ASPE-4 Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data Project

User Documentation

Prepared by: Sentinel Coordinating Center

**December 31, 2019**

ASPE-4 is funded by the US Food and Drug Administration through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I / HHSF22301006T.

**Table of Contents**

[I. background 1](#_Toc26982617)

[A. system overview 1](#_Toc26982618)

[B. Key Functional Components 2](#_Toc26982619)

[1. Metrics 2](#_Toc26982620)

[2. Measures 2](#_Toc26982621)

[3. Explore DQM 2](#_Toc26982622)

[II. Functionality 2](#_Toc26982623)

[A. Register 3](#_Toc26982624)

[B. login 3](#_Toc26982625)

[C. Dashboard 4](#_Toc26982626)

[1. My Metrics 4](#_Toc26982627)

[2. My Measures 5](#_Toc26982628)

[3. Visualizations 5](#_Toc26982629)

[4. Bookmarked Metrics 6](#_Toc26982630)

[D. resources 6](#_Toc26982631)

[1. General 6](#_Toc26982632)

[2. DQ Harmonization Framework Background 6](#_Toc26982633)

[3. Project description and funding source 6](#_Toc26982634)

[4. Community engagement 6](#_Toc26982635)

[5. Technical resources 7](#_Toc26982636)

[6. Link to GitHub for open source software 7](#_Toc26982637)

[7. Fast Healthcare Interoperability Resources (FHIR) 7](#_Toc26982638)

[8. Visualizations 7](#_Toc26982639)

[9. Data Model 7](#_Toc26982640)

[E. metrics 7](#_Toc26982641)

[1. Author a Metric 7](#_Toc26982642)

[F. Measures 10](#_Toc26982643)

[1. Submit Measures 11](#_Toc26982644)

[2. Manage Submitted Measures 12](#_Toc26982645)

[G. Explore dqm 13](#_Toc26982648)

[1. Qlik visualizations 13](#_Toc26982649)

[2. Register visualization 15](#_Toc26982650)

# 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.

## 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 al1. 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 2-6{, #94;Curtis, 2014 #52;Fleurence, 2014 #51;Vogel, 2014 #76;, 2018 #93}.

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 7-12. 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)

## Key Functional Components

### [Metrics](#_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](#_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](#_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.

# Functionality

The DQM System was instantiated as a [web portal](https://dqmetrics.healthdatacollaboration.net) with multiple pages of functionality.

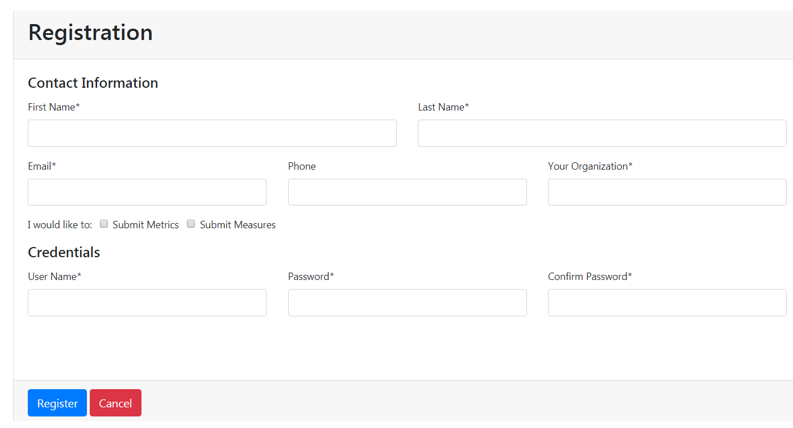
## Register

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



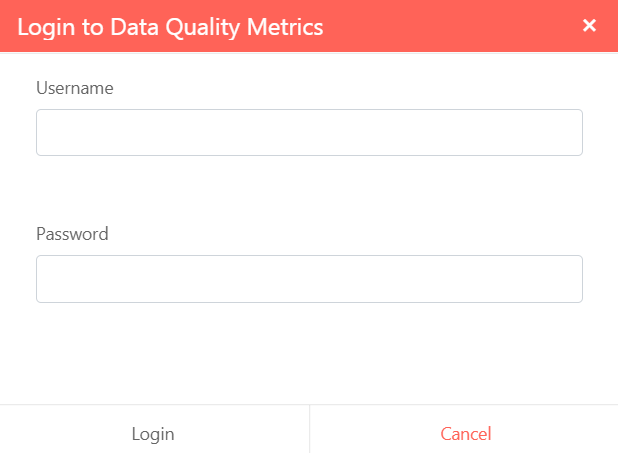
Requested information includes:

