Sentinel

Assessing Data Source Characteristics in Multi-site Analyses

Jeffrey S. Brown, PhD

Harvard Pilgrim Health Care Institute and Harvard Medical School

September 9, 2020

Problem & Purpose

- 10+ years of public funding to support health data networks
- There is no central place to find out about or how to engage with these growing data systems
- Networks have different types of data and varying data quality processes and definitions
- There are **no standard metrics for describing data across systems**

Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data
"Database fingerprinting framework"

https://aspe.hhs.gov/standardization-and-querying-data-quality-metrics-and-characteristics-electronic-health-data

Aim: create the infrastructure to curate, explore, and author standardized DQ metrics

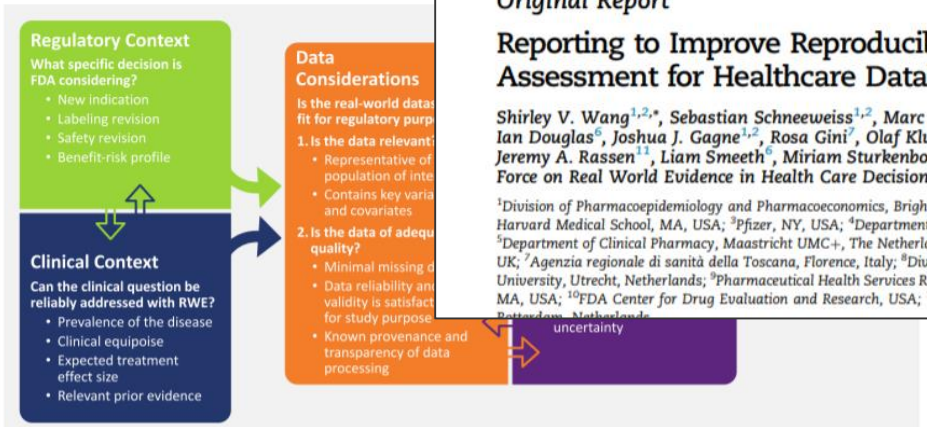
- Standard approach to capturing and sharing Data Quality Metrics (DQM) and associated measures
- Help researchers evaluate fitness for purpose across data sources
- Pilot Goals:
  - Operationalize the leading theoretical DQ harmonization framework* by developing a website that enables the capture and curation of DQMs and facilitates exploration of DMQ measures
  - Create a beta version of the platform with Sentinel and PCORnet as use cases
  - Collaborate with existing DQ stakeholder community and incorporate feedback on tools developed
  - Disseminate the platform as open source tools

*Kahn et al. 2016 DOI:http://dx.doi.org/10.13063/2327-9214.1244

DQM establishes a platform for the community to define and curate DQ metrics in standard ways

Duke | MARGOLIS CENTER for Health Policy

Figure 1. Considerations for gene



As manufacturers consider a RWE development strategy to support regulatory use, there are a number of considerations that should be addressed to ensure that an RWE approach is sensible. First, it is critical to examine the intended regulatory use and the clinical context within which RWE will be developed. Second, the strength of available RWD data sources and study methods for generating RWE that is fit for regulatory purposes must be considered. Matching data sources and appropriate methods to answer specific clinical and regulatory questions will result in different "types" of RWE for different use cases (figure modified from the 2017 paper).

https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf

VALUE IN HEALTH 20 (2017) 1009–1022

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Original Report

Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0

Shirley V. Wang[1,2,*], Sebastian Schneeweiss[1,2], Marc L. Berger[3], Jeffrey Brown[4], Ian Douglas[6], Joshua J. Gagne[1,2], Rosa Gini[7], Olaf Klungel[8], C. Daniel Mullins[9], Jeremy A. Rassen[11], Liam Smeeth[6], Miriam Sturkenboom[12], on behalf of the joint Force on Real World Evidence in Health Care Decision Making

