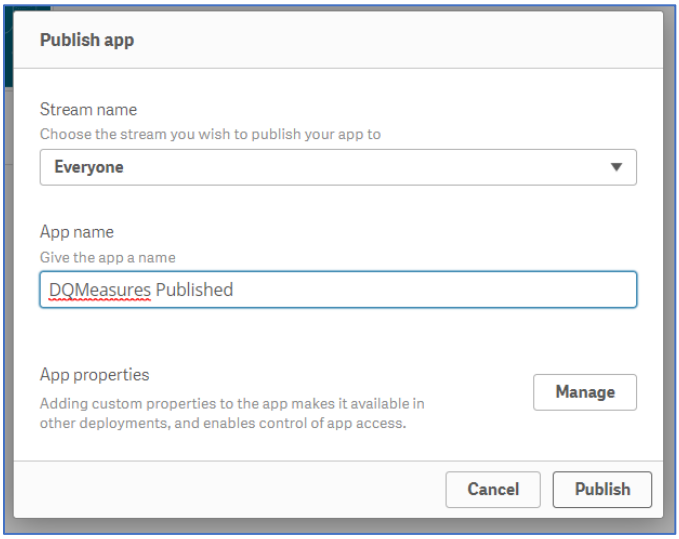
This document provides details on how to register a Qlik sheet into the DQM measures website for viewing.

STEPS

1.) Open up the Qlikdev hub and right click on the application you want to register, select publish.

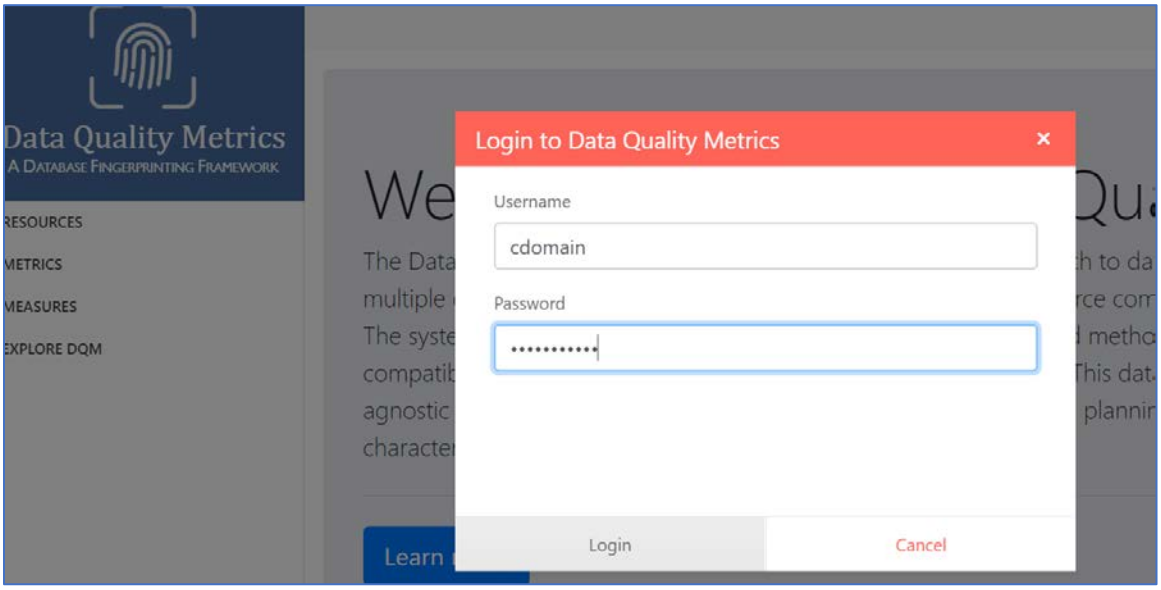


2.) Select the 'Everyone' stream and give the application a name, hit Publish.

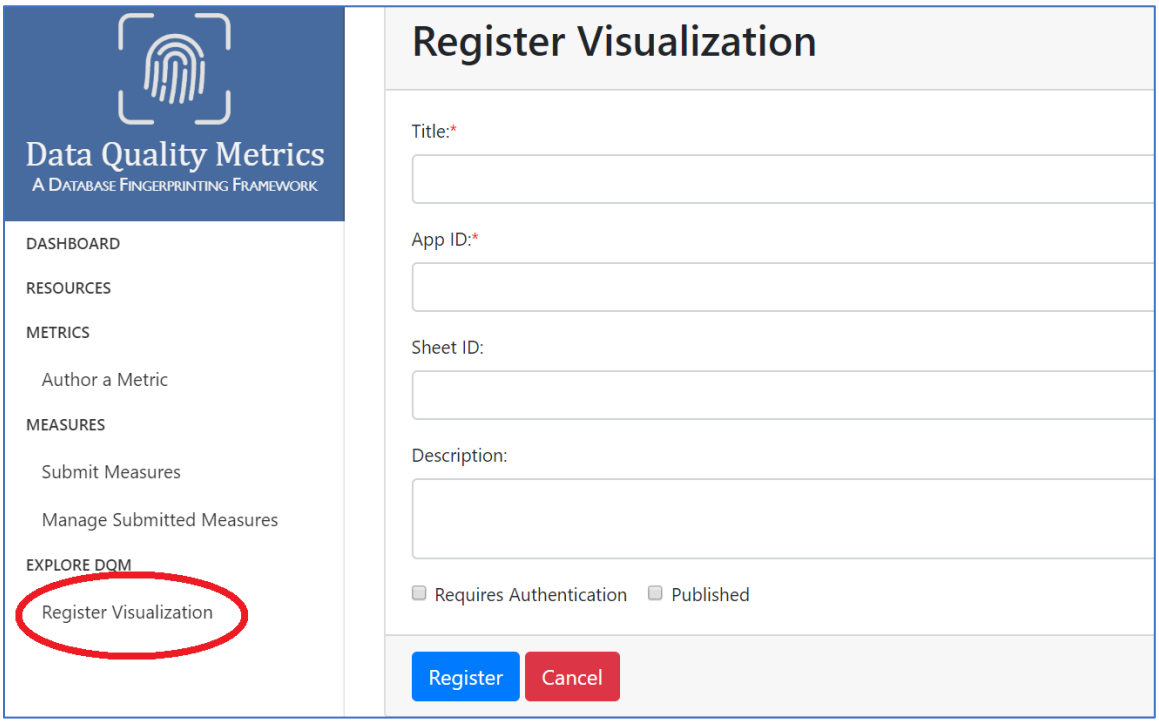


3.) Open the application you just published in the Everyone stream, keep this page open then open a new tab.

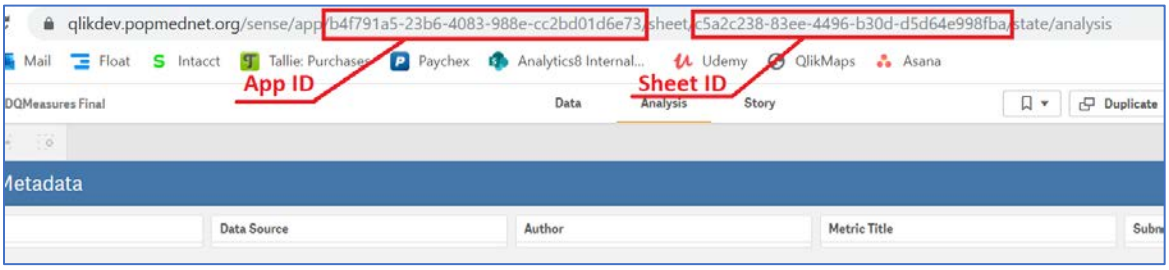
4.) Go to the DQM site: https://dataquality.healthdatacollaboration.net/ , click Login and enter your credentials, click Login.



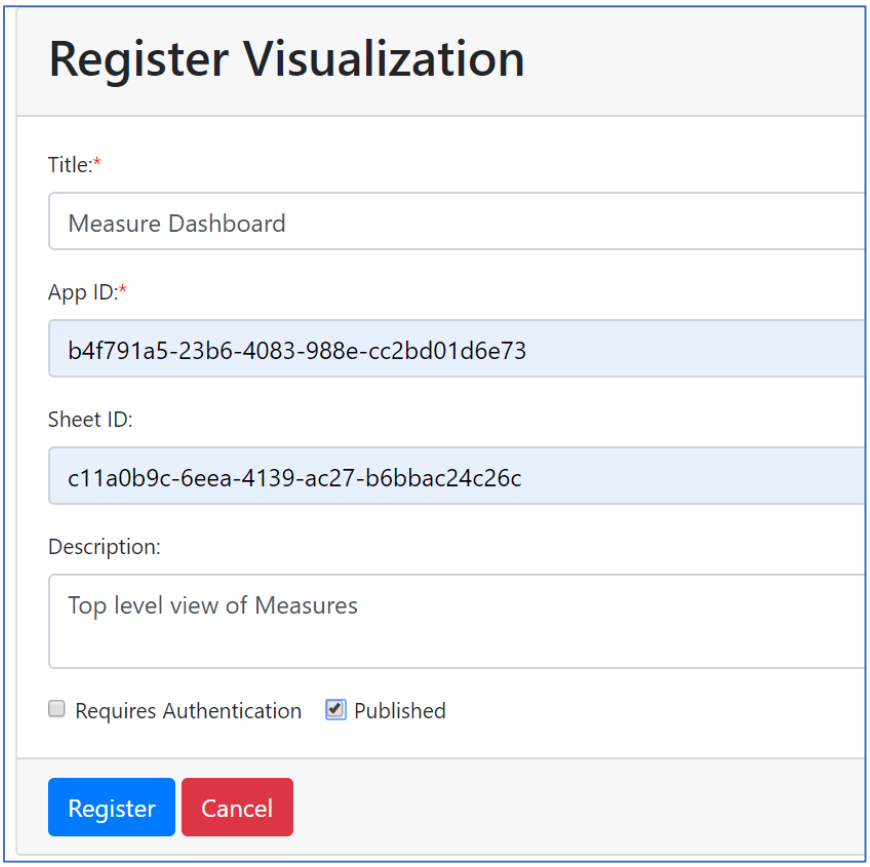
5.) On the bottom of the left menu select 'Register Visualization', you will be brought to the screen below.