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



## login

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



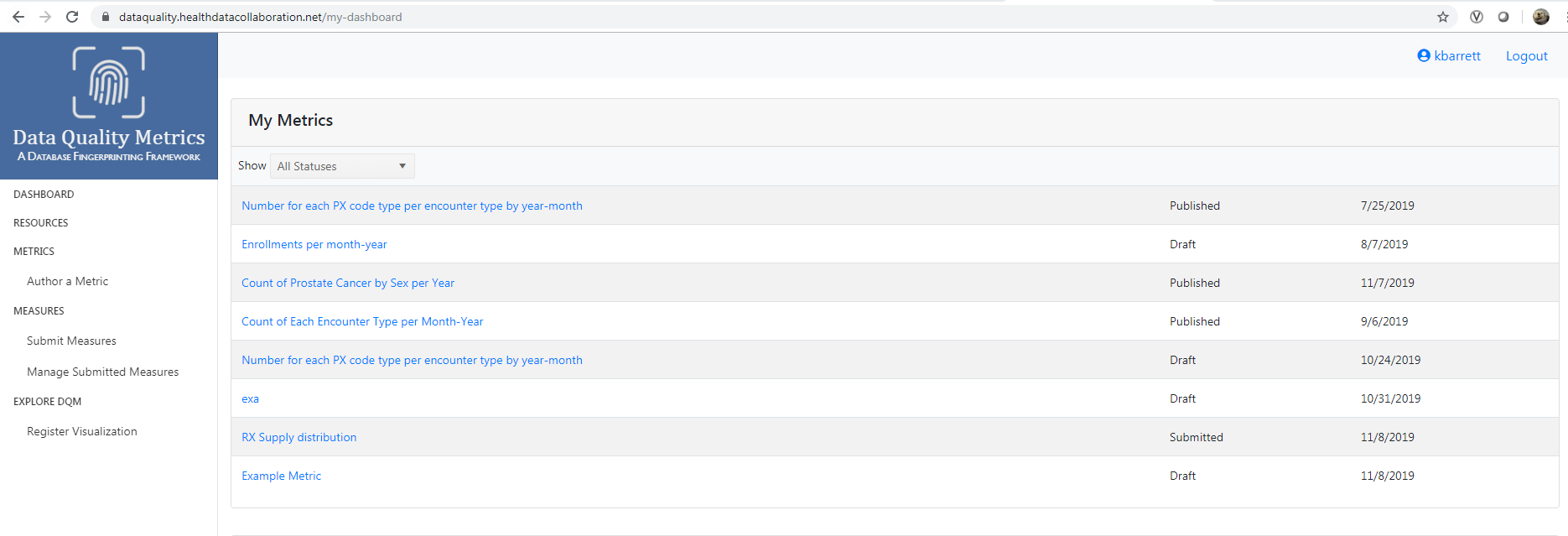
## Dashboard

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.

### My Metrics

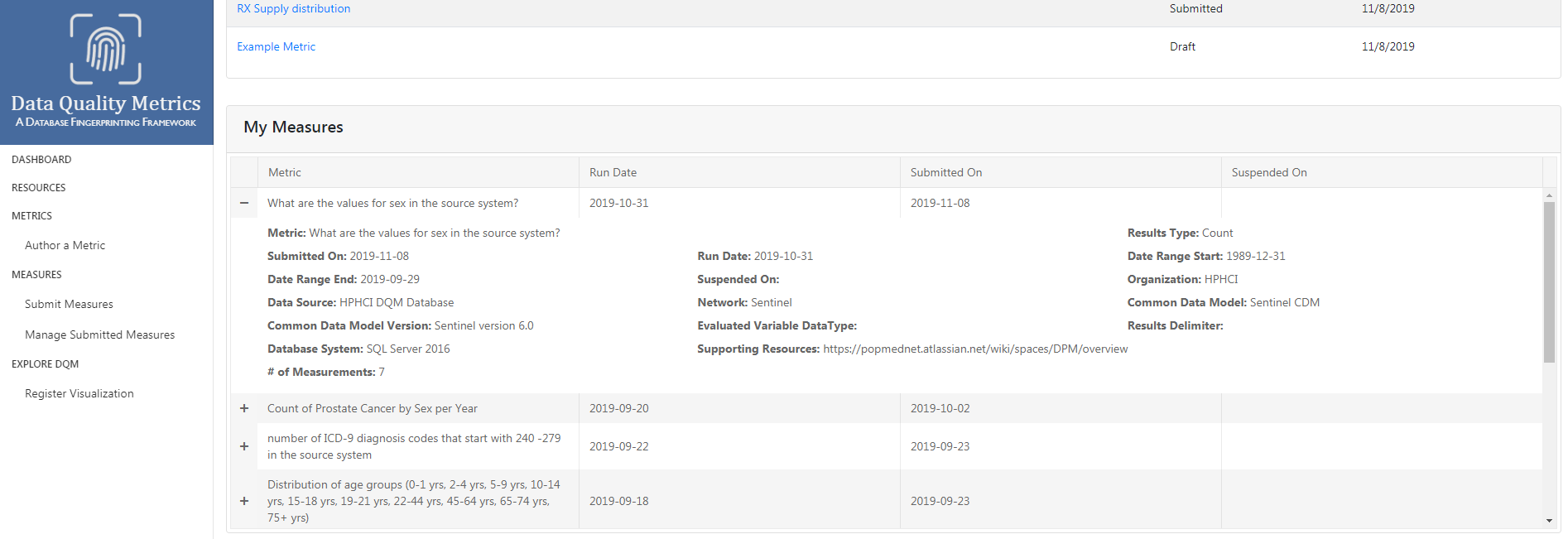
Logged in users can access a list of all Metrics they have submitted to the site by name, status, and date of submission. Filters can be enabled to further specify status:

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



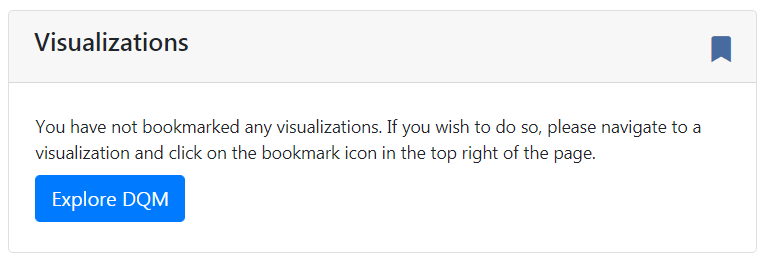
### My Measures

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.