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, MA, Harvard Medical School, MA, USA; [3]Pfizer, NY, USA; [4]Department of Population Medicine, Harvar [5]Department of Clinical Pharmacy, Maastricht UMC+, The Netherlands; [6]London School of Hygiene UK; [7]Agenzia regionale di sanità della Toscana, Florence, Italy; [8]Division of Pharmacoepidemiology & University, Utrecht, Netherlands; [9]Pharmaceutical Health Services Research Department, University MA, USA; [10]FDA Center for Drug Evaluation and Research, USA; [11]Aetion, Inc., NY, USA; [12]Erasm

Journal of the American Medical Informatics Association, 25(1), 2018, 17–24
doi: 10.1093/jamia/ocx109
Advance Access Publication Date: 23 October 2017
Research and Applications

Research and Applications

Exploring completeness in clinical data research networks with DQ^e-c

Hossein Estiri,[1,2,3] Kari A Stephens,[4,5] Jeffrey G Klann,[1,2,3] and Shawn N Murphy[1,2,3]

[1]Harvard Medical School, [2]Massachusetts General Hospital, [3]Partners HealthCare, Boston, MA, USA, [4]Department of Biomedical Informatics and Medical Education and [5]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

Staniford Street, Suite 750, Boston, MA

September 2017

uality assessment tool for evaluation
cord (EHR) data repositories.
itecture and gives an overview of its
ords with an encounter since January
artners HealthCare. All the code and
plementary Appendix.
leteness and conformance in a given
running the tool on the sample BPDR

Home › Software › ACHILLES for data characterization

ACHILLES for data characterization

Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) – descriptive statistics about an OMOP CDM v4 database.

We released our first open-source software application, ACHILLES, at the 2014 EDM Forum in San Diego, CA. Check out the interactive ACHILLES Demo.

ACHILLES is a platform which enables the characterization, quality assessment and visualization of observational health databases. ACHILLES provides users with an interactive, exploratory framework to assess patient demographics, the prevalence of conditions, drugs and procedures, and to evaluate the distribution of values for clinical observations.

ACHILLES is intended to be implemented by organizations that have patient-level observational health databases available in their local environment.

ACHILLES has two main components. The first component is implemented as an R package and runs securely within your local environment without disclosing any patient identifiable information. This R component requires that your database adheres to the OMOP common data model standard. The R package generates summary statistics which describe the quality and content of the patient-level observational health database and provides features to perform a simple review or bulk export of the summary statistics. The second component of ACHILLES is implemented as a HTML5 / JavaScript website with a series of interactive reports that allow for the exploration and visualization of the generated summary statistics. Summary statistics from multiple databases can be made available from a single installation of the ACHILLES web site.

Part 1: R package for summary statistics calculation (https://github.com/OHDSI/Achilles)

Part 2: Web interface to visualize ACHILLES summary statistics (https://github.com/OHDSI/AchillesWeb)

Project builds upon existing DQ activities: OHDSI example

- OHDSI* collaborative includes an active community of stakeholders who utilize the OMOP common data model (CDM)
- Several tools exist for the OHDSI community to use to evaluate data quality; tools are specific to the OMOP CDM
- DQM project created infrastructure to leverage the approaches from networks like OHDSI by enabling translation the OHDSI data quality metrics into a standardized format so the metrics can use used by others and compared across networks
  - The metrics should be data-model agnostic to enable cross network and cross data source comparisons

*OHDSI: Observational Health Data Sciences and Informatics | https://github.com/OHDSI | https://www.ohdsi.org/

Some Examples of Standard Data Checks

Example for two variables

ADate	Numeric (4)	SAS date	Encounter or admission date.
DDate	Numeric (4)	SAS date	Discharge date. Should be populated for all Inpatient Hospital Stay (IP) and Non-Acute Institutional Stay (IS) encounter types. May be populated for Emergency Department (ED) encounter types. Should be missing for ambulatory visit (AV or OA) encounter types.

