October 2015

DQ Code-A-Thon 2015 Orientation Packet

Data Quality Collaborative

PCORI Methods: Data Quality Collaborative (DQC) Code-A-Thon Orientation Packet
Updated: 23 October 2015

I. Meeting Logistics

- **Key contacts**: Michael Kahn's cell phone: (303) 324-2829
  Juli Barnard's email contact: Juliana.Barnard@ucdenver.edu
- **Meeting location**: Galvanize – Golden Triangle, 1062 Delaware Street, Denver CO 80204 Ph:(303) 823-4170
  http://www.galvanize.com/campuses/Denver-golden-triangle/#.Vfm5hhFVhBc
- **Date/time**: Friday, November 6, 2015 @ 5PM MT thru Sunday, November 8, 2015 @ 1PM MT
- **Accommodations**: (HOTEL, FLIGHTS, MEALS and SHUTTLES): All accommodations and travel will be reimbursed and arranged by Academy Health. Contact David Padgham (David.Padgham@AcademyHealth.org ; 202.292.6777) if you have not received email with travel details.
    - Meals will be provided. Let us know if you have dietary restrictions. There is a kitchen with refrigerators and microwaves at the meeting location that you can use if you wish.
    - We are planning a fun event for your entertainment during Saturday dinner (we promise nothing that will embarrass the extreme introverts amongst us)!

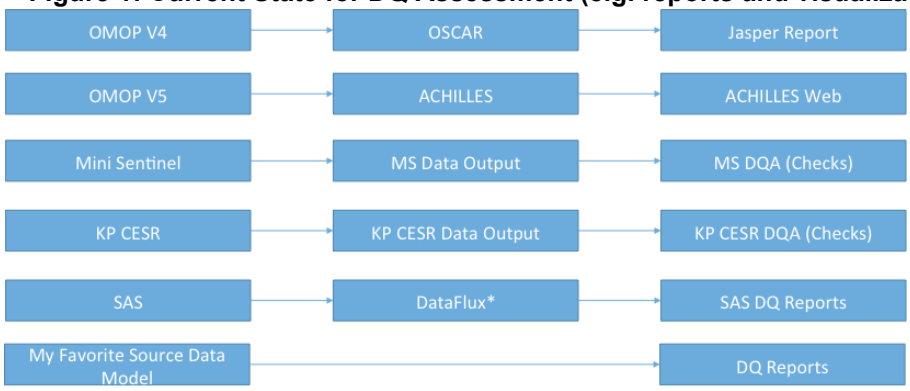
II. Data Quality Assessment: Background

- **Key Background Paper:** Participants are encouraged to read the manuscript entitled "A Harmonized Data Quality Assessment Terminology for the Secondary Use of Electronic Health Record Data" available for download at:
  https://drive.google.com/file/d/0By1tgphRY1wpMTUtcVpkcjUwT2s/view?usp=sharing
  Because this manuscript is under active review for publication, please do not distribute this manuscript to others who are not directly participating in the DQC Code-A-Thon. This paper describes the results of an arduous effort to harmonize the DQ terminology used to describe various aspects ("dimensions") of data quality. While not required, we hope that this organizational structure for data quality dimensions will be useful to guide your work.

- **Current State**: Data quality assessment (DQA) is often performed as an ad-hoc, one-off, non-systematic, not-documented, not-reported activity when it is performed at all. Only a few projects have invested significant resources to develop more formal DQA procedures and programs. Most of these programs have evolved over many years, reflecting substantial investments in developing DQA measures, reports, and illustrations.[1–4] Unfortunately, Figure 1 (see below) highlights the current situation with producing data quality (DQ) reports and visualizations. As illustrated, DQ programs are written by project-specific programmers against specific source data models. Typically, these programs generate DQ measures either directly from the source data model (Figure 1, last row), or write DQ summary statistics into project-specific summary tables (Figure 1, middle column). In either case, data quality routines are created that are project-specific. That is, innovative DQ routines or visualizations written by Project A are not available to communities using Project B. As the number of data-sharing networks and technologies grows, each project is faced with recreating DQ assessments done previously by other projects and unique data quality checks or visualizations cannot be shared across projects. Nor is it possible to assume that the terminology used in one

Project to describe a set of data quality checks is the same as the terminology used in a different Project (hence the creation of the Data Quality Harmonization paper listed above).

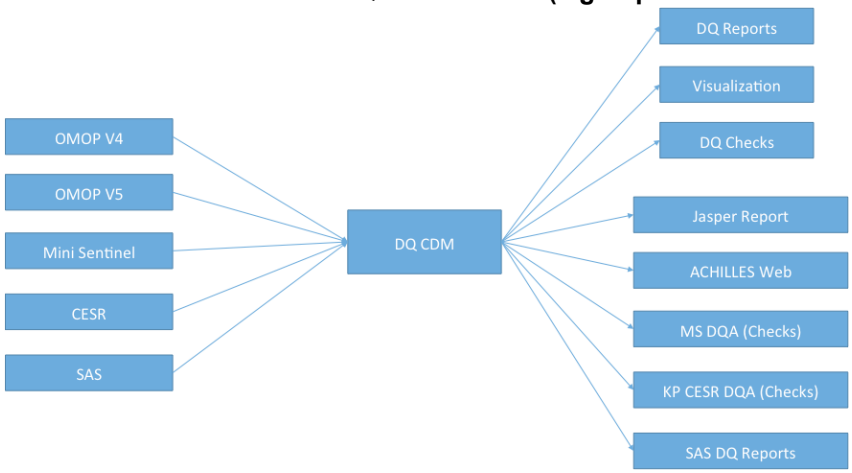
Figure 1: Current State for DQ Assessment (e.g. reports and visualizations)