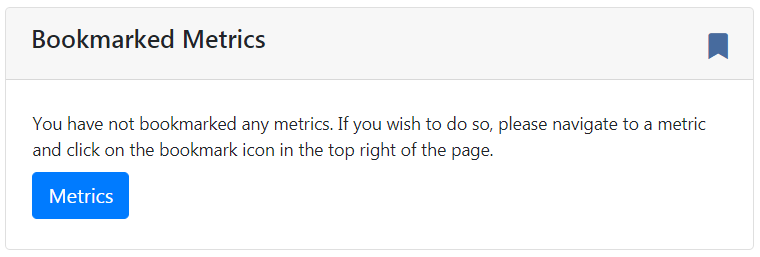
### Visualizations

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.



### 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.



## 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:

### 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.

### 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/)”

### Project description and funding source

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

### Community engagement

We are utilizing [Service Desk tickets](https://popmednet.atlassian.net/projects/DQMCB) to enable continued discussion among community members.

### Technical resources

The project team has developed detailed instructions on how to [submit Measures](#_Submit_Measures) either via the template or the APIs and instructions on how to [author a Metric](#_Author_a_Metric).

### Link to GitHub for open source software

Open source software and accompanying documentation can be found in the [DQM GitHub Repository](https://github.com/PopMedNet-Team/DataQualityMetrics).

### 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.

### Visualizations

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

### Data Model

Diagrams of the data model utilized by the DQM system, as well as documented descriptions of relationships between entities, are provided for reference.

## 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.

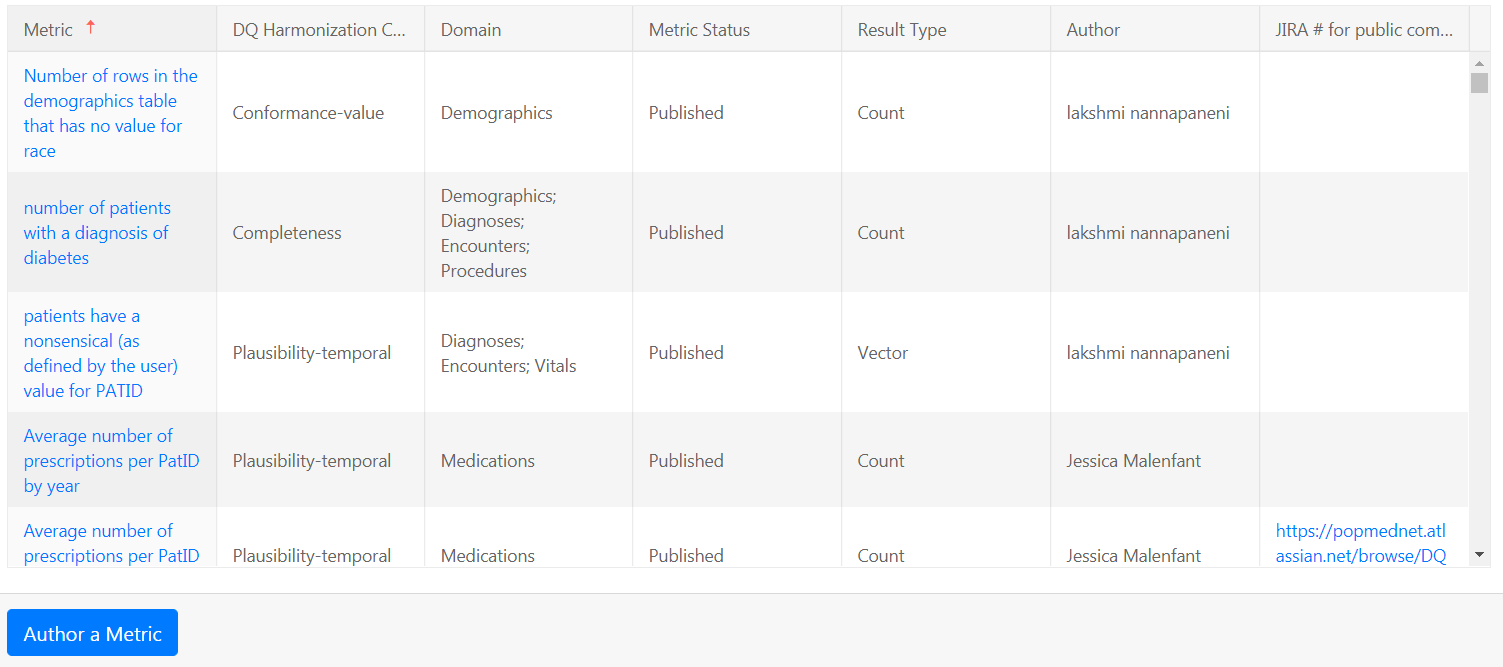
Each Metric contains a number of required an optional fields, further described in the instructions below for authoring a Metric.