6.) The title and description can be anything you want. To get App ID and Sheet ID navigate to the sheet you want to register, the ID's will be located in the URL.



7.) Once you enter all the information, check the 'Published' box and click register.



8.) To see the report simply select 'Explore DQM' from the menu and select the sheet you just registered!

References

1. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC).* 2016;4(1):1244.
2. Curtis LH, Brown J, Platt R. Four Health Data Networks Illustrate The Potential For A Shared National Multipurpose Big-Data Network. *Health affairs (Project Hope).* 2014;33(7):1178-1186.
3. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association.* 2014;21(4):578-582.
4. Vogel J, Brown JS, Land T, Platt R, Klompas M. MDPHnet: Secure, Distributed Sharing of Electronic Health Record Data for Public Health Surveillance, Evaluation, and Planning. 2014;104(12):2265-2270.
5. Electronic medical record Support for Public Health 2018. 2018.
6. Natinal Intitutes of Health (NIH) Strategic Plan for Data Science. 2016-2020.
7. Callahan TJ, Bauck AE, Bertoch D, et al. A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. *EGEMS (Washington, DC).* 2017;5(1):8.
8. Huser V, Kahn MG, Brown JS, Gouripeddi R. Methods for examining data quality in healthcare integrated data repositories. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing.* 2018;23:628-633.
9. Huser V, DeFalco FJ, Schuemie M, et al. Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets. *EGEMS (Washington, DC).* 2016;4(1):1239.
10. Liaw ST, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. *International journal of medical informatics.* 2013;82(1):10-24.
11. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association : JAMIA.* 2002;9(6):600-611.
12. Anderka M, Mai CT, Romitti PA, et al. Development and implementation of the first national data quality standards for population-based birth defects surveillance programs in the United States. *BMC public health.* 2015;15:925-925.

C. REQUIREMENTS, DESIGN, AND TESTING – JIRA TRACKING

The following table documents how all requirements, design specifications, bug reports, and updates to the DQM system were captured in JIRA throughout the project.

JIRA Issue(s)	Category	Description	Phase 1, 2, and/or 3
DQM-2	Discovery and Design	Map data quality metrics to the data model	Phase 1
DQM-3	Discovery and Design	Create JSON payload for Metrics	Phase 1
DQM-4, DQM-5	Discovery and Design	Data quality measures to data model mapping and payload	Phase 1
DQM-6	Discovery and Design	Create JSON payload for the measures/results	Phase 1
DQM-7	Discovery and Design	Wrap up of Data Quality Metrics Project Design phase	Phase 1
DQM-8	Discovery and Design	Using Qlik as the Visualization Tool	Phase 1
DQM-12	Discovery and Design	Epic for overall website requirements and desired pages	Phase 1
DQM-16, DQM-17, DQM-18	Discovery and Design	Stories for the implementation of the DQM website and dependencies (website, services, applications, etc.)	Phase 1
DQM-19, DQM-20, DQM-21, DQM-22, DQM-23, DQM-24, DQM-25, DQM-26, DQM-39, DQM-44	Discovery and Design	User Registration and link with the CNDS system. Includes user registration and how the CNDS system will be leveraged for this project	Phase 1
DQM-27	Discovery and Design	DQM website: sitemap	Phase 1
DQM-28, DQM-29, DQM-30, DQM-37	Discovery and Design	Design for Metrics aspect of site. Includes the ability to author, copy, edit, and submit metrics.	Phase 1
DQM-31, DQM-38	Discovery and Design	Design for the Measures aspect of the site. Includes the ability to populate a measure template and submit measures.	Phase 1
DQM-32, DQM-33, DQM-34	Discovery and Design	Design for the Explore DQM aspect of the site. Includes the need to register visualizations (apps)	Phase 1
DQM-35, DQM-36	Discovery and Design	DQM website: Project overview and objective, and landing page	Phase 1
DQM-40, DQM-41	Development and Testing	DQM website: set-up website in Azure cloud system	Phase 2
DQM-46	Discovery and Design	DQM website Ideal End State report	Phase 1 and 2
DQM-47, DQM-51	Development and Testing	Updates to site design, layout, and pages	Phase 2

JIRA Issue(s)	Category	Description	Phase 1, 2, and/or 3
DQM-49, DQM-114	Development and Testing	Create API calls	Phase 2
DQM-52, DQM-60, DQM-61, DQM-62, DQM-63, DQM-64, DQM-80, DQM-82, DQM-83, DQM-87, DQM-93	Development and Testing	Implementation of the following **Metric** related issues: • List (view) all metrics • Author functionality • Metric details • Submit metrics • Edit Metrics • Copy Metrics • Comment on Metrics • Add documents • When authoring, include a look-up tool for similar metrics • Ability to change metric status	Phase 2
DQM-53, DQM-73, DQM-76, DQM-89	Development and Testing	Implementation of the following **Measure** related issues: • Measure Template • Upload measure • Manage submitted measures • Auto-generate Measure template for each Metric	Phase 2
DQM-54, DQM-56, DQM-59, DQM-88	Development and Testing	DQM website: User profile and dashboard	Phase 2
DQM-55	Development and Testing	Changes to User Registration	Phase 2
DQM-57, DQM-127	Development and Testing	Implementation of the **Explore DQM** and visualization related issues: • List visualizations • Create visualization host • Visualization registration - Add ability to add and delete visualizations	Phase 2
DQM-58, DQM-68, DQM-70, DQM-84, DQM-85, DQM-86	Development and Testing	Implementation of the following: • User registration • Logging-in • Authentication • Authorization- • Link and integration of the CNDS system • Sync with the CNDS system	Phase 2

JIRA Issue(s)	Category	Description	Phase 1, 2, and/or 3
		• Configure default error pages and "Not Authorized" pages	
DQM-65, DQM-74	Development and Testing	Text for the Overview, Project Objective and site Landing Page	Phase 2
DQM-66	Development and Testing	Create and populate the Resources Page	Phase 2
DQM-67, DQM-126	Development and Testing	**Qlik** • Installation and set-up of Qlik server • Embed Qlik to website	Phase 2
DQM-69	Development and Testing	DQM data: Model for metric and dependencies	Phase 2
DQM-71	Development and Testing	DQM data: Document storage options	Phase 2
DQM-72	Documentation	DQM data: document metadata	Phase 2 and 3
DQM-78	Development and Testing	Create implementations for handling documents	Phase 2
DQM-79	Development and Testing	Updates to site based on feedback	Phase 2
DQM-90	Development and Testing	Submit Metric button covers the Metric List grid after 12 metrics have been added	Phase 2
DQM-91	Development and Testing	Explore DQM page: pop-up asking to translate page appears	Phase 2
DQM-92	Development and Testing	Replace the green box on top left corner	Phase 2
DQM-95	Development and Testing	Additional properties for the Measure Metadata	Phase 2
DQM-96	Development and Testing	Some external users cannot access the site	Phase 2
DQM-98	Development and Testing	Excel import for Measures was not read correctly	Phase 2
DQM-100	Documentation	Document the Technical Process of Importing Measures	Phase 2 and 3
DQM-101, DQM-103, DQM-112	Development and Testing	Unable to Upload Measures document	Phase 2
DQM-94, DQM-97, DQM-102, DQM-107, DQM-108, DQM-109, DQM-110, DQM-122, DQM-130, DQM-131, DQM-137, DQM-141,	Development and Testing	Updates to DQM site text and links	Phase 2

JIRA Issue(s)	Category	Description	Phase 1, 2, and/or 3
DQM-145, DQM-146. DQM-151			
DQM-105, DQM-111	Documentation	Technical documentation and database diagrams for the DQM data model	Phase 2 and 3
DQM-113	Development and Testing	DQM Website: Include Speed and Visual improvements when data is loading	Phase 2
DQM-115	Development and Testing	Populate DQM website with Metrics and publish Metrics	Phase 2
DQM-117	Development and Testing	Add links for recordings in the Community Engagement Section for the Stakeholder Meetings	Phase 2
DQM-118, DQM-120, DQM-121	Development and Testing	IE browser – Clicking on various links does not work	Phase 2
DQM-123	Development and Testing	Add new field for Metric to describe expected results based on stakeholder feedback	Phase 2
DQM-124	Development and Testing	Add new field to Measure Template Metadata tab for data resources based on stakeholder feedback	Phase 2
DQM-125	Development and Testing	Description of visualizations not appearing on Explore DQM page	Phase 2
DQM-128, DQM-129, DQM-132, DQM-133	Development and Testing	Add ability to bookmark visualizations and metrics	Phase 2
DQM-134	Development and Testing	Make API changes to fix html issue in Qlik	Phase 2
DQM-135, DQM-144	Development and Testing	Update .NET core for website	Phase 2
DQM-136	Development and Testing	Unable to delete draft metrics	Phase 2
DQM-138	Development and Testing	Update webpack to support production configuration on website	Phase 2
DQM-139	Development and Testing	Unable to Login when using a small screen, e.g. mobile phone	Phase 2
DQM-142, DQM-150	Documentation	Upload DQM source code to GitHub	Phase 3
DQM-143	Development and Testing	Date displayed in Uploaded Measure details on the user dashboard are incorrect	Phase 2
DQM-106	Consideration for future work	Functionality for system admins to manage the metadata elements (fields and value sets) for Metrics	Phase 3