- Future State (Long Term): Figure 2 highlights the desired (long-term) futures state. DQ assessment measures/programs developed by all projects store their intermediate DQ results into a universal common data model, called the Data Quality Common Data Model (DQ CDM). Existing and future novel data quality routines or visualizations are written to use the DQ CDM as a data input source for data quality measures. Under this future state, as long as a project writes their data quality measures into the DQ CDM, the project will have the ability to use data quality routines/visualizations written by any other project that also used the DQ CDM as the data source for their DQ tool. Note that the DQ CDM contains summary level data that is stored in a format that is independent of the data format of the source data model. The format of the DQ CDM is described in Section VI.

  By writing to and reading from a common data model for quality measures, innovations in new data quality checks, measures, graphics, reports, and interactive visualizations can be shared. In addition, as more tools are written to use the DQ CDM, a broader range of data quality assessments will be available to projects without having to rewrite code, leaving projects with more time/resources to investigate findings rather than in coding programs that have previously been written by others.

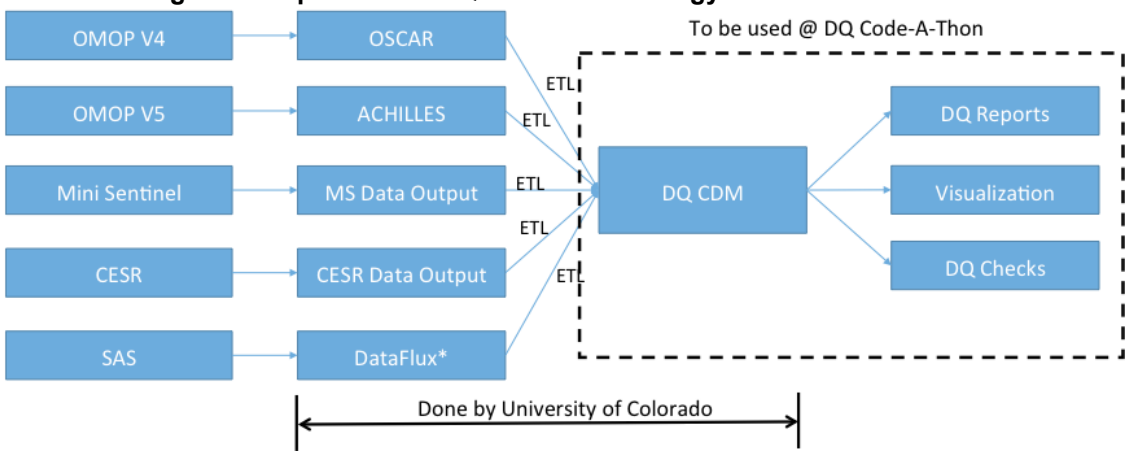
Figure 2: Desired Future State for DQ Assessment (e.g. reports and visualizations)



- **Intermediate State:** In order to prove the value of using the DQ CDM to enable cross-project reuse of data quality tools, reports and visualizations, Figure 3 illustrates the current method that the University of Colorado team has created to support the Code-A-Thon activities. Rather than

request/require existing DQ programmers to change their existing code to write their existing DQ summary measures into the DQ CDM structure, the University of Colorado has obtained copies of data files that contain DQ summary statistics from various projects and have created scripts that manually transfer DQ measures from the project-specific intermediate files (the "middle" boxes in Figure 1), into the DQ CDM. This interim approach allows current projects to keep their existing DQ summary file structures intact while also enabling members of the DQ Code-A-Thon to work with creating compelling tools that use DQ measures stored in the DQ CDM format.

Figure 3: Implemented DQ CDM methodology for the Code-a-Thon



III. DQC Code-A-Thon Rules, Resources and Requirements

- **Code-A-Thon Aim**: The goal of this weekend will be to have small teams of programmers, visualization experts and DQ project leads compete to create visualizations of DQ results produced by the project's Data Quality Common Data Model (DQ CDM). Be creative, take risks, be as wacko as you want. We want to come out of this meeting with ideas and prototypes that go far beyond the current tools and graphs. Reach high to go after "new data characterizations/visualizations to summarize data, as well as better mechanisms to explore the summary data so that unusual patterns can be more easily detected, diagnosed, and resolved" (Pat Ryan).

  Equally important, please have fun working with known and new colleagues who are equally passionate about data quality and data quality assessment. We are hoping that enduring collaborations will emerge from this long weekend together. The DQC Code-a-Thon will conclude with presentations of each team's DQ visualizations and code. *Prizes will be awarded on Sunday* to the team with the best and most creative visualizations.

- **How we will run the Code-A-Thon** (basic outline, we'll be revise as the event unfolds)
  - We will meet at Galvanize on Friday @ 5PM
  - Each participant's expertise will be denoted by a color dot on your name tag. The expertise "types" are: programmer, visualization expert, data quality assessment expert. Each team that is created for the Code-a-Thon must have a minimum of 1 DQ assessment expert, 1 visualization expert, and 1 programmer. For those of you who self-identified to more than one type, all of your expertise types will be noted on your name tag and you will be free to select one of your expertise types to fulfill for your team. It is out-of-bounds to swap color dots with other participants!