### Author a Metric

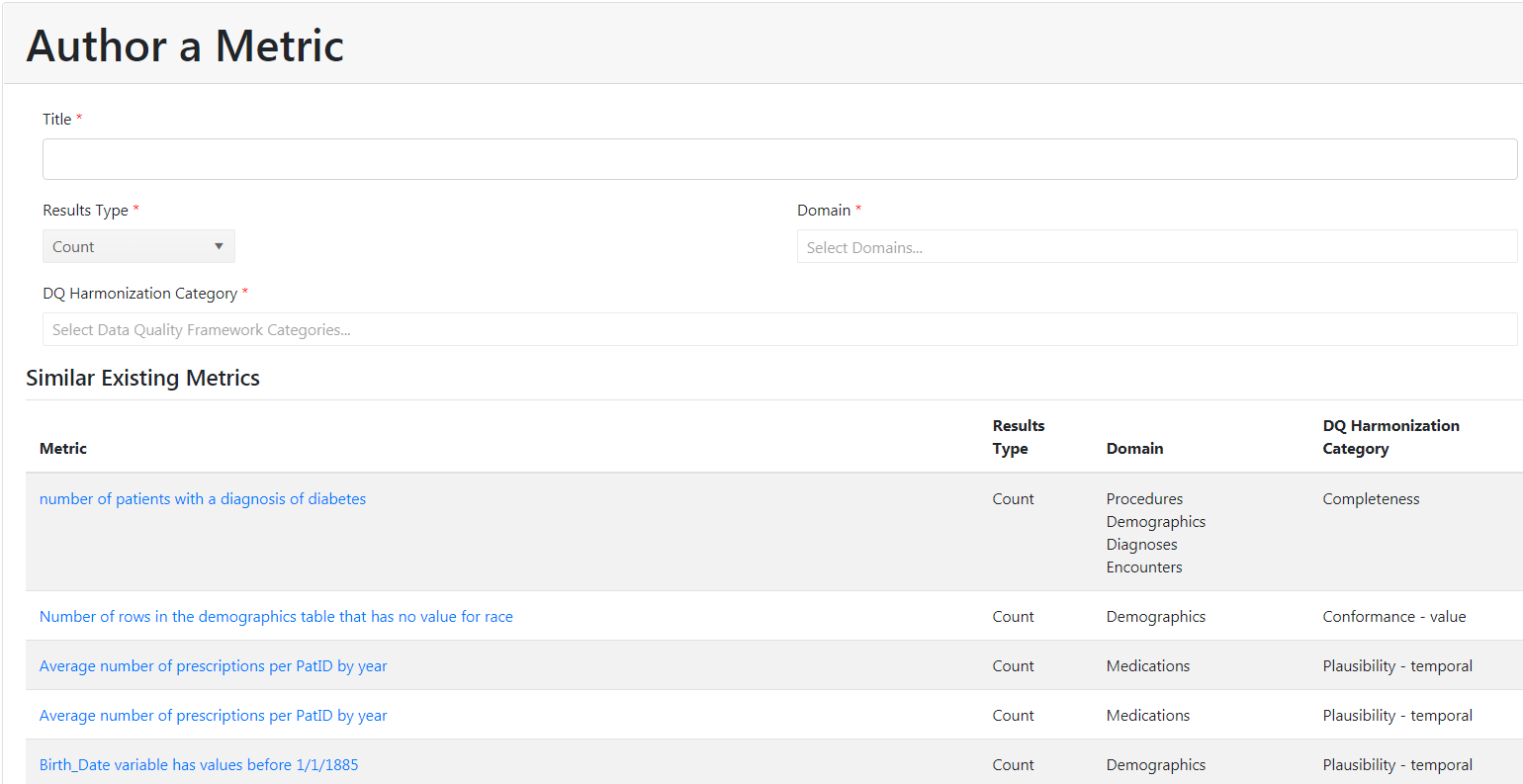
In order to author metrics, users must first register for an account with that ability. Existing users can request an update to their accounts via the [DQM Service Desk](https://popmednet.atlassian.net/servicedesk/customer/portal/6) to be granted access.

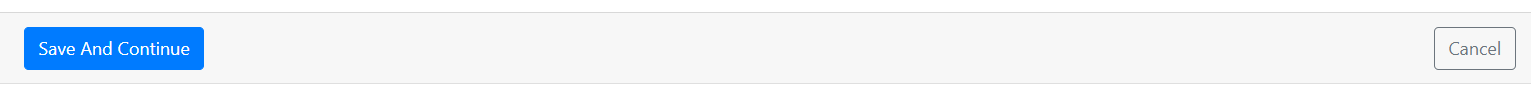
To author a Metric, users should first navigate to the Metrics page to review existing metrics.