Completeness:
- ADate variable has missing values

Validity:
- ADate variable is not SAS date value of numeric data type
- ADate variable is not of length 4
- DDate variable is not SAS date value of numeric data type
- DDate variable is not of length 4

Accuracy:
- ADate is after DDate (for IP and IS only)
- ADate and DDate variables have values before DP_MinDate

Integrity:
- DDate variable is missing for EncType value "IP"
- DDate variable is populated for records with EncType values other than "IP" or "IS"

Consistency:
- Problem with distribution of ADate (i.e. total number of records per year) within the ETL
- Problem with distribution of ADate (i.e. total number of records per year-month) within the ETL
- Significant change in number of records per ADate (year) across ETLs
- Significant change in number of records per ADate (year-month) across ETLs
- Problem with distribution of ADate (overall) within the ETL
- Problem with distribution of ADate (overall) across ETLs
- Problem with distribution of DDate (i.e. total number of records per year) within the ETL
- Problem with distribution of DDate (i.e. total number of records per year-month) within the ETL
- Significant change in number of records per DDate (year) across ETLs
- Significant change in number of records per DDate (year-month) across ETLs
- Problem with distribution of DDate (overall) within the ETL
- Problem with distribution of DDate (overall) across ETLs
- Problem with distribution of DDate variable by EncType per year
- Problem with distribution of DDate variable by EncType per year-month
- Problem with distribution of length of stay (DDate-ADate + 1) by EncType
- Problem with distribution of length of stay (DDate-ADate + 1) by EncType per year

Example for two variables

ADate	Numeric (4)	SAS date	Encounter or admission date.
DDate	Numeric (4)	SAS date	Discharge date. Should be populated for all Inpatient Hospital Stay (IP) and Non-Acute Institutional Stay (IS) encounter types. May be populated for Emergency Department (ED) encounter types. Should be missing for ambulatory visit (AV or OA) encounter types.

There are about 25 data checks for the admission and discharge date variables

Completeness:
- ADate variable has missing values

Validity:
- ADate variable is not SAS date value of numeric data type
- ADate variable is not of length 4
- DDate variable is not SAS date value of numeric data type
- DDate variable is not of length 4

Accuracy:
- ADate is after DDate (for IP and IS only)
- ADate and DDate variables have values before DP_MinDate

Integrity:
- DDate variable is missing for EncType value "IP"
- DDate variable is populated for records with EncType values other than "IP" or "IS"

- Admission date missing
- Discharge date is missing for encounter type of inpatient
- Problem with distribution of length of stay by encounter type by year

Consistency:
- Problem with distribution of ADate (i.e. total number of records per year) within the ETL
- Problem with distribution of ADate (i.e. total number of records per year-month) within the ETL
- Significant change in number of records per ADate (year) across ETLs
- Significant change in number of records per ADate (year-month) across ETLs
- Problem with distribution of ADate (overall) within the ETL
- Problem with distribution of ADate (overall) across ETLs
- Problem with distribution of DDate (i.e. total number of records per year) within the ETL
- Problem with distribution of DDate (i.e. total number of records per year-month) within the ETL
- Significant change in number of records per DDate (year) across ETLs
- Significant change in number of records per DDate (year-month) across ETLs
- Problem with distribution of DDate (overall) within the ETL
- Problem with distribution of DDate (overall) across ETLs
- Problem with distribution of DDate variable by EncType per year
- Problem with distribution of DDate variable by EncType per year-month
- Problem with distribution of length of stay (DDate-ADate + 1) by EncType
- Problem with distribution of length of stay (DDate-ADate + 1) by EncType per year

Sample metrics

- Distribution and missingness by variable by year
- Medical visits per month and per person per month
- Visits by visit type (inpatient, ambulatory, emergency department)
- Dispensings (prescriptions) per month and per person per month
- Distribution of days supplied and amount dispensed by year
- Proportion of encounters by disease category
- Out of range proportions (dates in the future or too far in the past)
- Ovarian cancer encounters by sex
- Rates of emergency department encounters that become inpatient hospital encounters

Example for Dispensing – Checking the Database

- Days supply and amount dispensed metrics
  - Missing
  - 0
  - < 0
  - 0 – 1
  - 1-30, 31-60, 61-90, 90-100, 100-999, 1000+
- Dispensings per year
- Dispensings per person per year
- Dispensings per person per year-month
- Dispensings per person per year by age group
- etc

Example for dispensing – Checking the study population

- Dispensings per person per year (period) for each medication of interest
- Metrics for days supply for medications of interest
- Metrics for amount dispensed for medications of interest
- Number of treatment episodes per person
- Length of treatment episodes (days)
- Days at risk by medication of interest
- Etc

Set of metrics for all variables of interest such as diagnoses, procedures, and key cohort and outcome phenotypes.