- There will be no prior assignment of folks to working groups.
- We will begin the evening as an amorphous group. After the initial pro-forma opening comments about how glad we are you have come and how much hope we are putting into your hands to make this a spectacular event, and some basic logistics, we will allow folks to self-aggregate into four or five working groups. The only "requirement" that we will place on groups is that there is at least one representative from each colored dot in each group. We strongly encourage diversity in team members, including folks who have not worked together before. Mix it up!!!
- The same people will work in the same group for the entire Code-A-Thon.
    - Some unassigned folks (Michael Kahn, Toan Ong, Juli Barnard) will be circulating to help groups that might get stuck.
- I will repeat much of the context that is presented here and answer any overall context questions. Toan Ong will be available to answer any questions about the DQ CDM.
- We will go to ~ 9PM Friday evening (recognizing that this is 11P for the East Coasters) and will be shuttled back to the hotel. Groups are free to continue working in the hotel if they wish.
- The shuttle will pick folks up at the hotel at 8:30AM on Saturday
- We will continue working at Galvanize until 9PM, shuttle back to the hotel. Again, the stalwarts in the group can continue to work in the hotel
- The shuttle will pick folks up at the hotel at 8:30AM on Sunday
- We continue until ~ 12N, do presentations, awards, and any wrap-up and then shuttle folks to the airport at 1PM.
- We do intend to have end-of-session prizes although we are still working on this aspect.
- We will provide Friday dinner, Saturday breakfast, lunch and dinner, and Sunday breakfast and lunch. The requisite Mountain Dew and Pizza will be our "Cardiology-income-enhancing" event for Saturday lunch.
- On Saturday, we'll have a 1 hour entertainment event at 7PM. No specific attire is required. Please hold back any displeasure with our selection – it's costing us a more than we thought…….. And yes, it is tasteful (some of our options were more marginal than others…….).

- **Resources:**
    - The University of Colorado has developed instances of the DQ CDM from at least five source data models:

        - ACHILLES → DQ CDM from a pediatric OMOP V5 CDM
        - ACHILLES → DQ CDM from a CMS SYNPUF OMOP V5 CDM developed by Lee Evans
        - Fake MiniSentinel Summary SAS files → DQ CDM from HarvardPilgrim
        - Child Health Corporation Pediatric Health Information Systems (PHIS) data quality measures for Children's Hospital Colorado
        - Fake PCORnet Summary SAS files → DQ CDM from HarvardPilgrim
        - There may be other DQ CDM instances available by the time of the Code-A-Thon

    Prior to the Code-A-Thon, CSV and PostgreSQL DBMS versions of these DQ CDM instances will be uploaded to the EDM Forum wiki page (URL to be available soon).

    In addition to the DQ CDM instances that the University of Colorado will make available for use, some Code-A-Thon participants have requested the ability to use their own source data (left box in Figure 3) to write their own set of data quality summary statistics (left middle box in Figure 3). These "personal" source data sets may be known to have interesting DQ findings that are not represented in the UCD-provided data sets or they may have data domains that support novel DQ summary measures that are not present in the UCD DQ CDM instances.

Participants are free to use their own source data to create their own summary statistics **but it is an absolute requirement of the grant that's paying for all of this(!) that the summary statistics generated from "personal" data sets must be formatted into the DQ CDM structure (right middle box in Figure 3) before proceeding to writing DQ reports, graphs or visualizations based on their data.**

- Check out the Galvanize web site to see all of the cool stuff that is available there, including:
  - Cash bar with multiple local brews on tap (a requirement for Pat Ryan's participation)
  - Kitchen with refrigerators and microwaves, coffee or tea (didn't see Expresso but Metropolis Coffee is around the corner)
  - Multiple rooms for groups to work (each space has its own video display; Galvanize will have HDMI-VGA and Mac HDMI dongles but bring your own HDMI dongle just-in-case)
  - Lots of alternative spaces for spreading out if you wish, couches for snoozing, and smaller rooms for actually doing work.
  - Two ping pong tables
  - We will have our own private WiFi network that is dedicated to our use only.
  - 3M large post-it notes easel boards and large tables with dry erase surfaces for doodling and drafting ideas. We will have markers for all on site.
  - A printer will be available but it will be a bit awkward – we will need to email the file to be printed to the Galvanize employee who will be watching over us to make sure we don't get too rowdy.

- **Requirements**
  - Teams are free to use any programming environment, tools, or software, including commercial software such as SAS, Tableau or Spotfire.
  - Individuals bring their own programming environments
  - Teams agree to use the DQ CDM as the sole data input source for all programming. (see comment about "personal" data sets in Section III).
  - This is a diverse group and there will be differences of opinion. That's totally fine! And if you need guidance from one of our code-a-thon judges (Juli and me!), just ask.
  - ***Teams agree to have their post-Code-A-Thon source code and output posted on a public-domain web site*** and to allow BSD-like open source access to their source code by the general public. Programming environments that do not readily support this type of distribution will be discussed on a case-specific basis.

IV. Sample Use Cases

Participants are free to develop novel data quality checks, reports or visualizations that are of most interest to them. There are ***no boundaries to what folks may pursue***. However, some participants have asked to be provided with some sample use cases in order to help frame their thinking about the general scope of the desired work.