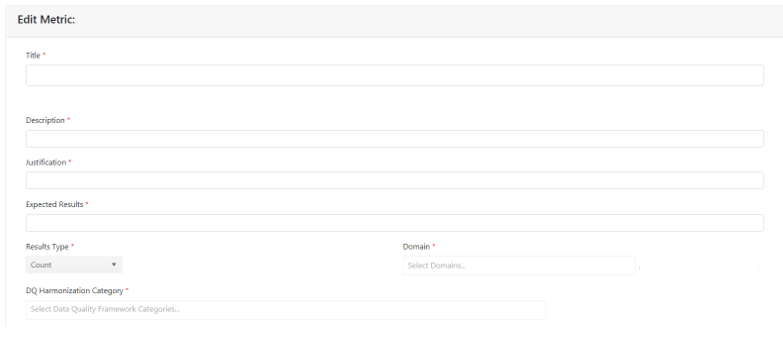


To submit a new metric, click “Author a Metric” and begin by entering a brief description of the Metric. You can then select the Results Type, Domain, and DQ Harmonization Category from the drop-down menus. Additional information on the DQ Harmonization Categories can be found in the Resources page to assist with that selection.

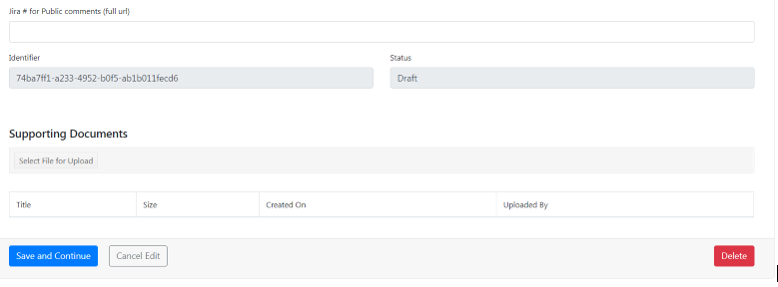




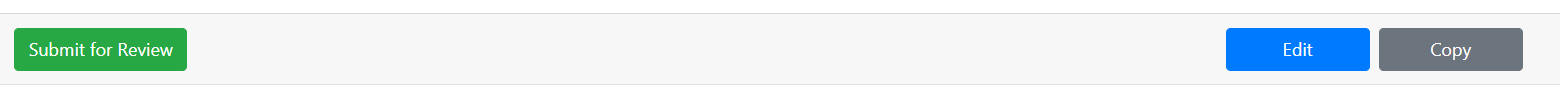
* A list of similar existing metrics will populate the panel below based on the information entered for you to review. Please confirm that this is a new metric and not a duplicate of an existing metric.
* Click “Save and Continue” to move to the Metrics Details form and fill out the following optional 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.



## 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](https://popmednet.atlassian.net/servicedesk/customer/portal/6) to be granted access.

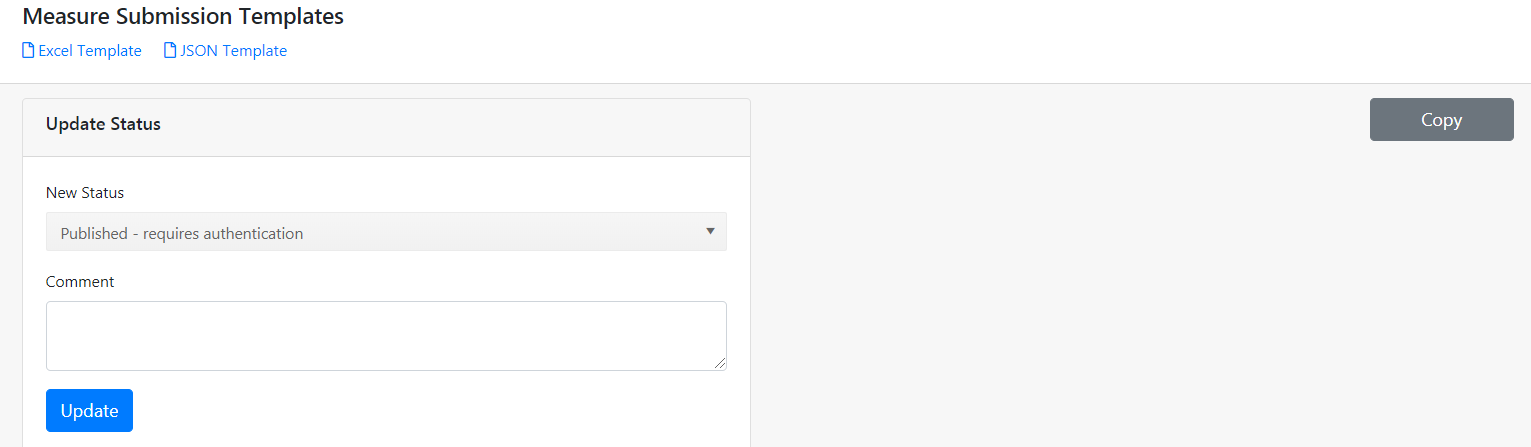
A Measure is the numeric representation of a Metric that has been executed against a data source. Measures include the data characteristics defined in the Metric, as well as metadata about the data source, metric details, and information about when the measurement was calculated. The measures can be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

### 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.



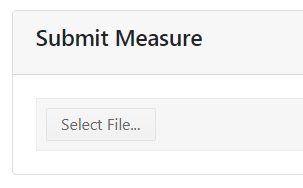
Once selected, scroll down the metric description to locate the attached Measure template and download it.