Platelet count units of measure across Sentinel

Platelet count original result units[‡]

Blank	FL	TH/UL	X10(3)
%	K/CMM	THOU/CMM	1000/UL
/100 W	k/cmm	thou/cmm	X10(3)/MCL
/CMM	K/CU MM	thou/mm3	X10(3)/UL
CMM	K/CUMM	THOU/UL	X10(6)/MCL
10 3L	K/MCL	THOUS/CU.MM	X10*9/L
10X3UL	K/mcL	THOUS/MCL	X10E3/UL
10^3/UL	K/UL	THOU/mcL	X1000
10*3/uL	k/uL	THOUS/UL	X10X3
10?3/uL	KU/L	Thou/uL	X10^3/UL
10E3/uL	K/MM3	THOUSA	x10
10e3/uL	K/mm3	THOUSAND	X10?3/ul
10e9/L	LB	THOUSAND/UL	X10E3/UL
E9/L	PLATELET CO	U	X10E3
BIL/L	T/CMM	X 10-3/UL	K/A?L
bil/L	TH/MM3	X 10(3)/UL	K/B5L
CU MM	th/mm3	X10 3	

Raebel MA, Haynes K, Woodworth TS, Saylor G, Cavagnaro E, Coughlin KO, Curtis LH, Weiner MG, Archdeacon P, and Brown JS**.** Electronic Clinical Laboratory Test Results Data Tables: Lessons from Mini-Sentinel. Pharmacoepidemiol Drug Saf. 2014 Feb;23(6):609-18.

NEGATIVE
POSITIVE
UNDETERMINED
BORDERLINE
BORDERLI
252.3
278
28
3178.2
5 Int
DETECTED
INDETERM
N
NOT DETE
Neg
Negative
Negatvie
P
Positive
SPRCS
TNP
N
Neg
Negative

.
820
840
1615
ABNORMAL
BOARDERL
BODERLIN
CANCELLE
DUPLICAT
EQIVOCAL
EQUIVOCA
NE-CHECK
NEAGTIVE
NEG (-)
NEGA
NEGA T I
NEGA TIV
NEGAT IV
NEGATAIV
NEGATIAV
NEGATIBE
NEGATIE
NEGATRIV

NEGATTVE
NEGATVIE
NEGAVTIV
NEGITIVE
NEGTIVE
NETGATIV
NORM
NORMAL
POA
POPSITIV
POSIITIV
POSITIFV
POSITTVE
POSITVE
POSOTIVE
POSTIVE
PSOITIVE
REPEAT
STAT
URINE