There are different "customers" that we envision would benefit from the data quality assessment products that we will develop at the DQC Code-A-Thon:

- Data analysts who are responsible for understanding the strengths and weaknesses of data sets that they receive from data owners. These individuals tend to be "in the data" as part of their daily responsibilities. They are comfortable with data exploration tools and are used to looking at data quality results for anomalies that might indicate data quality issues
- Data users, typically (in our setting) clinical investigators, who are more focused on using data to answer a specific research question. While these individuals are involved in using data for their research, they tend to not be as well versed in the nuances of complex data sets nor do they tend to spend time digging into the nitty gritty features of data sets

- Data consumers, which for our work, are defined as patients, patient advocates, and health care policy makers. These individual use the results of data analyses to understand who these results affect their care (patients, patient advocates) or may alter health policy (patient advocates, health care policy makers). Most in this group are used to working with highly "processed" data in aggregate form, usually as summarized by statistical models, plots and other high level summary. Their interest in data quality is more indirect – what do I need to know about the underlying data quality that might affect how I interpret these findings for myself or for setting new policies. A brief perspective on data consumer needs is available from Erin MacKay of the National Partnership for Women and Families, a national consumer group.

Work products developed for these customers are likely to be very different as these groups have markedly different experience with and exposure to detailed large-scale clinical data. Telling the data quality "story" ("What do these findings mean to me?") is the core challenge to be addressed by this DQC Code-A-Thon. You are free to pick whichever customer type (or more than one customer type if you are very brave), to frame your work. **However, we strongly encourage working towards creating product for the data consumers group since (1) It is a requirement in our grant which is funding the Code-a-Thon, (2) it probably is a more challenging community to engage in data quality findings, and (3) It is likely to get you extra points to win a prize on Sunday!**

The use cases below are combined/modified from more detailed use cases provided by Pat Ryan (OHDSI) and Meredith Nahm (Duke). Their original use cases will be posted on the EDM Forum wiki. These use cases are provided only as examples and are not intended to be directive or limiting – folks should feel free to wander down their own paths if they so choose.

- Can we find anomalies in prevalence trends over time (e.g. by month/by year) with adequate sensitivity/specificity?
- Can we find Are there improbable/implausible values for categorical distributions to flag?
- Can we find implausible values (crazy lab measures, diagnoses in the wrong gender, physically impossible such as negative weights, etc.) Can we drill down to which site, which data element, which encounter?
- Implausible health patterns, such as outpatient visits subsumed within inpatient visits, visits/procedures recorded after death date or before birth date
- How to describe missingness – amount and patterns
- Can we view results of comparisons between two independent datasets to identify areas of significant inconsistency? For example comparison of Diagnoses from CMS claims data and Health Records, or between problem lists and Diagnoses, or Diagnoses for same patients between two different facilities (or in our setting, difference across two different DQ CDM instances)
- Can we use temporal disturbances (temporal gaps) to signal potential exit by population members. Ability to compare by facility

The Data Quality Harmonization paper mentioned in Section II also contains examples of other data quality assessments. Additional examples can be found in the work by Kahn [5] and by Brown[1]. Both papers are open access and are available for download on Google.
Kahn: https://drive.google.com/file/d/0By1tgphRY1wpZmtUQkRudXNFQTg/view?usp=sharing
Brown: https://drive.google.com/file/d/0By1tgphRY1wpOGE3cDdCY1ZuYVk/view?usp=sharing

V. References:

1 Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care* 2013;**51**:S22–9.

2 Botsis T, Hartvigsen G, Chen F, *et al.* Secondary use of EHR: Data quality issues and informatics opportunities. *AMIA Summits Transl Sci Proc* 2010;**2010**:1–5.

3 Bonney W, Scobbie D, Nind T, *et al.* Profiling Clinical Datasets for Data Quality Assessment and Improvement. http://www.bcs.org/upload/pdf/ewic_his14_full_paper1.pdf (accessed 8 Jun2015).

4 Mini-Sentinel Coordinating Center. Mini-Sentinel Data Quality Review and Characterization Procedures and Finding Report Version 1.0. Published Online First: 20.http://www.mini-sentinel.org/work_products/Data_Activities/Mini-Sentinel_Year-1-Data-Quality-and-Characterization-Procedures-and-Findings-Report.pdf

5 Kahn MG, Raebel MA, Glanz JM, *et al.* A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care* 2012;**50 Suppl**:S21–9. doi:10.1097/MLR.0b013e318257dd67

VI. DQ CDM Technical Description/Specification

The common data model for data summarization is a star schema. The Result table is the fact table, which contains the results of the summary statistics. There are two dimension tables: Measures and Dimension_Set. The Dimension Set table is a denormalized table that contains information for up to 6 dimension by which the data in the Result table were aggregated. The Measure table contains meta-data descriptions of a measure.