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/submit-measure). Select the file of interest and submit the completed template.





### 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.

### 

## Explore dqm

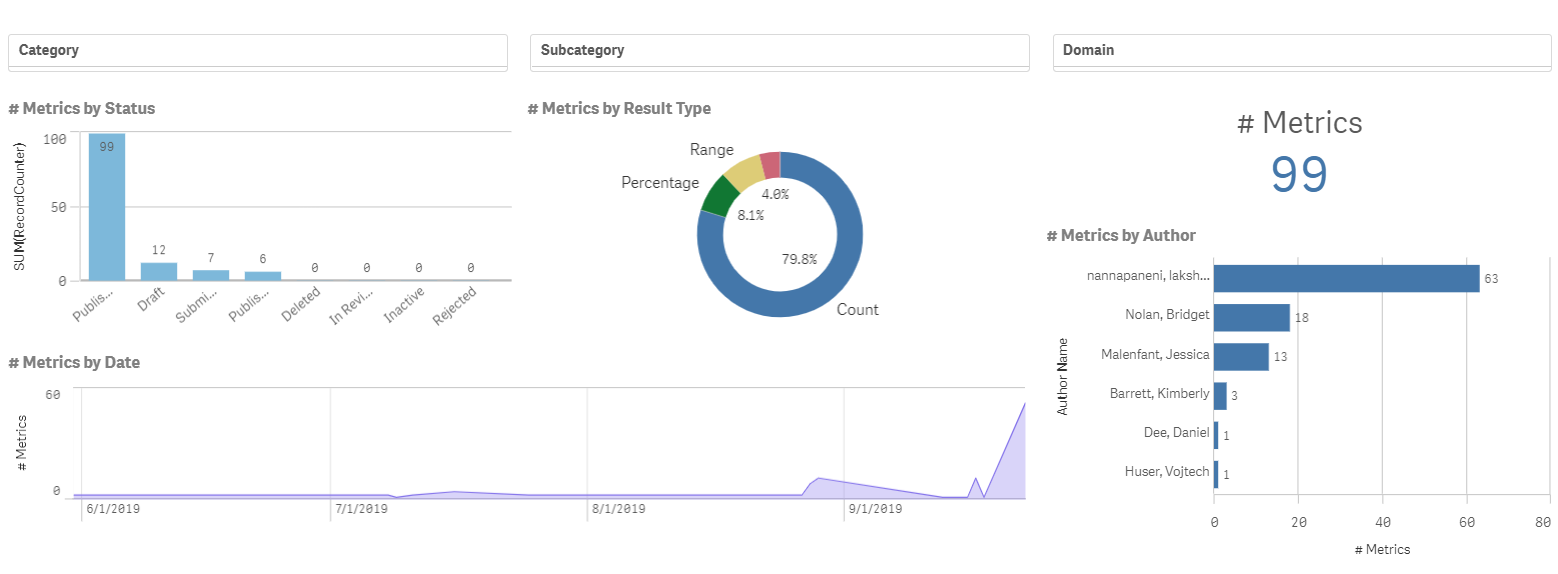
### Qlik visualizations

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](https://github.com/PopMedNet-Team/DataQualityMetrics).