JIRA Issue(s)	Category	Description	Phase 1, 2, and/or 3
DQM-147	Consideration for future work	Leverage the CNDS and PMN infrastructure for adoption	Phase 3
DQM-148	Consideration for future work	Enhance governance based on feedback from stakeholders	Phase 3
DQM-149	Consideration for future work	Based on site governance, design or write specifications for a distributed DQM System. This is based on stakeholder feedback.	Phase 3
DQM-152	Documentation	Document DQM set-up in Azure Environment	Phase 3

D. STAKEHOLDER SUMMARY

The stakeholder summary documents the stakeholder engagement activities, including documentation of stakeholder comments and disposition of comments. This feedback informed additional testing and updates to the system to ensure end user goals were addressed. Recordings of stakeholder sessions can be found within the "Community Engagement" section of the DQM Resources page: https://dataquality.healthdatacollaboration.net/resources

I. BACKGROUND

Four stakeholder sessions were held and recorded in September 2019 to demonstrate a beta-version of the software. The sessions addressed the following topics: 1) demonstration and discussion related to authoring data quality metrics - these two sessions were targeted to stakeholders that are interested in the creation and discussion of metrics that can be utilized for multiple data sources and research questions; and 2) demonstration and discussion regarding exploring database fingerprints - these two sessions were targeted to stakeholders that are interested in evaluating fitness for use of various data sources or for various research questions.

This report represents the deliverable Objective 5 as described in the Statement of Work and has been prepared according to the updated deliverable schedule reviewed with FDA in April of 2019.

The tables contained in this document detail the summarized feedback by subject area, as well as the follow up and response from the project team.

The appendices of this document include:

1. Meeting summaries from stakeholder sessions
2. Previous Discovery and Design deliverable

II. METRICS

The following table is a summary of the feedback received from the two stakeholder meetings that focused on a demonstration and discussion related to authoring data quality metrics. The goal of these sessions was for the project team to: review and discuss the metadata fields captured for each Metric, discuss engaging community members to author Metrics, and discuss sharing of resources as they relate to Metrics.

Feedback	Disposition
Be more prescriptive in the Metric to enable implementation and interpretation of the Measure. • Individuals may run Metrics differently and obtain different counts. • Include a mechanism to describe specific use of a Metric and its strengths and weaknesses in a specific setting.	Feedback addressed: Added additional field for users to describe the expectations of a metric (e.g. For encounters over time, we would expect to see an increase) *(internal JIRA # DQM-123)* and developed a community discussion board to share implementation details and resources.
We have description and justification fields, but need to be more clear about where users should document the “why” of the Metric so that others understand the significance of implementing it.	Feedback addressed: Added additional field for users to describe the expectations of a metric (e.g. For encounters over time, we would expect to see an increase) *(internal JIRA # DQM-123)*

Feedback	Disposition
It would be helpful to decide whether details of a Metric are included in the webpage *or* supporting documentation.	Feedback addressed: Removed fields related to implementation details (e.g. Network or Project, Tables of Interest) to avoid confusion and drive users to the community board for implementation discussions *(internal JIRA # DQM-122)*
We may have to rely on the community to tell us how they execute Metrics to further inform the details.	Feedback addressed: Added an optional field in the template for submitting measures to allow data holders to link to any shareable code related to the query (*internal JIRA # DQM-124)*
"Metric vs. measure vs. check" concepts may need to be presented clearly upfront to set expectations.	Metric and Measure concepts are defined on the DQM site home page, as well as the respective sub-pages.

III. MEASURES

The following table is a summary of the feedback received from the two stakeholder meetings that focused on a demonstration and discussion regarding exploring database fingerprints. The goal of these sessions was for the project team to: discuss the process for submitting data to the site, discuss community engagement, and discuss sharing of resources as they relate to running queries and sharing Measures.

Feedback	Disposition
Field experience reveals edge cases that were not previously considered in research work and queries, so it is hugely important to include the voice of data owners.	Feedback addressed: We have developed a community discussion board to share implementation details and resources.
We may need to consider versioning based on field experience.	This item is beyond the scope of the pilot project and will be documented as potential for Future Directions.
It would be very useful for contributors to make their code available. • E.g. include information on how the Metric was executed, such as SQL queries, R package, SAS program, etc.	Feedback addressed: Added an optional field in the template for submitting measures to allow data holders to link to any shareable code related to the query (*internal JIRA # DQM-124).* The community board can also facilitate discussions and information sharing related to running a metric.
We need transparency about what is done from real raw data	Feedback addressed: We have developed a community discussion board to share implementation details on transformation of data.

Feedback	Disposition
• E.g., having a convention for something that is missing may be necessary even if some data models enforce values • E.g., some understanding of a health care system is useful to capture the data, the ETL decisions made, and the skillset of the requester	Describing the upstream raw data is beyond the scope of the pilot project and will be documented as potential for Future Directions.

IV. GOVERNANCE & ENGAGEMENT

The following table is a summary of the feedback received from all four stakeholder meetings. During all four meetings, the project team had a goal of understanding incentives and barriers to participation, discussing strategies and materials that would engage community members, and determining what contributors would expect for governance and access controls.

Feedback	Disposition
Consider additional questions for stakeholders and community members on governance, oversight, and sustainability.	This item is beyond the scope of the pilot project and will be documented as potential for Future Directions.
Showing comparative metrics, identified or not, will require a lot of discussion on governance. Various sites may respond differently to the idea of sharing this kind of data due to small cell counts, business risk, etc.	This item is beyond the scope of the pilot project and will be documented as potential for Future Directions.
Insofar as anyone has the resources for the governance process, we could make the option available. • Note what went through an approval workflow and was vetted	This item is beyond the scope of the pilot project and will be documented as potential for Future Directions.
Think about visualizations that are helpful to an individual organization, e.g., the organization that submitted compared to all others. • If data is updated or changed, what is the motivation or incentive for end users to keep the information current in all the places it lives?	Feedback addressed: Example visualizations have been developed to compare one organization to the average of all others, and further this discussion. Requiring contributors to maintain current documentation of data is beyond the scope of the pilot project and will be documented as potential for Future Directions.
Further discussions are needed on the incentive for sites to engage with the system; many sites characterize their data locally or in a central network, and the DQM system is an additional arena to do so.	This item is beyond the scope of the pilot project and will be documented as potential for Future Directions.

E. USER DOCUMENTATION

The User Documentation below provides detailed user documentation information related to the use of the web-based DQM system. It can be found in the DQM GitHub repository: https://github.com/PopMedNet-Team/DataQualityMetrics

Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data Project

User Documentation

Prepared by: Sentinel Coordinating Center

December 31, 2019

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I. This project was funded by the FDA through HHS Mini-Sentinel contract number HHSF223200910006I. This work was supported by the Office of the Secretary PCORTF under Interagency Agreement #750016PE060001.

DATA QUALITY METRICS

User Documentation

I. BACKGROUND

The goal of Data Quality Metrics project and system was to provide a harmonized approach to data characterization across multiple data sources to enable researchers to better understand candidate data sources before querying and analyzing them. This work included the creation of a system that operationalizes existing data quality (DQ) parameters and methodologies in a way that is compatible across multiple Common Data Models (CDMs) to increase research planning efficiency and improve the interpretability of analytic results.

We created and implemented a data quality data model to contain a set of metadata standards and metrics describing: 1) Data quality and characteristics; 2) Data sources and institutional characteristics; and 3) Fitness-for-use. These standards were the basis for a flexible data quality collation system that is able to incorporate data metrics from any data source. The system was designed to enable flexible exploration of DQ characteristics for multiple data sources at the same time.

Together, the information contained in the data model provides a standardized data source "fingerprint" that can be expanded to provide additional granularity. Additionally, the DQM system was enabled to maintain and query the data model and is available as open source web-based technology such that the system provides approaches to access the data model and can use any business intelligence tool of choice to interact with the data and explore and describe the quality, completeness, and stability of data sources.

A. SYSTEM OVERVIEW

We proposed a pragmatic approach to developing consistent data quality metrics through development of an extensible data model based on a collection of data quality standards and metrics included in the Harmonized Data Quality framework put forth by Kahn et al[1]. An extensible data quality data model must be flexible and independent of the source data model. The Kahn framework describes and defines data quality standards and metrics in a general and harmonized fashion and this system applies it to a variety of data sources and research needs. Operationalizing that framework and developing a tool for analyses allows researchers to evaluate data quality at any life stage of a data source in a consistent manner, and to effectively compare data sources based on the same metrics. A standard data quality metric data model will assist researchers in determining fitness-for-use of various data sources and research purposes.

We have demonstrated our "data fingerprinting" system using synthetic data sets that reflect those used by existing networks, such as PCORnet and Sentinel, with consideration as to how our system can be used by an open network where anyone can review, contribute to, and utilize the DQ data model and explore database fingerprints approved for public consumption— a priority interest for the NIH community and others [20-24].

Although several groups and researchers have done thorough evaluations of DQ metrics for specific data sources (e.g., birth defect surveillance systems, primary care data, medical registries), to our knowledge there is not currently a data model in place for generic quality measures that can be tailored to specific data sources [10-15]. While study-specific data characterization work provides a framework to evaluate data, it lacks a focus on extensibility

and generalizability. Our model will enable users to add any data quality metric of value from their work, thus expanding the initial DQ metrics included in this reference implementation.