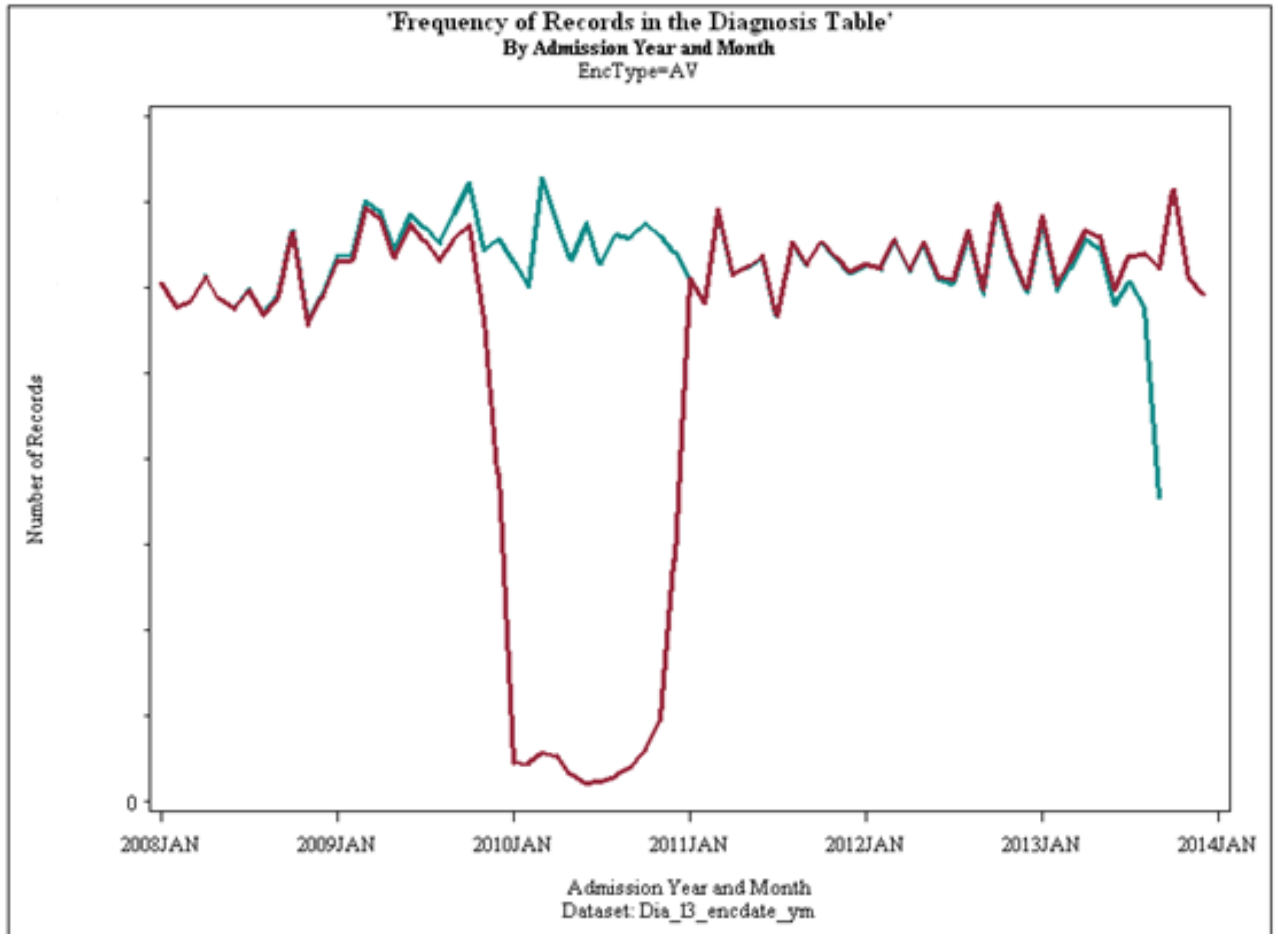
Examples of variations in qualitative pregnancy result units in source data across Sentinel

(I removed some rows…)

Why check after every refresh?

- Underlying data sources are dynamic
- Want to identify changes in data source transformation processes or data availability

Why check after every refresh?



Green: records from prior refresh

Red: record from new refresh under review

Problem:

Loss of 2010 observed in the **Diagnosis** table. Was due to an issue with loss of information in enrollment file.

Outcome:

The Partner was asked to recreate the refresh including 2010 data.