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

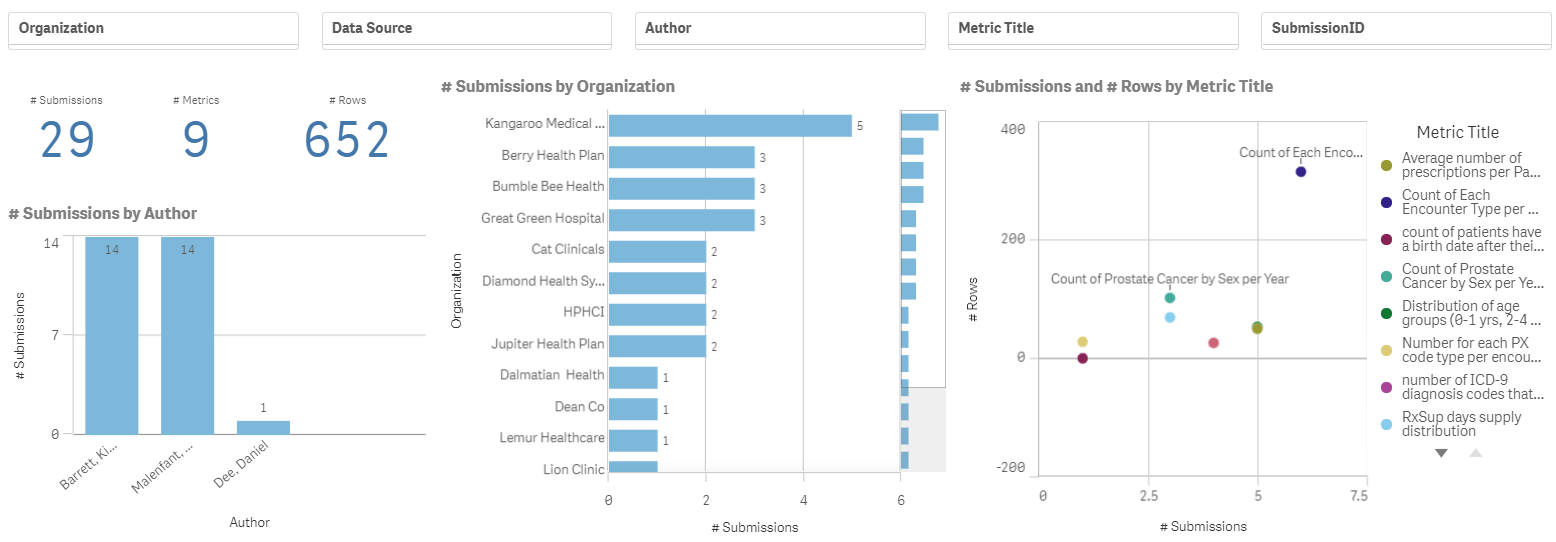


#### [Data Quality Metrics Drill Down](https://dataquality.healthdatacollaboration.net/visual/6166a651-71c0-4d62-9b94-aae300edecae)

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

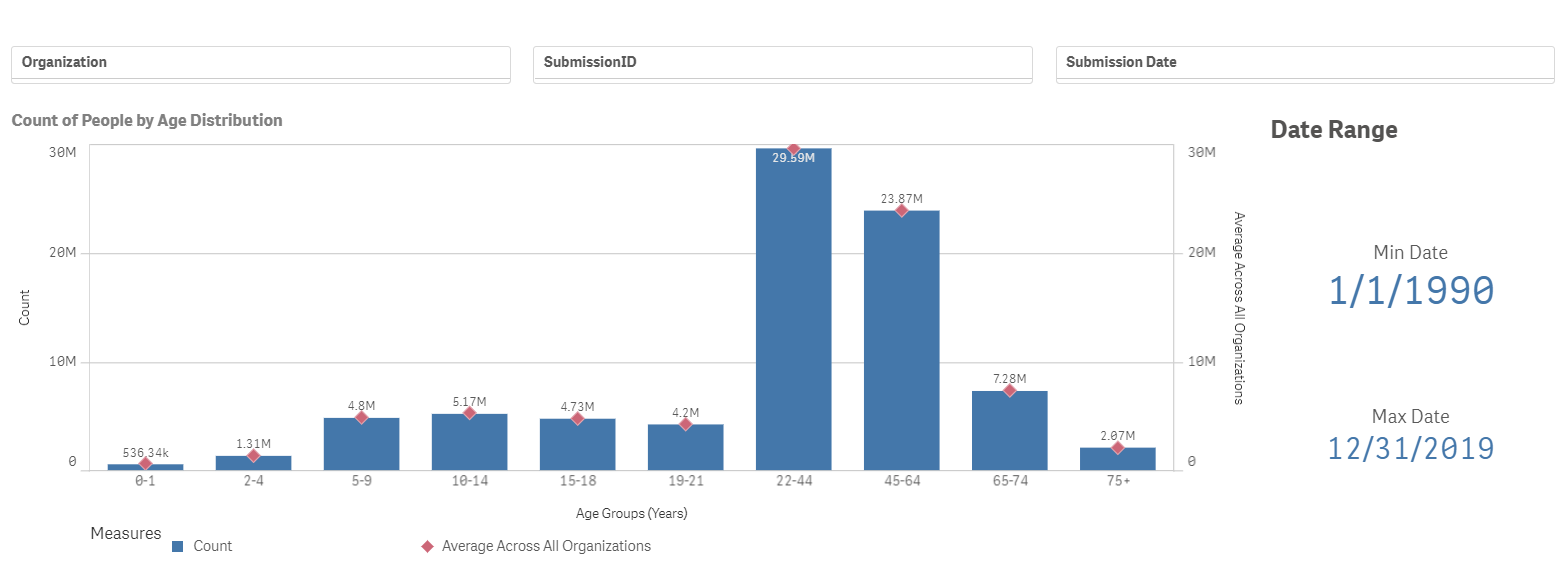


#### [Data Quality Measures Drill Down](https://dataquality.healthdatacollaboration.net/visual/0349a2bb-b36d-4057-a08e-aae300ef5821)

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.