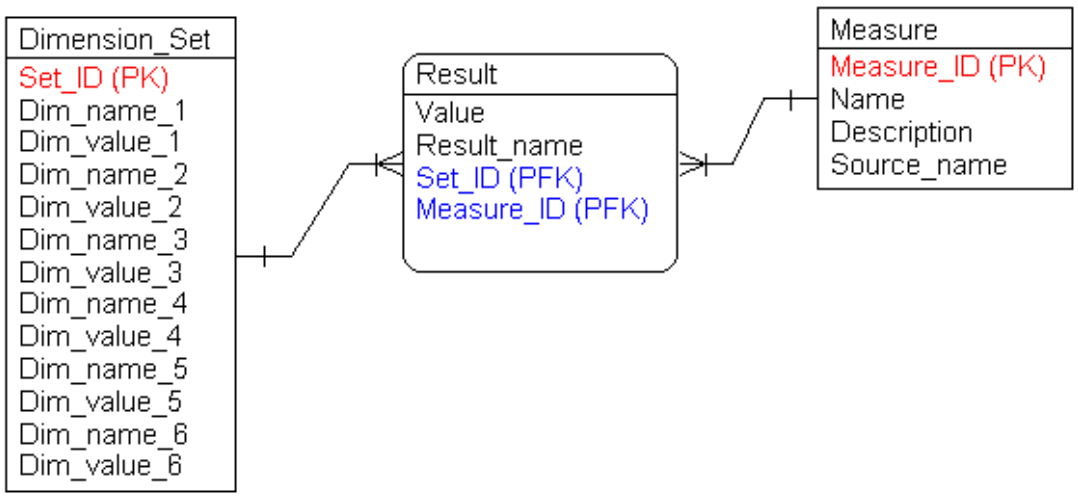


Table Specifications

Measure

Column name	Data type	Required	Description
Measure_ID	Integer	YES	A unique identifier of each measure.
Name	Text (50)	YES	A short name of the measure. Examples: DIA1.3.3, Diagnosis_null_date
Description	Text (200)	NO	Text description of the measure. Example: Count of diagnosis dates which are NULL.
Source_name	Text (50)	YES	Name of the source data. For examples: KAISER-NW, CHCO

- Examples

Measure_ID	Name	Description	Source_name
1	ENR1.1.1	Count of patient ids that contain special characters in the Enrollment table	VDW
2	ENC_MISSING_AS	Count of encounters with missing admitting source information	PCORnet
3	IP_VISIT_AVG_DIA	Average number of diagnoses per inpatient visit	CHCO-OMOP

Dimension_Set

Column name	Data type	Required	Description
Set_ID	Number	YES	A unique identifier of each dimension set.
Dim_name_1	Text (50)	YES	A short name of dimension 1
Dim_value_1	Text (50)	YES	Value of dimension 1
Dim_name_2	Text (50)	NO	A short name of dimension 2
Dim_value_2	Text (50)	NO	Value of dimension 2
Dim_name_3	Text (50)	NO	A short name of dimension 3
Dim_value_3	Text (50)	NO	Value of dimension 3
Dim_name_4	Text (50)	NO	A short name of dimension 4
Dim_value_4	Text (50)	NO	Value of dimension 4
Dim_name_5	Text (50)	NO	A short name of dimension 5
Dim_value_5	Text (50)	NO	Value of dimension 5
Dim_name_6	Text (50)	NO	A short name of dimension 6
Dim_value_6	Text (50)	NO	Value of dimension 6

- Notes:
  - Each dimension set must have at least one dimension
  - Each dimension must have one value
  - The combination of dimension name and dimension value is required to be unique in the Dimension_Set table. That is, two sets must not have the exact same combination of dimension name and dimension value.
  - For summary statistics based on total counts (e.g. count(*)), use the following convention:
    - Dim_name_1 = ALL, Dim_value_1 = *

- Examples (For simplification, only 3 dimension columns were included in the example)

Set_ID	Dim_name_1	Dim_value_1	Dim_name_2	Dim_value_2	Dim_name_3	Dim_value_3
1	Year	2013				
2	Year	2014				
3	Year	2015				

4	Quarter	1				
5	Quarter	2				
6	Year	2013	Quarter	1		
7	Year	2013	Quarter	2		
8	Year	2014	Quarter	1		
9	Year	2014	Quarter	2		
10	Year	2013	Month	1	Day	2
11	Year	2013	Month	1	Day	3
12	Year	2013	Month	1	Day	4
13	Year	2013	Month	2	Day	1
14	Year	2013	Month	2	Day	2
15	ALL	*				

Result

Column name	Data type	Required	Description
Set_ID	Integer	YES	A foreign key to the Dimension set which Value is aggregated by.
Measure_ID	Integer	YES	A foreign key to the Measure to which the results belong.
Value	Numeric	NO	The result of the summary statistics. Null indicates value couldn't be computed.
Result_name	Text	YES	Description of the value. For examples: Max, Min with Age>65

- Conventions:
  - Results with the same semantic should have the same name.
  - Result_name contains meta-data of the result. Data elements contained in the result_name are triple-pipe delimited and stored using the following structure.

Metadata 1 name=Metadata 1||| Metadata 2 name=Metadata 2||| Metadata 3 name=Metadata 3

- Examples of result_name:
  - name=COUNT|||unit=patient
  - name=AVERAGE

- Examples:
  - Count of patient ids that contain special characters in the Enrollment table in 2013
  - Count ALL patient ids that contain special characters in the Enrollment table
  - Average number of diagnoses per inpatient visit in Q1 2013

Set_ID	Measure_ID	Value	Result_name
1	1	5	name=COUNT
15	1	100	name=COUNT_ALL
6	3	3.5	name=AVERAGE

VII. Current datasets in DQ CDM format

Dataset 1: Children's Hospital Colorado de-identified dataset

Description: EHR data from Children's Hospital Colorado was de-identified and converted into the OMOP v4 CDM. From this dataset, summary statistics were generated using ACHILLES[1]. Data in ACHILLES result table were transformed into the DQ CDM.