Networks and data sources have their own definitions and value sets for the same domains

Race	Char (1)	0 = Unknown 1 = American Indian or Alaska Native 2 = Asian 3 = Black or African American 4 = Native Hawaiian or Other Pacific Islander 5 = White

DEMOGRAPHIC Table Specification

Field Name	*RDBMS Data Type*	*SAS Data Type*	*Predefined Value Sets and Descriptive Text for Categorical Fields*
RACE	RDBMS Text(2)	SAS Char(2)	01=American Indian or Alaska Native 02=Asian 03=Black or African American 04=Native Hawaiian or Other Pacific Islander 05=White 06=Multiple race 07=Refuse to answer NI=No information UN=Unknown OT=Other

Sentinel

pcornet®

DQM provides a platform for data sources to describe data characteristics using common terms despite how their data are defined locally

This approach does not disrupt existing network-specific processes; researchers determine if the data fields in different networks have the same sematic meaning (e.g., sex vs gender)

OMOP	
concept	**description**
38003600	African
38003599	African American
38003573	Alaska Native
38003572	American Indian
8657	American Indian or Alaska Native
38003616	Arab
8515	Asian
38003574	Asian Indian
38003601	Bahamian
38003575	Bangladeshi
38003602	Barbadian
38003576	Bhutanese
38003598	Black
8516	Black or African American
38003577	Burmese

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

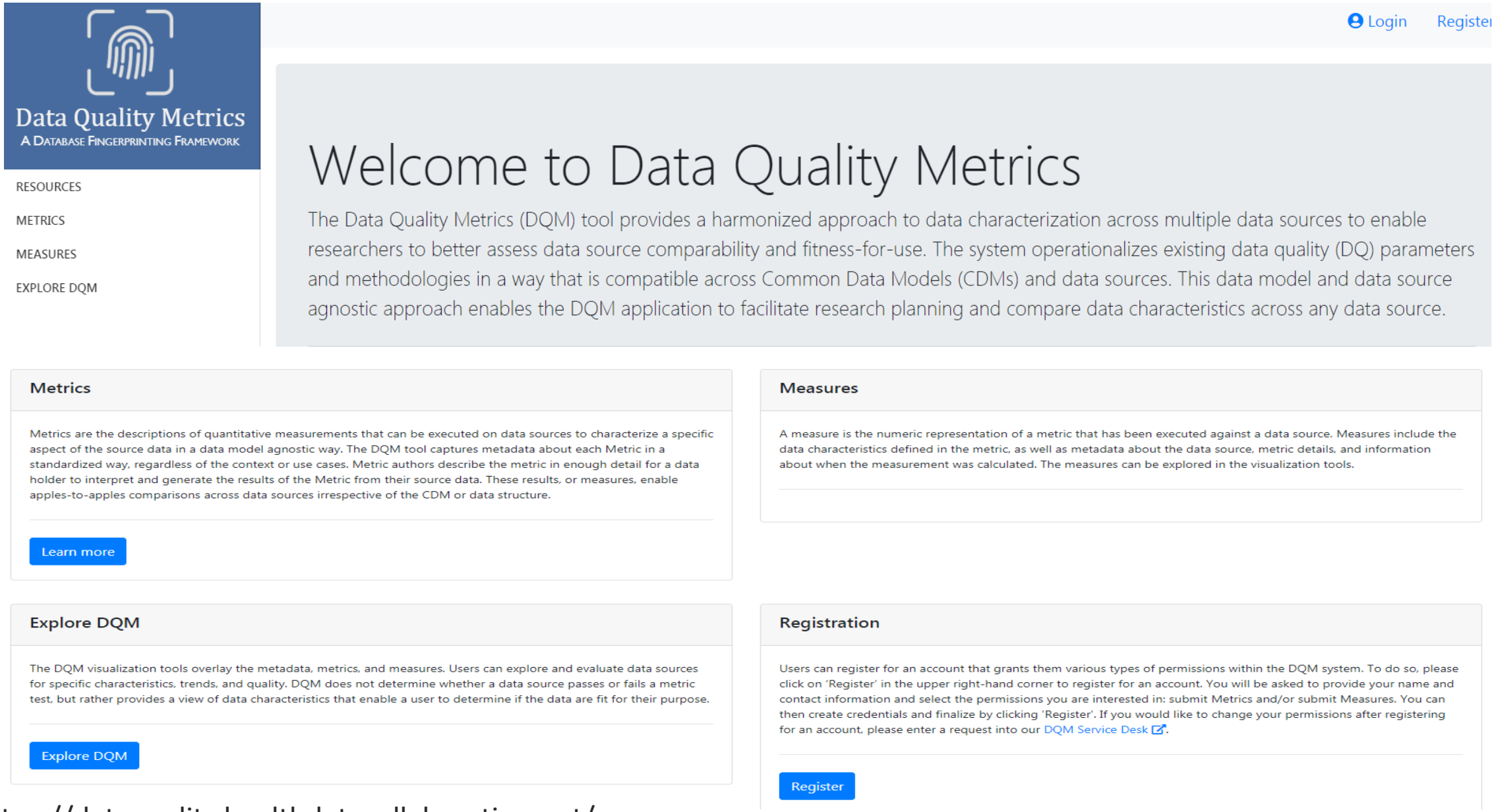
DQM project definitions and terminology: METRICS