We articulated 78 use cases to support development of the data quality metric data model and open-source toolkit (the DQM system). In addition to the specific metrics used as use cases, the implemented DQM system captures 25 items of interest (metadata) describing the source data system and its measures, as well as 15 items of metadata describing each metric. This information informed the development of the data quality data model and design of the DQM system. Based on the use cases and review of current data quality standards, we identified the following structures to contextualize the quality of data:

- Time component (e.g., number of encounters by clinical setting per year)
- Person-based construct (e.g., number of prescriptions ordered per person per year)
- External context (e.g., rates of asthma by age compared to expected population rates)

B. KEY FUNCTIONAL COMPONENTS

1. Metrics

Metrics are the descriptions of quantitative measurements that can be executed on data sources to characterize a specific aspect of the source data in a data model agnostic way. The DQM tool captures metadata about each Metric in a standardized way, regardless of the context or use cases. Metric authors describe the metric in enough detail for a data holder to interpret and generate the results of the Metric from their source data. These results, or measures, enable apples-to-apples comparisons across data sources irrespective of the CDM or data structure.

2. Measures

A Measure is the numeric representation of a metric that has been executed against a data source. Measures include the data characteristics defined in the metric, as well as metadata about the data source, metric details, and information regarding when the measurement was calculated. The Measures can be explored in the visualization tools found in Explore DQM.

3. Explore DQM

The DQM visualization tools overlay the metadata, metrics, and measures. Users can explore and evaluate data sources for specific characteristics, trends, and quality. DQM does not determine whether a data source passes or fails the executing of a metric, but rather provides a view of data characteristics that enable a user to determine if the data are fit for their purpose.

II. Functionality

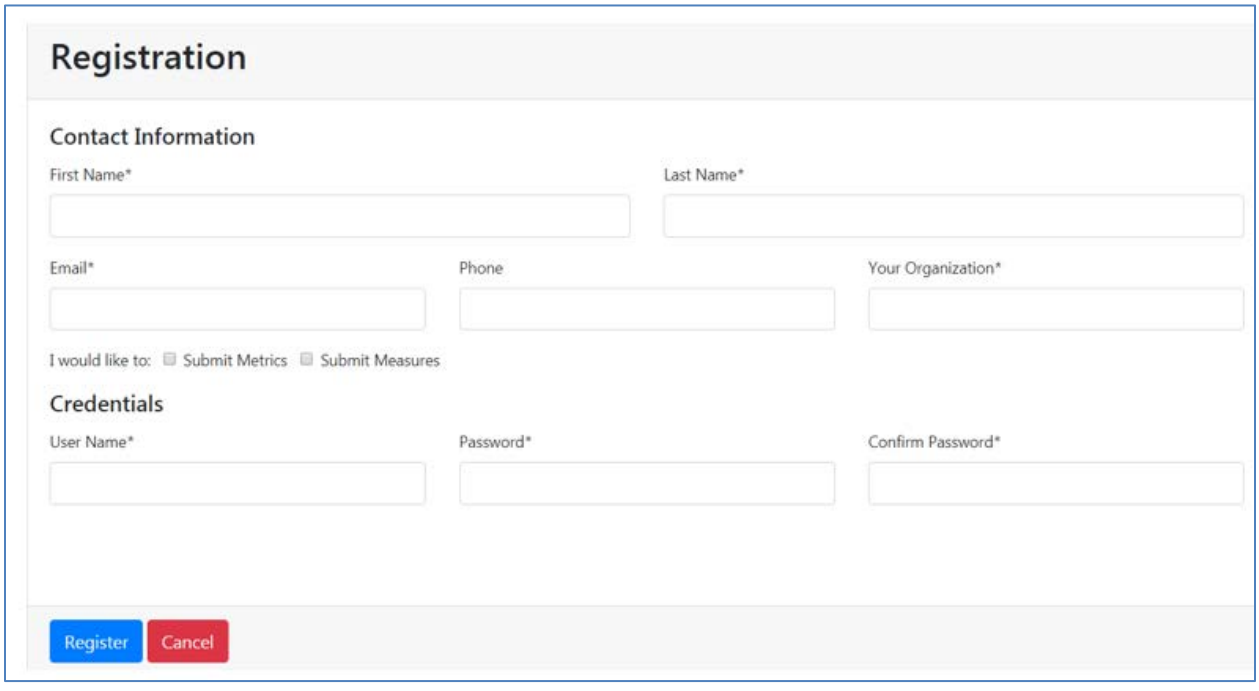
The DQM System was instantiated as a web portal with multiple pages of functionality.

A. Register

Users can navigate to the DQM system landing page and select the "Register" button to create a user profile and request permissions for functionality within the site.

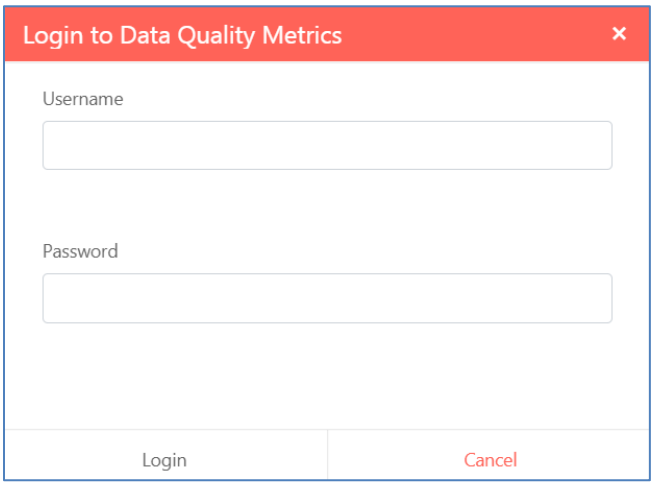
Requested information includes:

- First and last name
- Email address
- Phone
- Your organization
- Requested permissions
  - Submit Metrics (i.e. Author Metrics)
  - Submit Measures
- Credentials
  - User name
  - Password
  - Confirmation of Password



B. LOGIN

Upon registration, any time a user navigates to the site, they are able to login and access additional pages within the site.



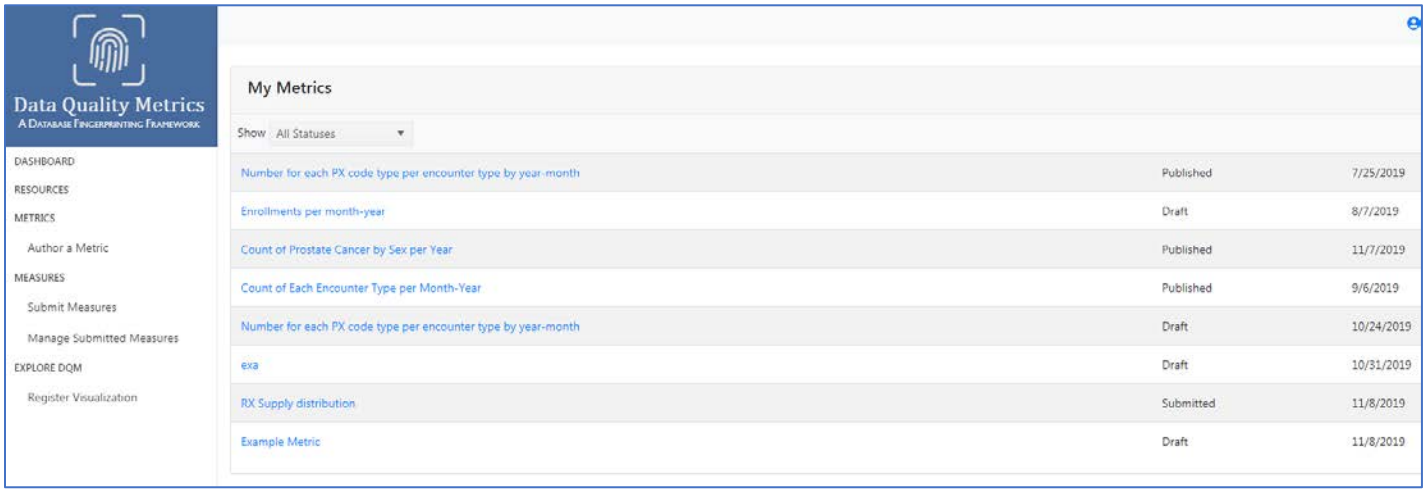
C. DASHBOARD

Once logged-in, users will have access to a personal Dashboard. Navigating to the Dashboard allows a user to interact with metadata specific to their individual use of the DQM system related to the Key Functional Components.

1. My Metrics

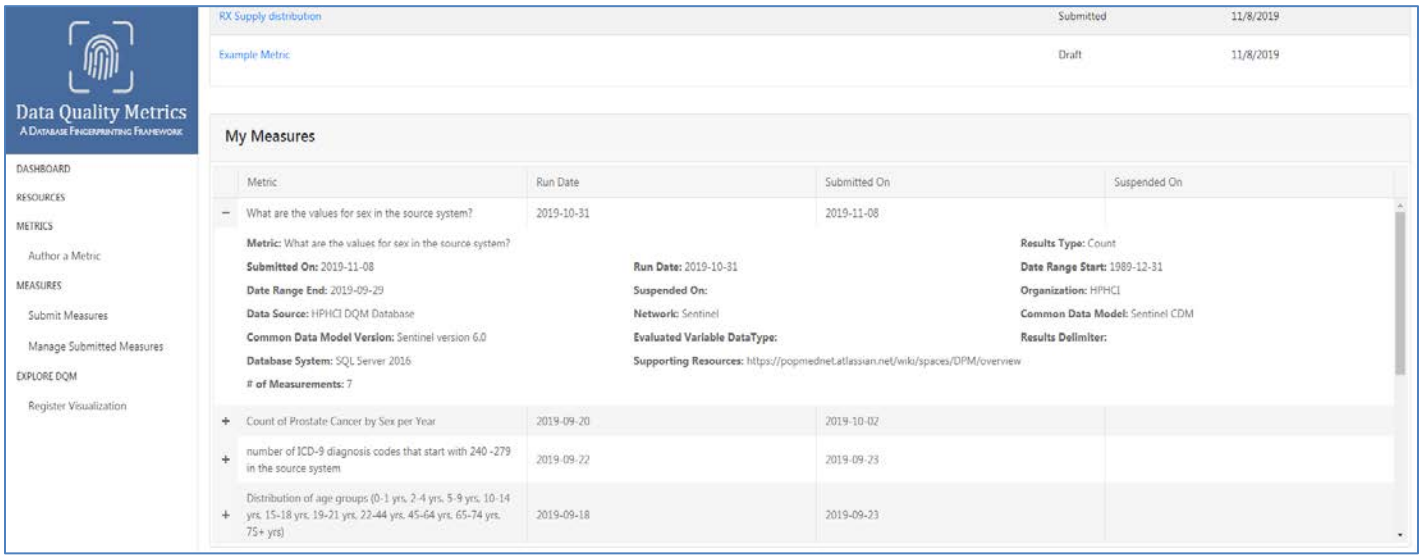
Logged in users can access a list of all Metrics they have submitted to the site by name, status, and date of submission. Filters can be enabled to further specify status:

- All Statuses
- Draft
- Submitted
- In Review
- Published
- Published – requires authentication
- Rejected
- Inactive
- Deleted



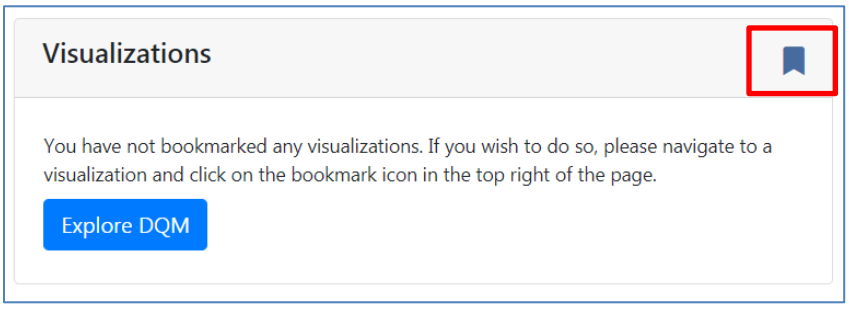
2. My Measures

Logged in users can access a list of all Measures they have submitted to the site. In this section, users can expand each of their submitted measures to see the relevant metadata, such as when the measure was submitted, the date range of the database, database system, etc. The raw data and measurements are not available to view on the Dashboard. The raw data can be viewed in the Measures Drill Down application in Explore DQM.



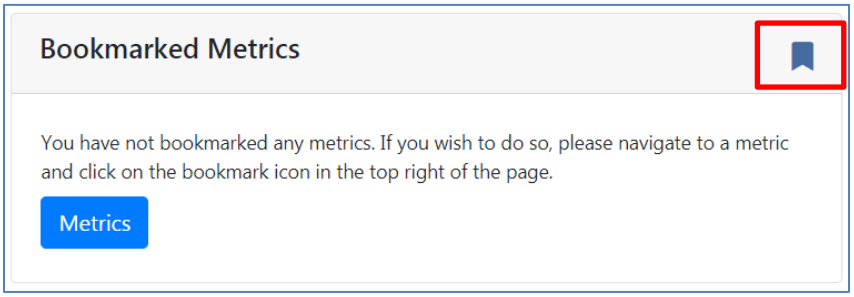
3. Visualizations

Logged in users have the ability to bookmark visualizations of interest. To do so, they must navigate to the Explore DQM section of the website to select a particular visualization, and click the bookmark icon.



4. Bookmarked Metrics

Logged in users have the ability to bookmark Metrics of interest. To do so, they must navigate to the Metrics section of the website to select a particular Metric, and click the bookmark icon.



D. RESOURCES

The Resources page contains information as it relates to the project itself, the framework on which it is based, engagement, and technical resources and details:

1. General

The Data Quality Metrics (DQM) project leverages the data quality harmonization framework (Kahn, 2016) to implement a new platform that enables standardization of data quality metrics and assessment and visualization of data quality output.

2. DQ Harmonization Framework Background

Additional information on the DQ categories and subcategories is provided from the Kahn et al. 2016 manuscript, "Data Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051581/)

3. Project description and funding source

The page includes additional context on the project and details on the funding source.

4. Community engagement

We are utilizing Service Desk tickets (https://popmednet.atlassian.net/jira/servicedesk/projects/DQMCB) to enable continued discussion among community members. Additionally, the recordings of four Stakeholder sessions which include demonstrations of the site are available for the public to view.

5. Technical resources

The project team has developed detailed instructions on how to submit Measures either via the template or the APIs and instructions on how to author a Metric.

6. Link to GitHub for open source software

Open source software and accompanying documentation can be found in the DQM GitHub Repository.

7. Fast Healthcare Interoperability Resources (FHIR)

Information on the project team's investigation of the Fast Healthcare Interoperability Resources (FHIR) standards (https://www.hl7.org/fhir/overview.html) is noted; while we did not formally use FHIR services, there may be opportunities to structure the DQ payload in ways that align with current FHIR data structures.

8. Visualizations

Qlik Sense was selected as the visualization tool for users to explore the characteristics of data sources.

9. Data Model

Diagrams of the data model utilized by the DQM system, as well as documented descriptions

E. METRICS

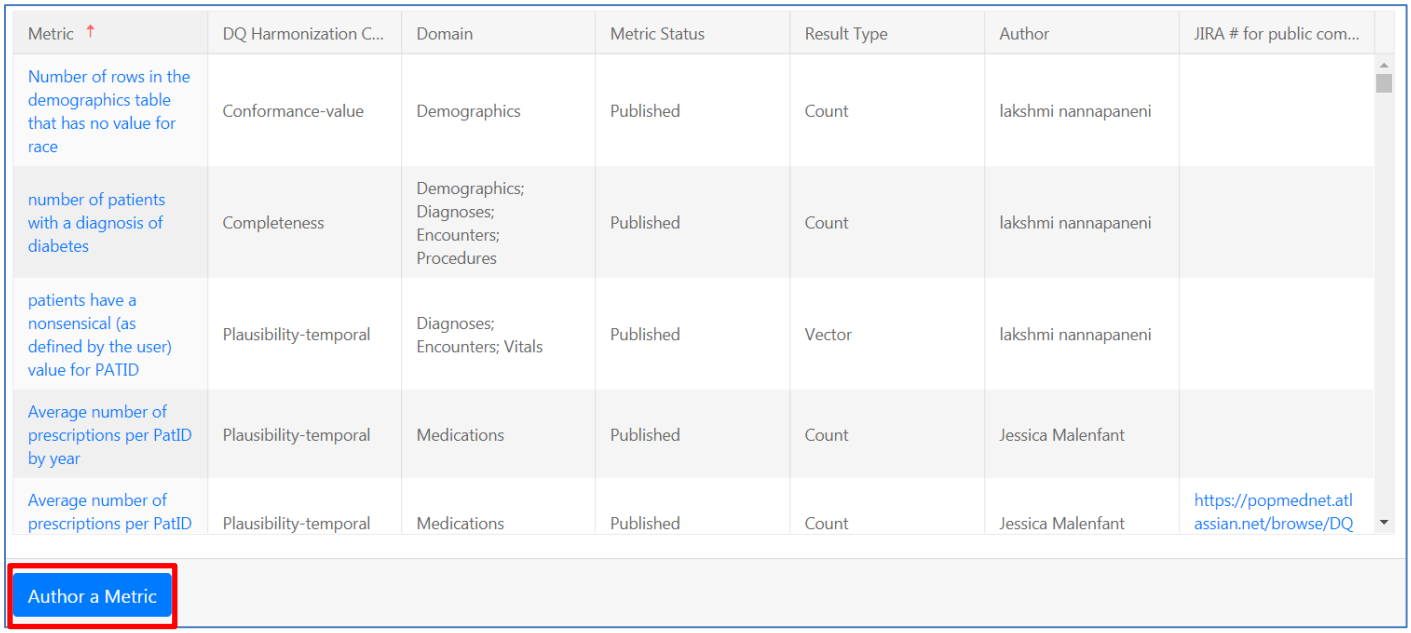
Metrics are the descriptions of quantitative measurements that can be executed on data sources to characterize a specific aspect of the source data in a data model agnostic way. The DQM tool captures metadata about each Metric in a standardized way, regardless of the context or use cases. Metric authors describe the metric in enough detail for a data holder to interpret and generate the results of the Metric from their source data. These results, or measures, enable apples-to-apples comparisons across data sources irrespective of the CDM or data structure.

Each Metric contains a number of required an optional fields, further described in the instructions below for authoring a Metric.