Measure.Source_name = 'OMOP4DEID'

Table 1 - Possible result_names in the Result table (ACHILLES)

Values	Description
name=COUNT|||type=RESULT	Count from the ACHILLES Result table
name=COUNT|||type=DIST	Count from the ACHILLES Dist table
name=STDEV|||type=DIST	Standard deviation from the ACHILLES Dist table
name=MAX|||type=DIST	Max from the ACHILLES Dist table
name=AVG|||type=DIST	Average from the ACHILLES Dist table
name=MIN|||type=DIST	Min from the ACHILLES Dist table
name=MEDIAN|||type=DIST	Median from the ACHILLES Dist table
name=p10|||type=DIST	10 percentile from the ACHILLES Dist table
name=p25|||type=DIST	25 percentile from the ACHILLES Dist table
name=p75|||type=DIST	75 percentile from the ACHILLES Dist table
name=p90|||type=DIST	90 percentile from the ACHILLES Dist table

Sample query 1 (in PostgreSQL dialect)

This query recreates the ACHILLES_dist table from the DQ CDM tables:

```
WITH
r1 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=COUNT|||type=DIST'),
r2 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=MIN|||type=DIST'),
r3 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=MAX|||type=DIST'),
r4 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=AVG|||type=DIST'),
r5 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=STDEV|||type=DIST'),
r6 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=MEDIAN|||type=DIST'),
r7 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=p10|||type=DIST'),
r8 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=p25|||type=DIST'),
r9 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=p75|||type=DIST'),
r10 as (SELECT * FROM dqcdm.result WHERE result_name = 'name=p90|||type=DIST')
select r1.value as "COUNT", r2.value as "MAX", r3.value as "MIN", r4.value as "AVG", r5.value as
"STDEV", r6.value as "MEDIAN", r7.value as p10, r8.value as p25, r9.value as p75, r10.value as p90
FROM r1 JOIN r2 ON r1.measure_id=r2.measure_id AND r1.set_id = r2.set_id
        JOIN r3 ON r1.measure_id=r3.measure_id AND r1.set_id = r3.set_id
        JOIN r4 ON r1.measure_id=r4.measure_id AND r1.set_id = r4.set_id
        JOIN r5 ON r1.measure_id=r5.measure_id AND r1.set_id = r5.set_id
        JOIN r6 ON r1.measure_id=r6.measure_id AND r1.set_id = r6.set_id
        JOIN r7 ON r1.measure_id=r7.measure_id AND r1.set_id = r7.set_id
```

[1] http://www.ohdsi.org/analytic-tools/achilles-for-data-characterization/

```
JOIN r8 ON r1.measure_id=r8.measure_id AND r1.set_id = r8.set_id
JOIN r9 ON r1.measure_id=r9.measure_id AND r1.set_id = r9.set_id
JOIN r10 ON r1.measure_id=r10.measure_id AND r1.set_id = r10.set_id
```

Dataset 2: 1000 person SynPUF data

Description: 1000 person data set from the CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) on the CMS website[2] and converted it to the OMOP Common Data Model Version 5 format. OHDSI ACHILLES results of dataset 2 were generated and used to populate the DQ-CDM.

Measure.Source_name = 'SYNPUF'

Table 2 - Possible result_names in the Result table (ACHILLES)

Values	Description
name=COUNT|||type=RESULT	Count from the ACHILLES Result table
name=COUNT|||type=DIST	Count from the ACHILLES Dist table
name=STDEV|||type=DIST	Standard deviation from the ACHILLES Dist table
name=MAX|||type=DIST	Max from the ACHILLES Dist table
name=AVG|||type=DIST	Average from the ACHILLES Dist table
name=MIN|||type=DIST	Min from the ACHILLES Dist table
name=MEDIAN|||type=DIST	Median from the ACHILLES Dist table
name=p10|||type=DIST	10 percentile from the ACHILLES Dist table
name=p25|||type=DIST	25 percentile from the ACHILLES Dist table
name=p75|||type=DIST	75 percentile from the ACHILLES Dist table
name=p90|||type=DIST	90 percentile from the ACHILLES Dist table

Example query 2 (in PostgreSQL dialect): Count number of visit occurrence records by visit occurrence start month in 2008

```
SELECT d.dim_name_1, d.dim_value_1, r.value,
FROM dimension_set d join result r on d.set_id = r.set_id join measure m on r.measure_id =
m.measure_id
WHERE left(dim_name_1,4) = '2008' and dim_name_2 is null AND result_name =
'name=COUNT|||type=RESULT' AND m.measure_id = 220;
```

Dataset 3: Mini sentinel synthetic datasets

Description: Two synthetic datasets (100K and 500K) were used to generate the summary statistics to perform 224 MS data quality checks. The original data in SAS were loaded into PostgreSQL tables, which were transformed into the DQ CDM.