- **Metrics describe quantitative measurements that characterize a specific aspect of the source data in a data model agnostic way**
  - Eg, outpatient pharmacy dispensings per health plan member per year
- DQM tool captures metadata about each metric
- Metric authors describe the metric in enough detail for a data holder to generate the data for the metric from source data source
- **Enable apples-to-apples comparisons across data sources regardless of the CDM or data structure**
- **As importantly, helps avoid inadvertent apples-to-orange comparisons**

DQM project definitions and terminology: MEASURES

- A **measure is the numeric representation of a metric that has been executed against a data source**
- Measures have associated metadata that includes:
  - **Target metric**: outpatient pharmacy dispensings per health plan member per year
  - **Data source (model)**: Harvard Pilgrim health plan claims database (Sentinel)
  - **Calculation details**: count of filled outpatient dispensings in year / number of health plan members with any medical or drug coverage enrollment in the year
  - **Timing of measure creation**: August 2019
- Measures can be explored using visualization tools

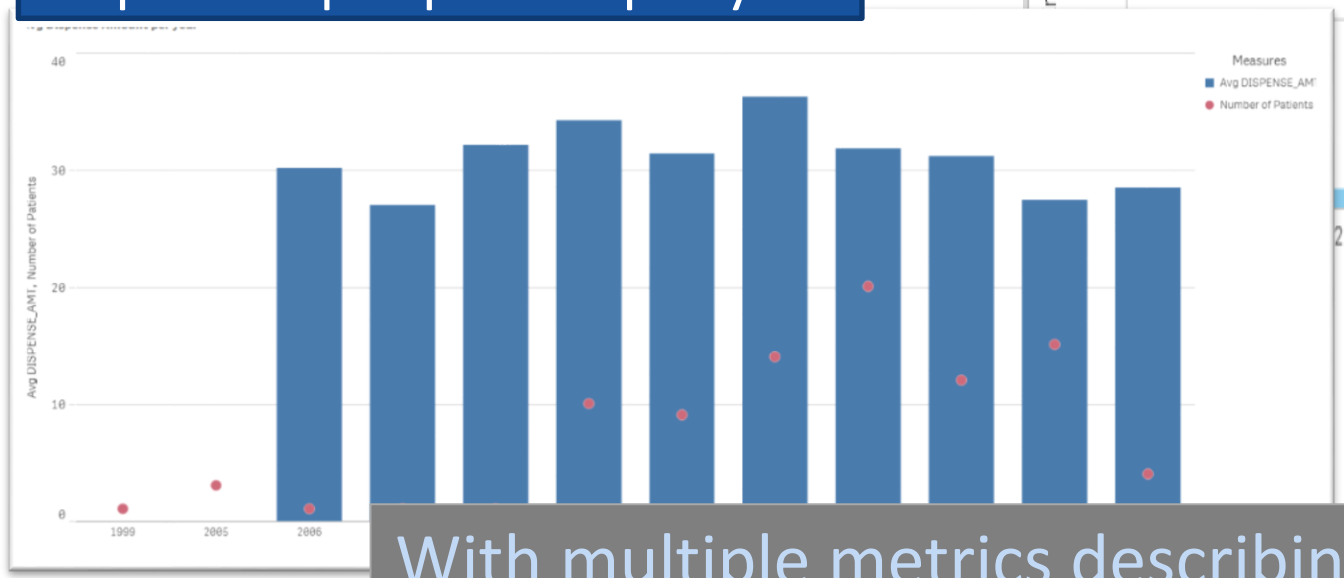
DQM project: Proof of concept



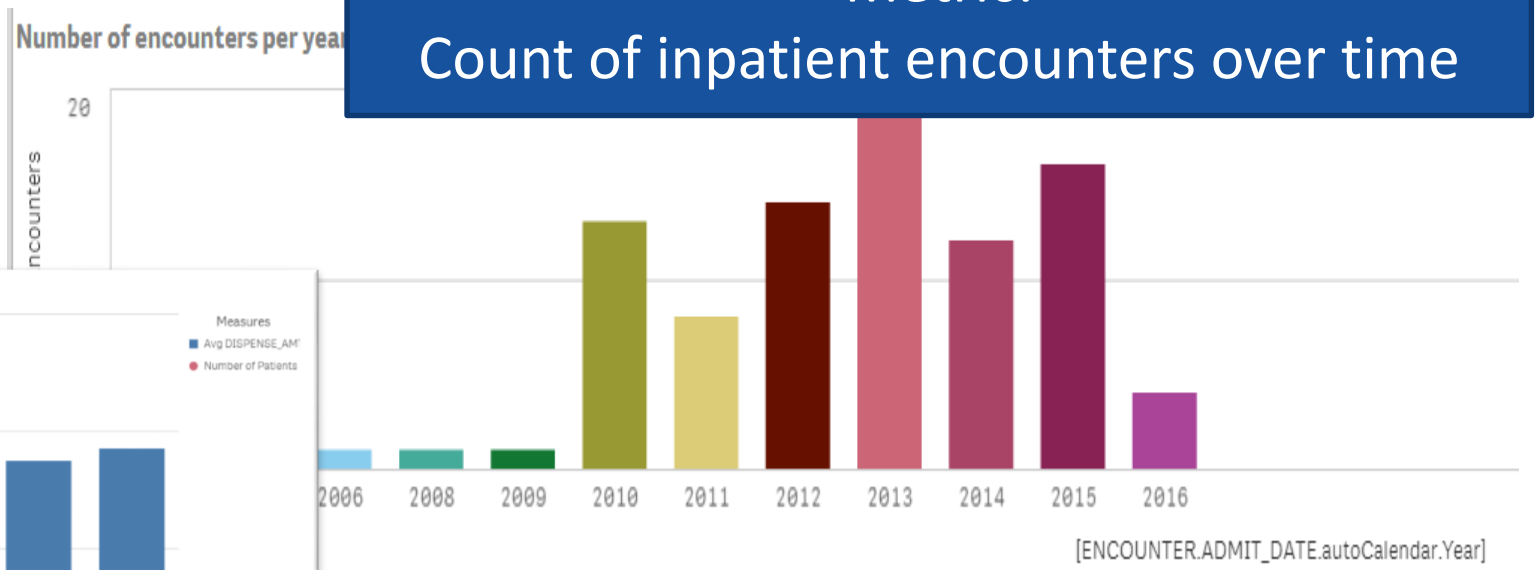
https://dataquality.healthdatacollaboration.net/

DQM helps researchers find the right data sources for specific studies

Metric:
Number of medications dispensed per patient per year



Metric:
Count of inpatient encounters over time



With multiple metrics describing data characteristics of interest, a user can explore data with interactive analytics tools to drill down into data sources that may be fit for their purpose

Summary

- Characterization of data sources and cohorts is critical, especially in multi-site research
- Be careful if using new data extracts or frequently updated data
- Each RUF Accelerator project should share as much data characterization information as possible
  - Beyond Table 1 comparisons
- Each project should develop a core set of data characterization metrics as part of the study synopses along with other study parameters
- Metrics should focus on the study cohort and key variables, can be extended to the source population