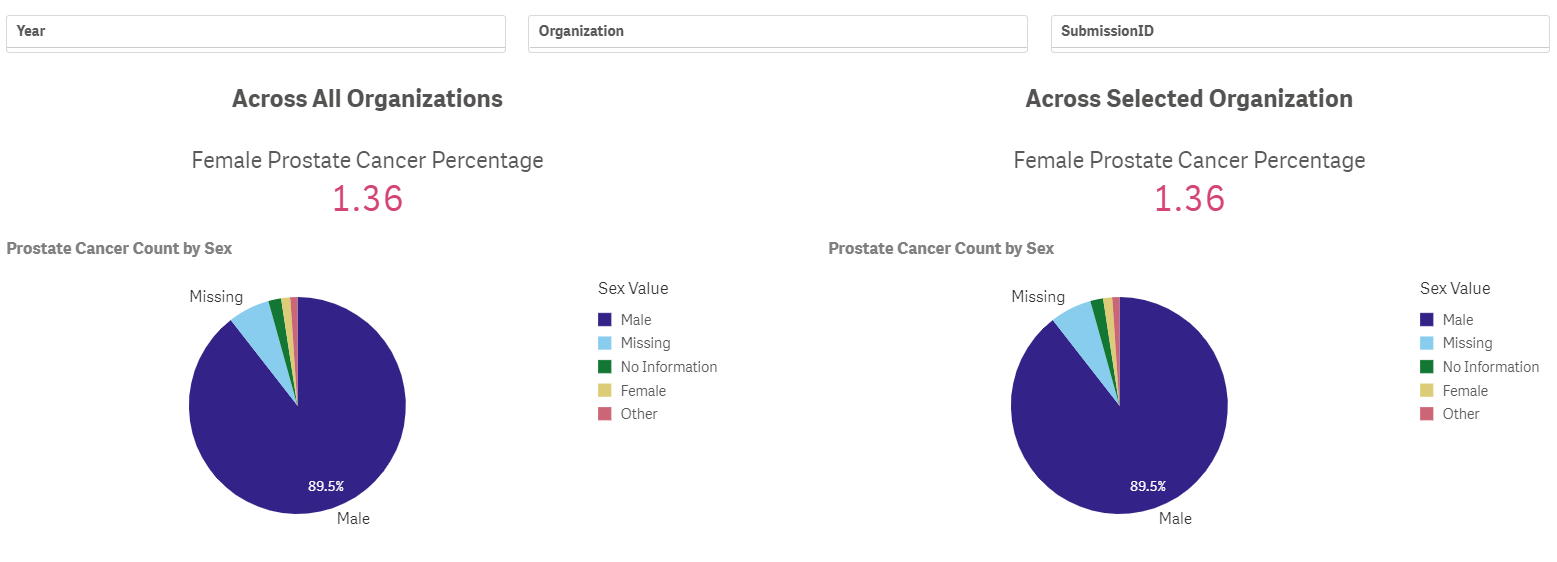


#### [RX Days Supply Distribution](https://dataquality.healthdatacollaboration.net/visual/9e1a7337-210d-4273-93c9-aae300f0d674)

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)

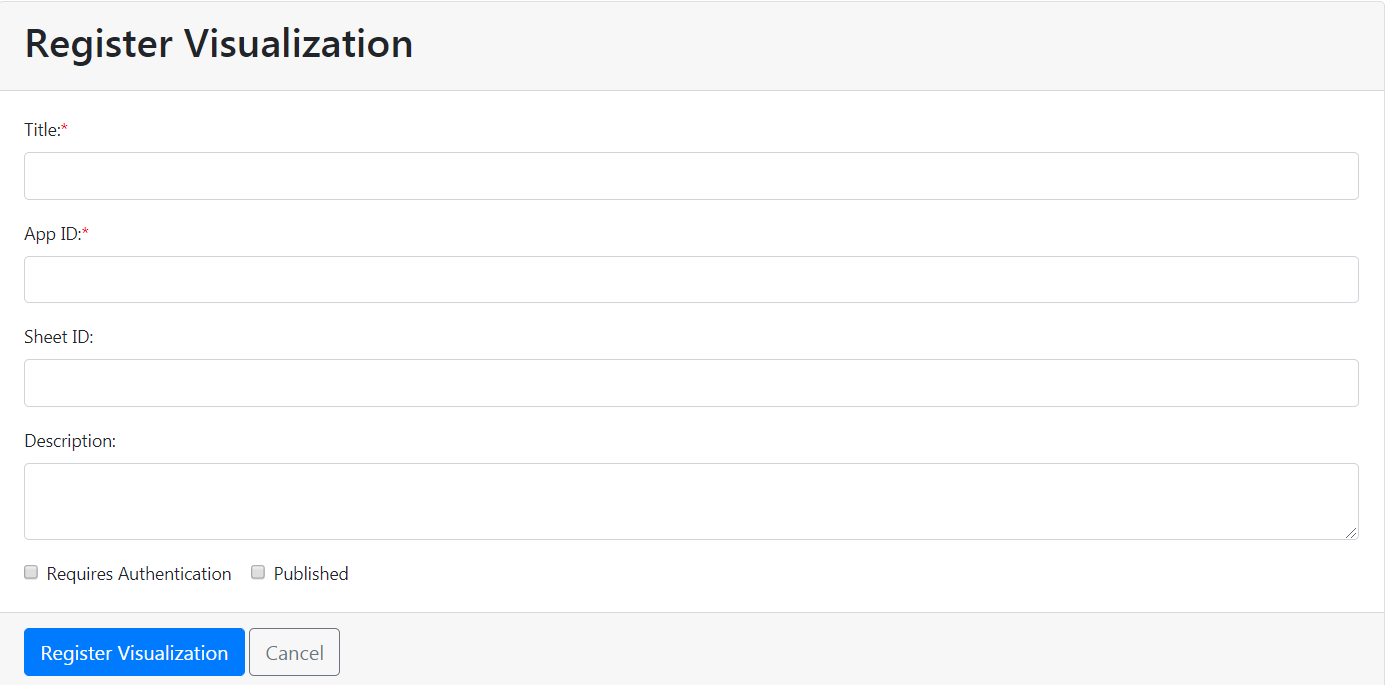
View the encounter count broken down by encounter type. To see the line chart, select a single organization at a time; you can also filter by year and by encounter type.

#### [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.

### 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.



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

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

4. 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.

5. 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.

6. 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.

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.