Measure.Source_name = '100K' for the 100K dataset

Measure.Source_name = '500K' for the 500K dataset

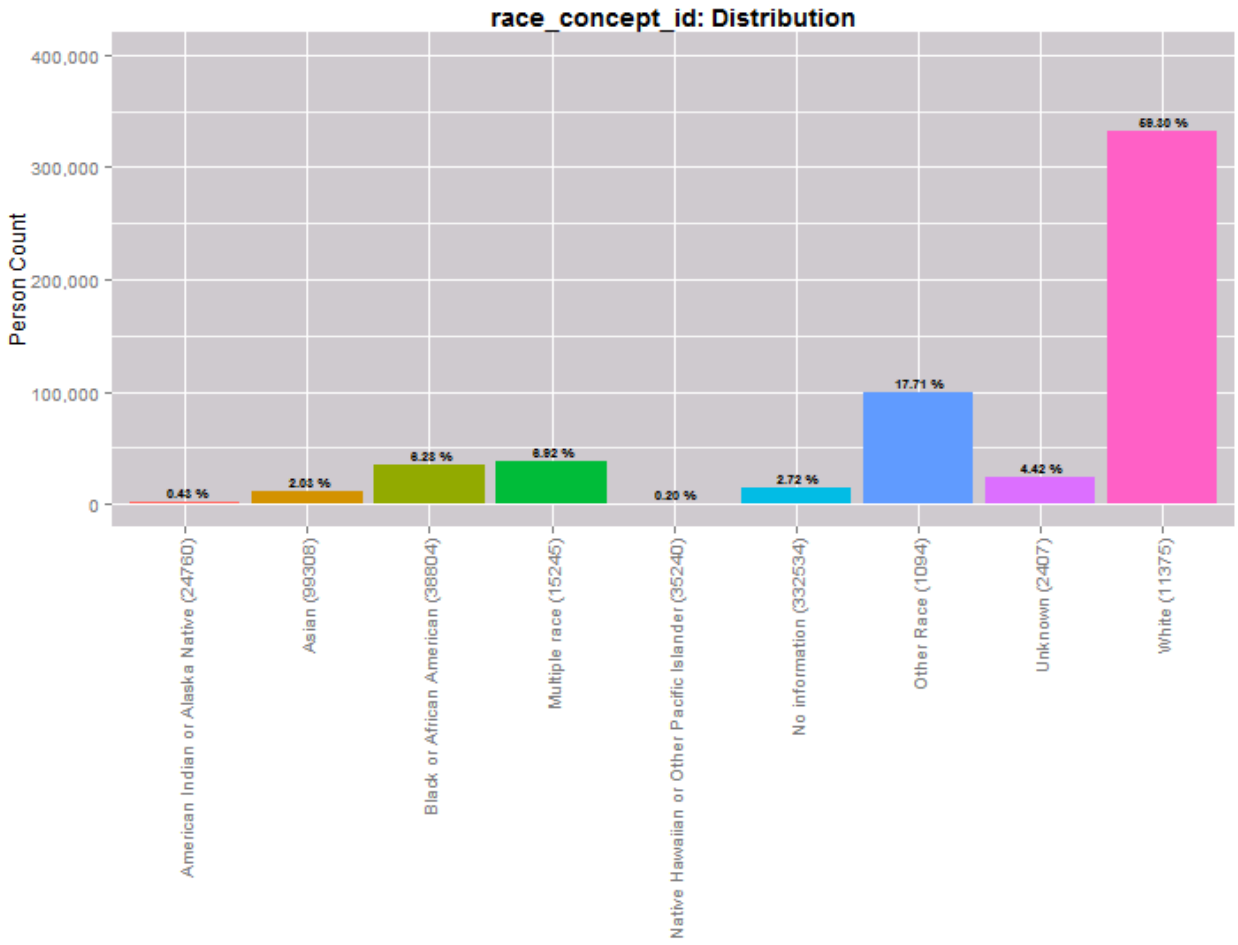
[2] https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html

Table 2 - Possible result_names in the Result table (Mini Sentinel)

Values	Description
name=COUNT	COUNT
name=COUNT DISTINCT	COUNT DISTINCT
name=MAX	MAX
name=MEDIAN	MEDIAN
name=MIN	MIN
name=p1	1 percentile
name=p5	5 percentile
name=p25	25 percentile
name=p75	75 percentile
name=p95	95 percentile
name=p99	99 percentile

Example query 3 (in PostgreSQL dialect): Count number of patients by race

```
SELECT d.dim_name_1, d.dim_value_1,c.concept_name, r.value, r.result_name
FROM dimension_set d join result r on d.set_id = r.set_id join omop4deid.concept c on d.dim_value_1 =
c.concept_id::text
WHERE dim_name_1 = 'race_concept_id' and dim_name_2 is null AND result_name =
‘name=COUNT|||type=RESULT’;
```



Bar plot generated from sample query 3 data

Dataset 4: Pediatric Health Information System (PHIS):

Description: Clinical data from Children's Hospital Colorado were stored in PHIS data quality format (Excel spreadsheet) for 1856 DQ measures. Data associated with these measures include the summary statistics from CHCO and PHIS median. PHIS data in Excel format were transformed into the DQ CDM in PostgreSQL.