1. Author a Metric

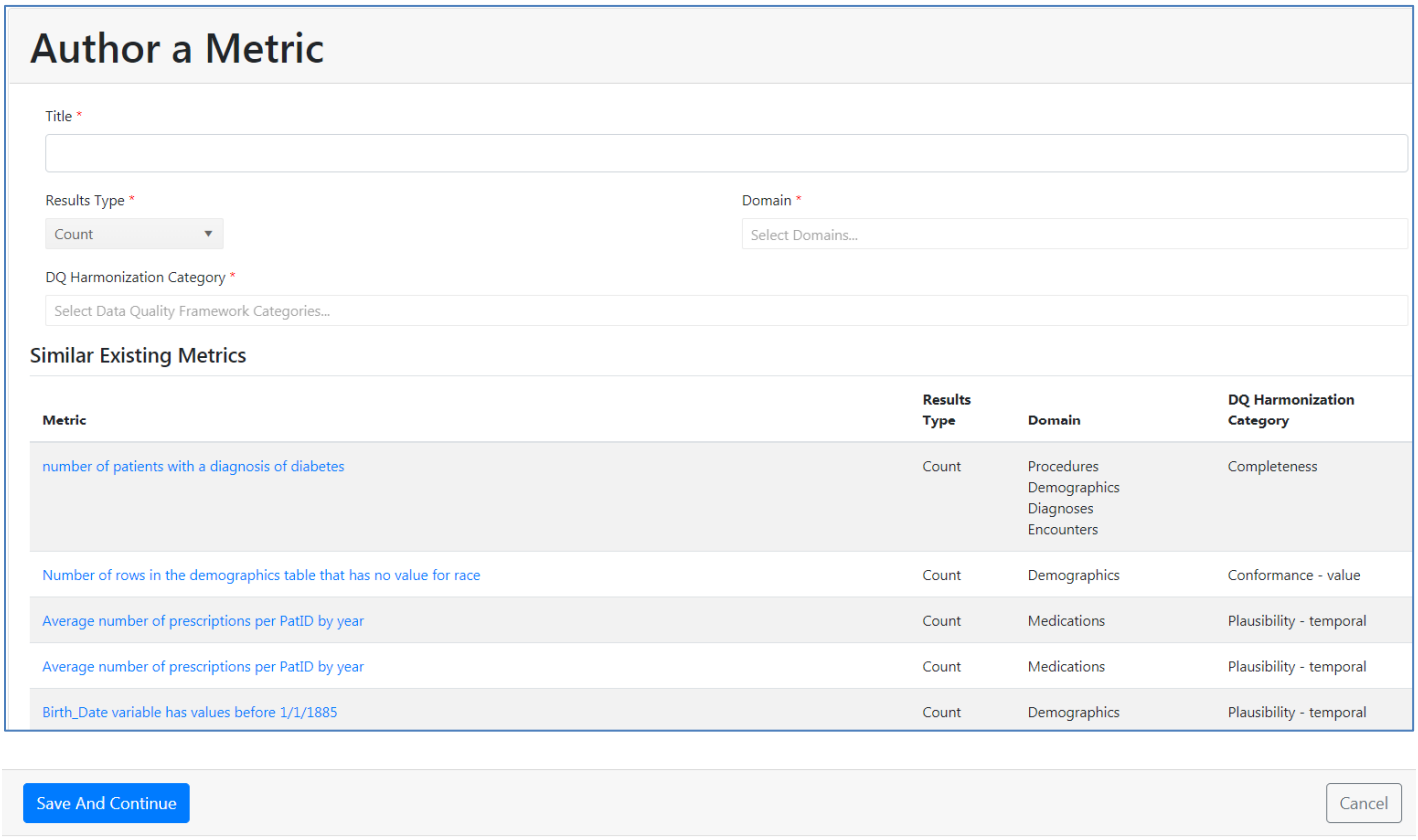
In order to author metrics, users must first register for an account with that ability. Existing users can request an update to their accounts via the DQM Service Desk to be granted access.

To author a Metric, users should first navigate to the Metrics page to review existing metrics.

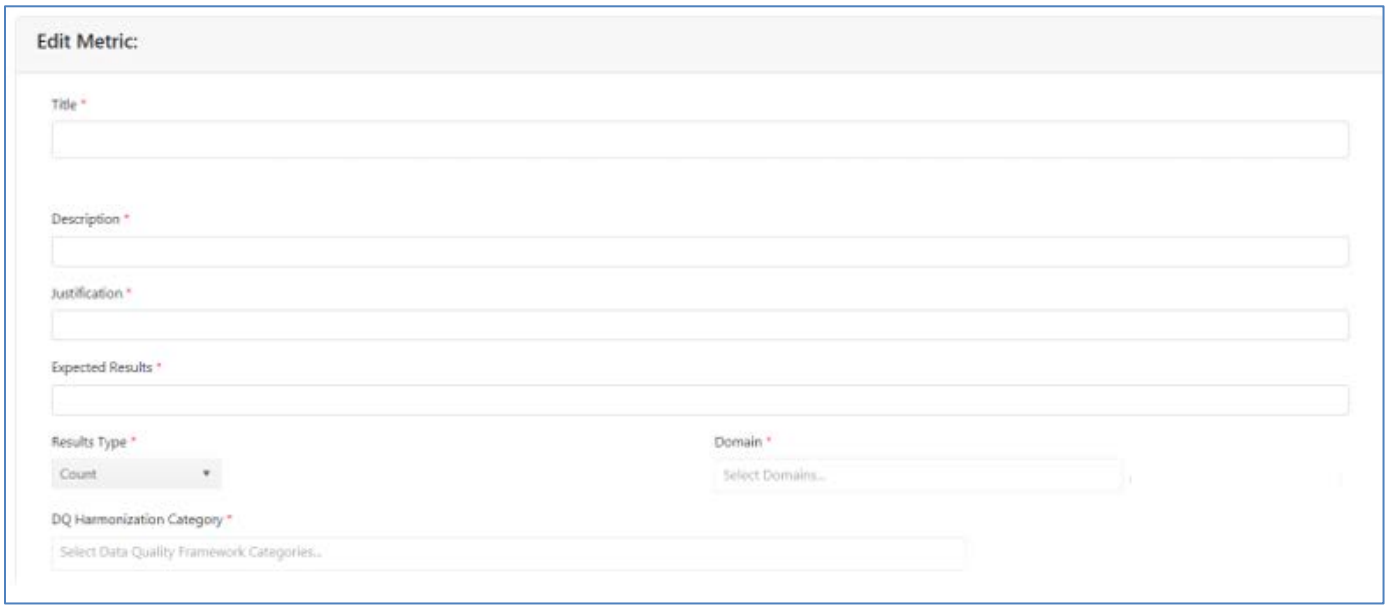


To submit a new metric, click “Author a Metric” and begin by entering a brief description of the Metric. You can then select the Results Type, Domain, and DQ Harmonization Category from the

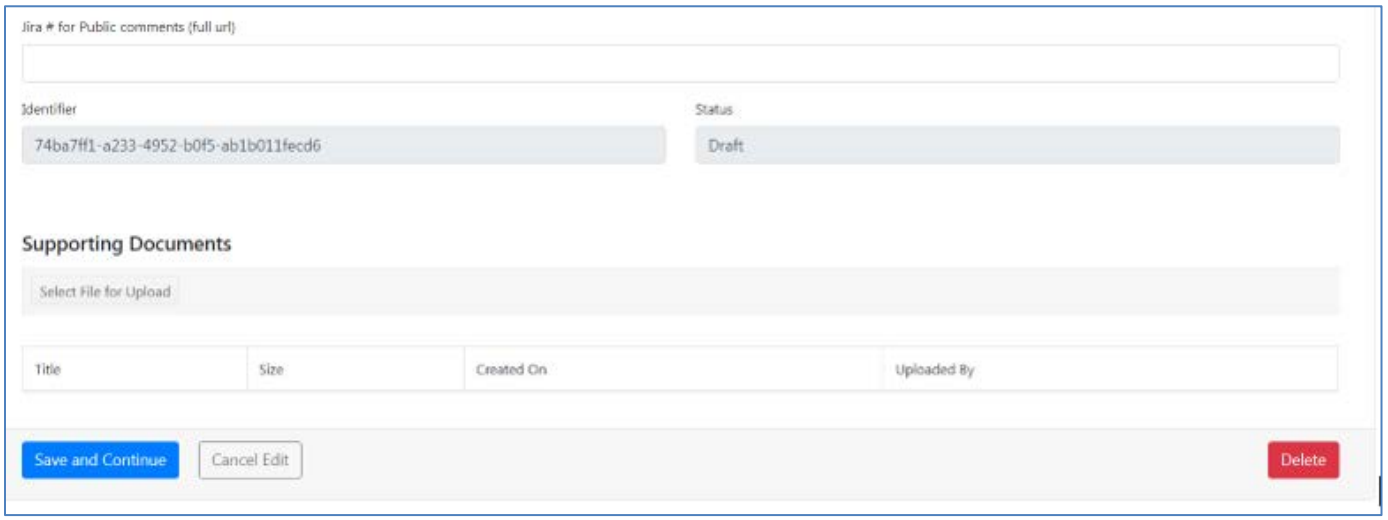
drop-down menus. Additional information on the DQ Harmonization Categories can be found in the Resources page to assist with that selection.



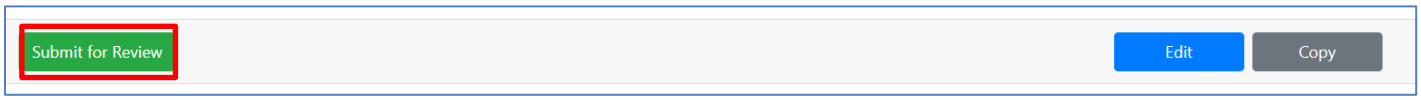
- A list of similar existing metrics will populate the panel below based on the information entered for you to review. Please confirm that this is a new metric and not a duplicate of an existing metric.
- Click “Save and Continue” to move to the Metrics Details form and fill out the following optional fields:
  - Description—details on the purpose of the metric
  - Justification—additional context or reasoning for creation of the metric
  - Expected Results –description of what the author is expecting as a result of executing the metric against a data source
  - Results type
  - JIRA # for Public Comments –a ticket will be created to enable discussion on the specific metric. Users can go to the ticket and share resources and feedback on the particular metric.



- Once the details of the metric have been filled in, select "Save and Continue"



- On the Metric Summary page, choose to either "Submit for Review". You will be able view all of your submitted and draft metrics on your Dashboard.