Measure.Source_name = PHIS

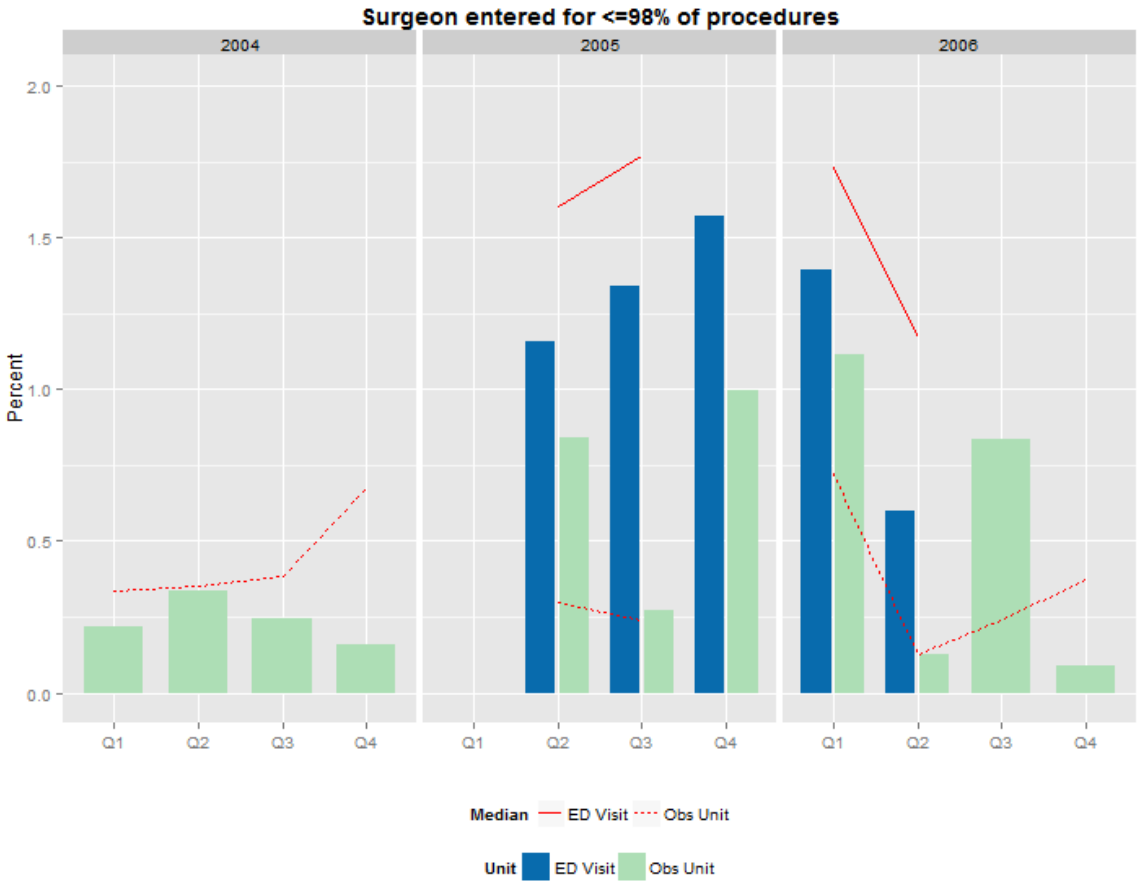
Table 4 - Possible result_names in the Result table (PHIS)

Value	Description
name=HOSPITAL VALUE|||unit=	• HOSPITAL VALUE = Summary statistics from Children's Hospital Colorado data • PHIS MEDIAN = The median value across all hospital within PHIS • Some results don't have a unit
name=HOSPITAL VALUE|||unit=$	
name=HOSPITAL VALUE|||unit=%	
name=HOSPITAL VALUE|||unit=BNs	
name=HOSPITAL VALUE|||unit=cases	
name=HOSPITAL VALUE|||unit=codes	
name=HOSPITAL VALUE|||unit=date	

name=HOSPITAL VALUE|||unit=days	
name=HOSPITAL VALUE|||unit=LOS	
name=HOSPITAL VALUE|||unit=MRNs	
name=HOSPITAL VALUE|||unit=PPx	
name=HOSPITAL VALUE|||unit=Px	
name=PHIS MEDIAN|||unit=	
name=PHIS MEDIAN|||unit=$	
name=PHIS MEDIAN|||unit=%	
name=PHIS MEDIAN|||unit=BNs	
name=PHIS MEDIAN|||unit=cases	
name=PHIS MEDIAN|||unit=codes	
name=PHIS MEDIAN|||unit=date	
name=PHIS MEDIAN|||unit=days	
name=PHIS MEDIAN|||unit=LOS	
name=PHIS MEDIAN|||unit=MRNs	
name=PHIS MEDIAN|||unit=PPx	
name=PHIS MEDIAN|||unit=Px	

Sample query 4 (in PostgreSQL dialect): It is expected that surgical procedures should contain a reference to the principle surgeon (the individual most responsible for the outcome of the surgical procedure). Thus, when a surgeon is not identified in a procedure record, this is a data quality issues – specifically missingness. A data quality check performed by the PHIS system determines if a surgeon is identified (associated with) a surgical procedure..

```
SELECT m.name, d.dim_name_1,d.dim_value_1,d.dim_name_2,split_part(d.dim_value_2, ' ', 1) AS
YEAR, split_part(d.dim_value_2, ' ', 2) AS QTR,d.dim_name_3,d.dim_value_3,
d.dim_name_4,d.dim_value_4, r1.value as hospital_value, r2.value as phis_median,
split_part(split_part(r1.result_name,'|||',2),'=',2) as unit
FROM
result r1 join dimension_set d on r1.set_id = d.set_id join measure m on r1.measure_id = m.measure_id
join result r2 on r1.measure_id=r2.measure_id and r1.set_id = r2.set_id
where m.name Like 'Surgeon entered for <=98%' AND split_part(split_part(r1.result_name,'|||',1),'=',2)
='HOSPITAL VALUE' AND split_part(split_part(r2.result_name,'|||',1),'=',2) ='PHIS MEDIAN'
```



Bar plot generated from sample query 4 data