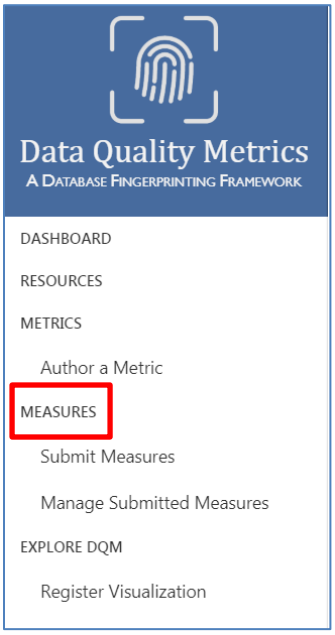
F. MEASURES

In order to submit measures, users must first register for an account with that ability. Existing users can request an update to their accounts via the DQM Service Desk to be granted access.

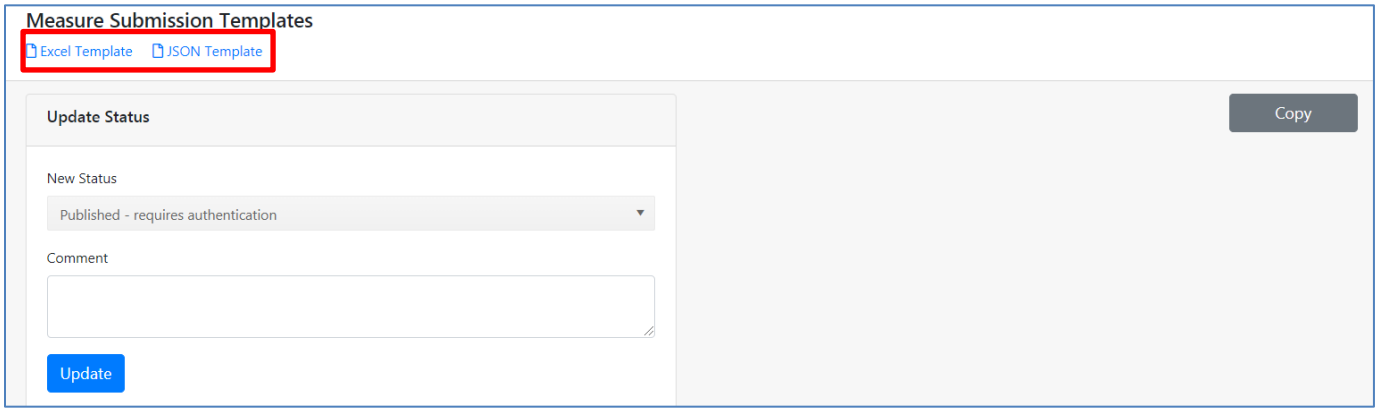
A Measure is the numeric representation of a Metric that has been executed against a data source. Measures include the data characteristics defined in the Metric, as well as metadata about the data source, metric details, and information about when the measurement was calculated. The measures can be explored in the visualization tools.

1. Submit Measures

To submit a Measure, users should first navigate to the Metrics page (https://dataquality.healthdatacollaboration.net/metrics) to select a metric of interest.



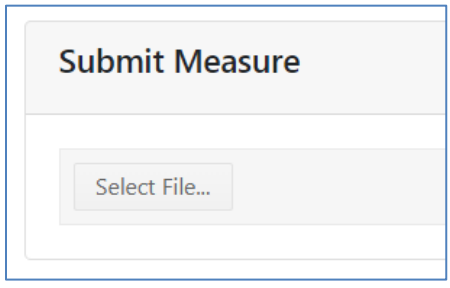
Once selected, scroll down the metric description to locate the attached Measure template and download it.



Populate the provided template with your data according to the following:

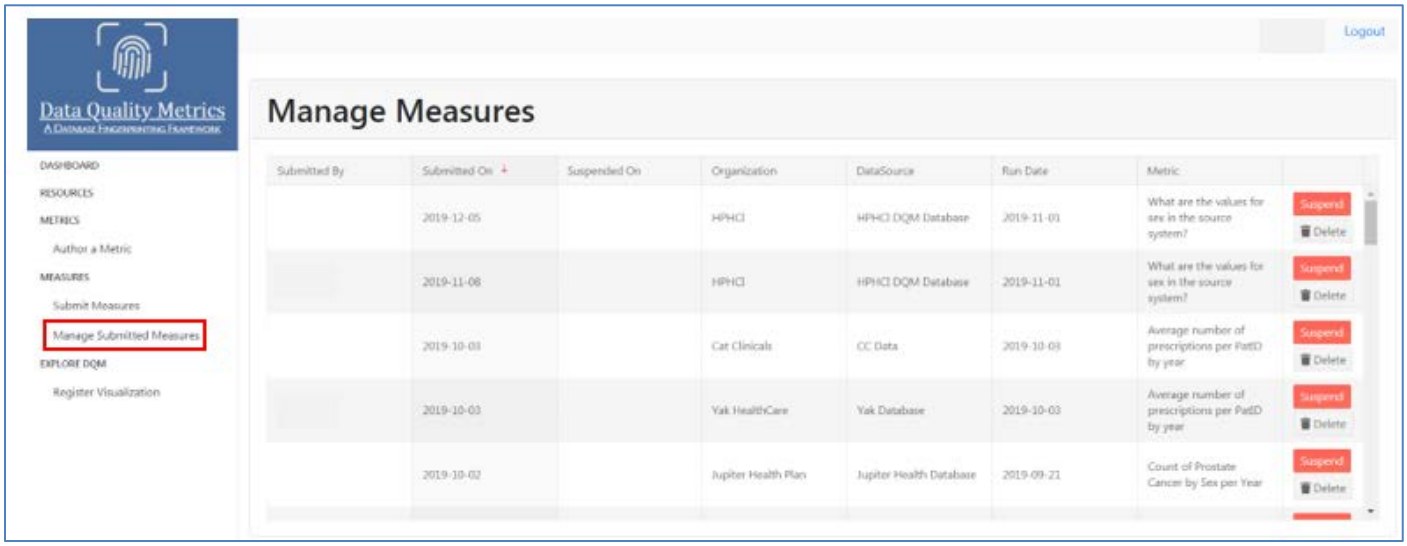
- When you have downloaded the template, fill out Tab 1 according to the included metadata descriptions. Fill out Tab 2 based on the following column definitions:
    - Raw value - predefined value-set. For example, a SEX value set may contain the following: “M”, “F”, “A”, “OT”.

  - Definition - descriptive text for the raw values. Following the above example, the definition for each raw value would be: “Male”, “Female”, “Ambiguous”, and “Other” respectively.
  - Measure - based on the result type (count vs. percentage); result or answer to the metric of interest.
  - Total - overall count/percentage of Measures
- It is necessary to download the template from the specific metric which is being executed.
- Once the template has been populated, navigate to the Submit Measures page (https://dataquality.healthdatacollaboration.net/Home/Index?ReturnUrl=%2Fsubmit-measure). Select the file of interest and submit the completed template.



2. Manage Submitted Measures

This page is only accessible by the DQM site administrators. DQM site administrators can use this page to suspend or delete measures from the system. Data sources and users can request that one or more of their submitted measures be removed either temporarily (suspended) or permanently from the system.



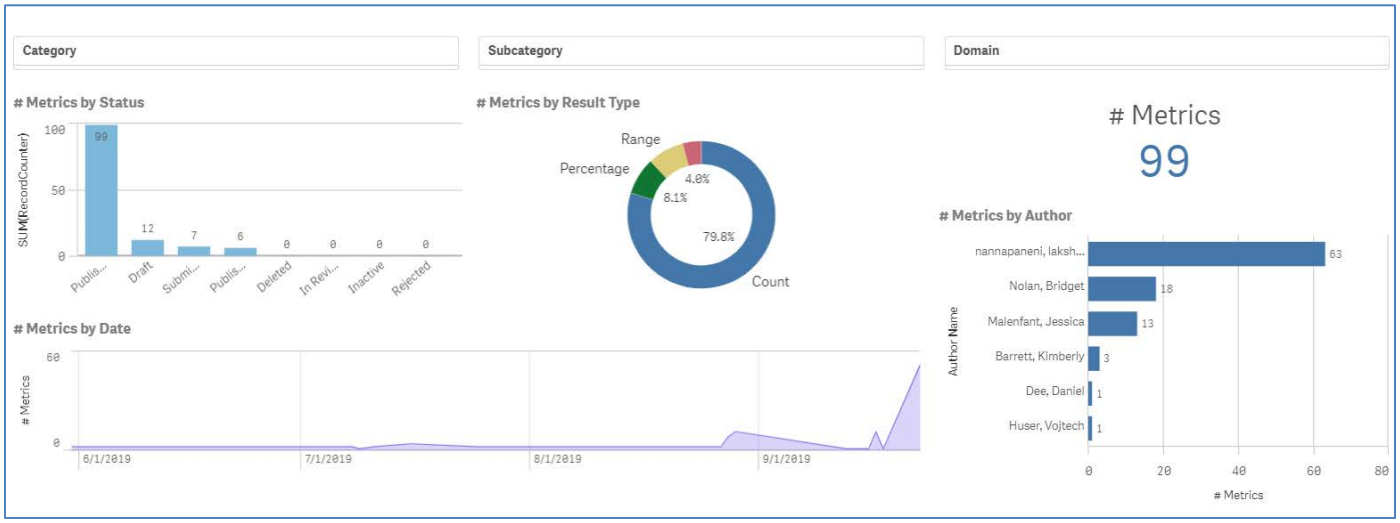
G. EXPLORE DQM

1. Qlik visualizations

Qlik Sense was selected as the visualization tool for users to explore the characteristics of data sources. Qlik can connect to data sources using standard APIs, and the assumption is that other analytic tools able to load data via an API (e.g. Tableau) could be used in place of Qlik. Technical documentation on Qlik and the available APIs are posted in the GitHub repository.

A number of apps have been developed to visualize metadata about the DQM system and a set of use cases as selected by the project team:

- **Data Quality Metrics Dashboard**
(https://dataquality.healthdatacollaboration.net/visual/aa366737-48aa-4e6c-8bc6-aae1015e2ae3)
A top-level view of the Metric submission metadata

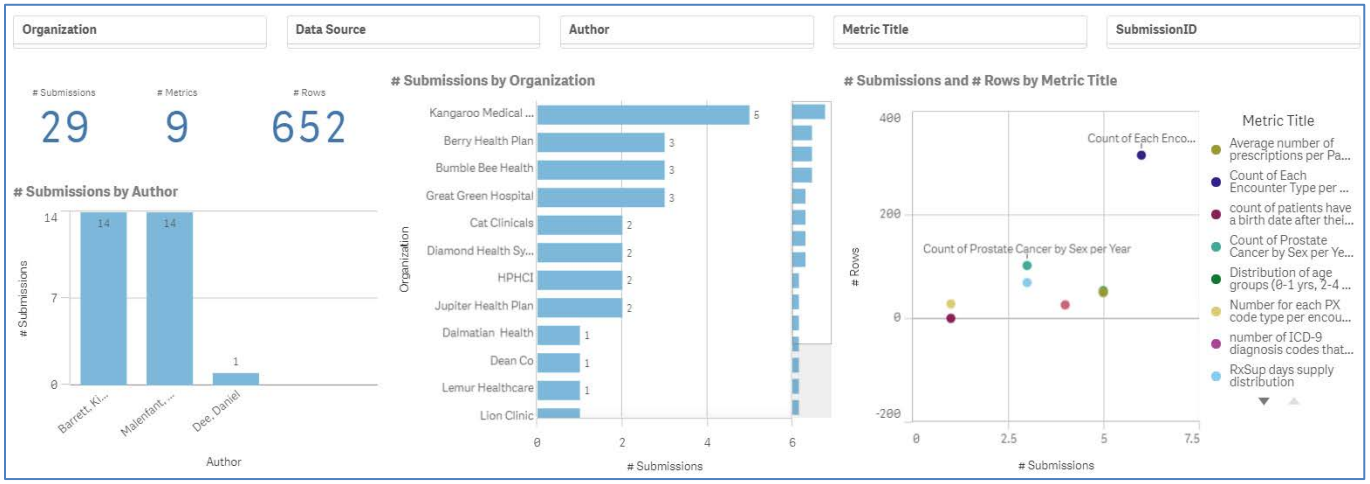


- **Data Quality Metrics Drill Down**

(https://dataquality.healthdatacollaboration.net/visual/6166a651-71c0-4d62-9b94-aae300edecae)
Drill down to see the metadata of an individual metric submission

- **Data Quality Measures Dashboard**

(https://dataquality.healthdatacollaboration.net/visual/c2f51fa4-8f0b-4512-8972-aae300eea9b9)
Dashboard view of the data quality measures metadata

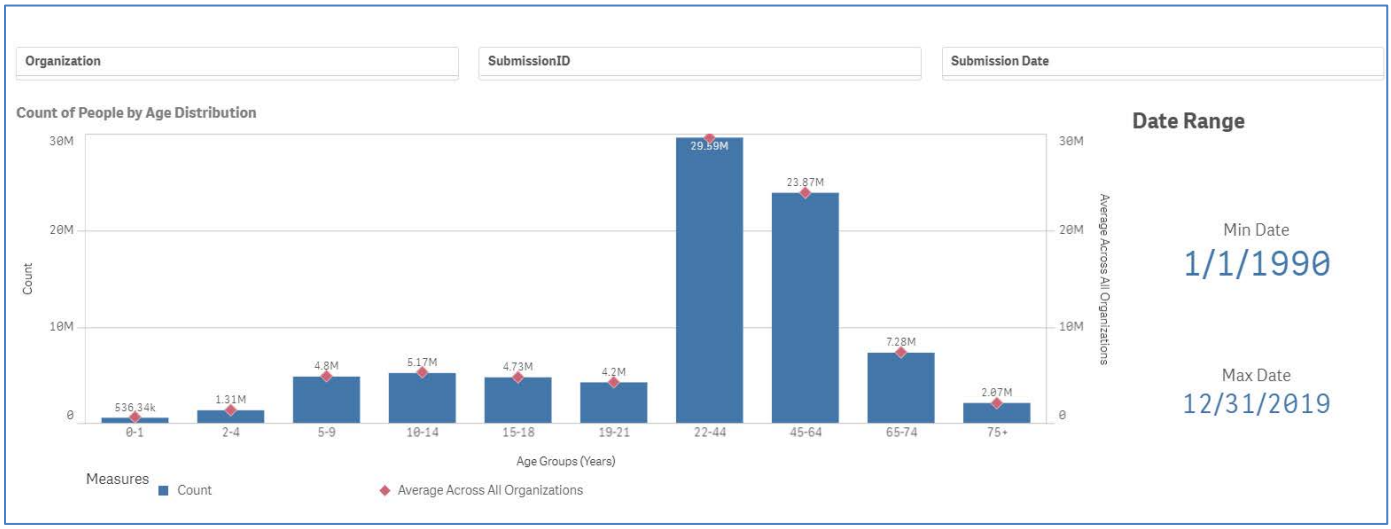


- **Data Quality Measures Drill Down**

(https://dataquality.healthdatacollaboration.net/visual/0349a2bb-b36d-4057-a08e-aae300ef5821)
Drill down to see the raw data for a single measure submission. To see the data, you must filter down to a single submission.

- **Age Distribution Metrics**

(https://dataquality.healthdatacollaboration.net/visual/d69b8cd4-1a86-4425-90e8-aae300f0102d)
Compare an organizations age distribution data against an average of all the age distribution data.

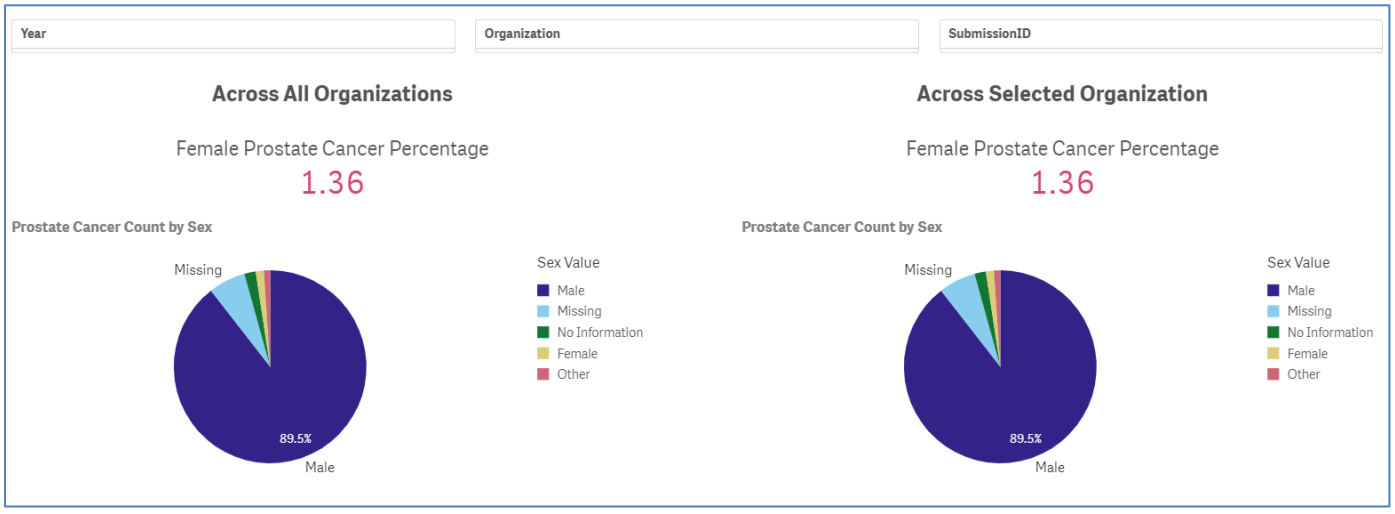


- **RX Days Supply Distribution**

(https://dataquality.healthdatacollaboration.net/visual/9e1a7337-210d-4273-93c9-aae300f0d674)
See the prescription count for each days supply metric, broken up by organization.

- **Prostate Cancer by Sex Per Year**

(https://dataquality.healthdatacollaboration.net/visual/c939fc5d-0338-4c98-98c8-aae300f1dfa1)
See the rate of prostate cancer by sex per year. The left side shows the average across all organizations, and the right side allows you to filter to a specific organization's data and a specific year.



- **Encounter Type Per Year-Month**

(https://dataquality.healthdatacollaboration.net/visual/4b3b02ee-2e1a-4919-bb33-aae300f2b5d6)
View the encounter count broken down by encounter type. To see the line chart, select a single organization at a time; you can also filter by year and by encounter type.

- **Average Number of Prescriptions Per Patient By Year**

(https://dataquality.healthdatacollaboration.net/visual/be186ff8-6f21-41a9-b109-aae300f9b989)
See the average number of prescriptions per patient per year. To see the chart, filter to a single organization submission.

2. Register visualization

As part of potential future work, we have enabled the ability to "Register Visualization" to add more visualizations to the DQM system. This function requests a title, App ID, a sheet-level ID, and a description. We envision that this would be the responsibility of a Coordinating Center in an operationalized version of the system; additional details are contained in the project's Technical Documentation.


Register Visualization
Title:*
App ID:*
Sheet ID:
Description:
Requires Authentication
Published
Register Visualization
Cancel