UNIVERSITY of WASHINGTON

ITHS Institute of Translational Health Sciences
Accelerating Research. Improving Health.

Testing Data Completeness – DQe-c

10/13/18

Tim Bergquist, Graduate Student
Biomedical Informatics & Medical Education
University of Washington

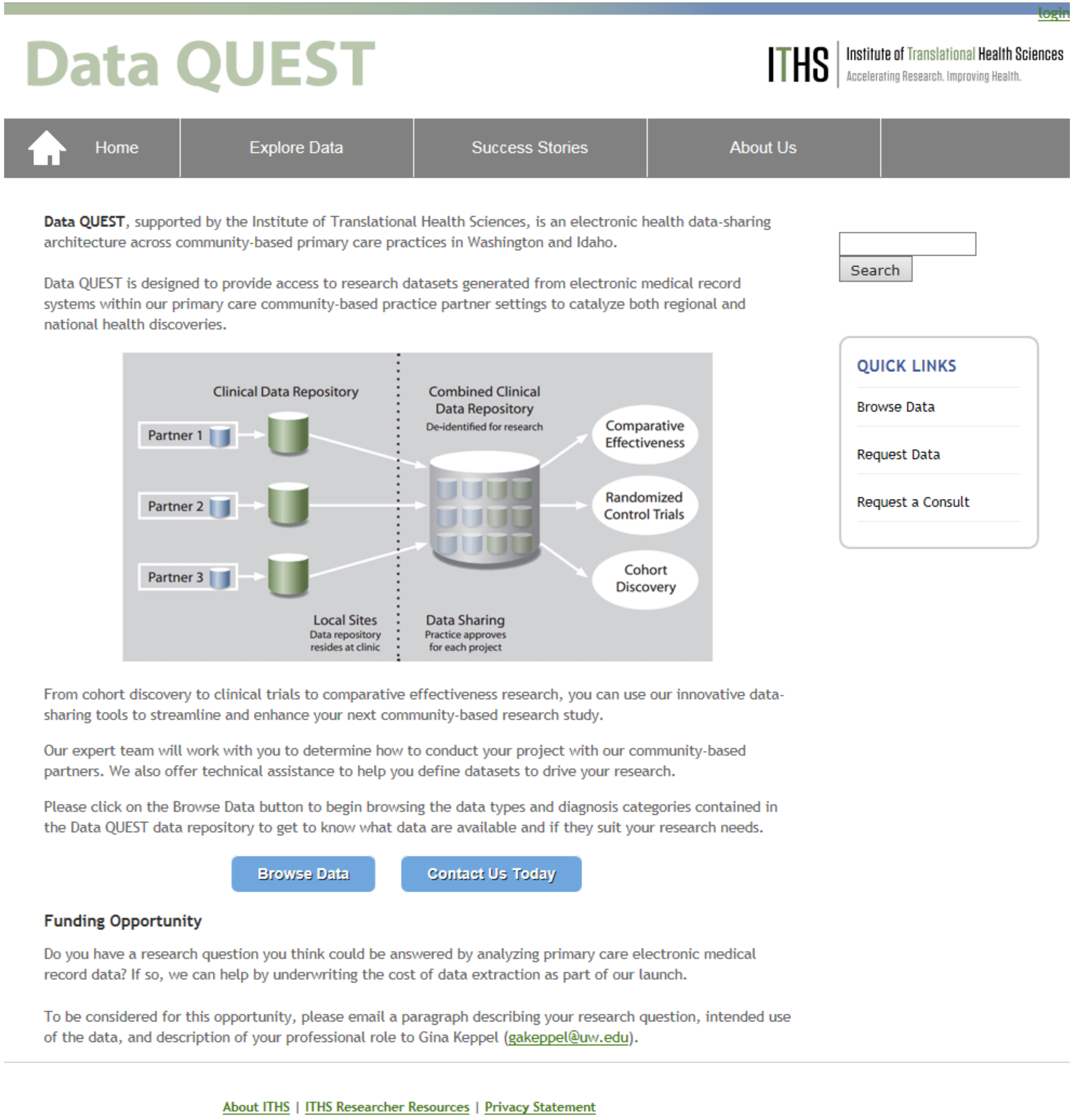
NATIONAL CENTER
FOR DATA TO HEALTH

WWAMI region Practice & Research Network

WPRN

Collaborative Research. Innovative Care.

- 60+ Primary care WWAMI clinics
- ~20 data connected clinics
- CHCs and RHCs
- Underserved populations
- Many serving rural populations
- Collaboration with national network of practice based research networks
- Data QUEST represents over 250,000 patients https://dataquest.iths.org/



login

Data QUEST

ITHS Institute of Translational Health Sciences
Accelerating Research. Improving Health.

Home | Explore Data | Success Stories | About Us

Data QUEST, supported by the Institute of Translational Health Sciences, is an electronic health data-sharing architecture across community-based primary care practices in Washington and Idaho.

Data QUEST is designed to provide access to research datasets generated from electronic medical record systems within our primary care community-based practice partner settings to catalyze both regional and national health discoveries.

Clinical Data Repository
Partner 1
Partner 2
Partner 3
Local Sites
Data repository resides at clinic
Combined Clinical Data Repository
De-identified for research
Data Sharing
Practice approves for each project
Comparative Effectiveness
Randomized Control Trials
Cohort Discovery

From cohort discovery to clinical trials to comparative effectiveness research, you can use our innovative data-sharing tools to streamline and enhance your next community-based research study.

Our expert team will work with you to determine how to conduct your project with our community-based partners. We also offer technical assistance to help you define datasets to drive your research.

Please click on the Browse Data button to begin browsing the data types and diagnosis categories contained in the Data QUEST data repository to get to know what data are available and if they suit your research needs.

Browse Data | Contact Us Today

Funding Opportunity

Do you have a research question you think could be answered by analyzing primary care electronic medical record data? If so, we can help by underwriting the cost of data extraction as part of our launch.

To be considered for this opportunity, please email a paragraph describing your research question, intended use of the data, and description of your professional role to Gina Keppel (gakeppel@uw.edu).

Search

QUICK LINKS
- Browse Data
- Request Data
- Request a Consult

About ITHS | ITHS Researcher Resources | Privacy Statement

Data QUEST supports numerous grants

Across 14 clinical domains

> **17 awarded**

> **Large trials to small training grants for junior investigators**

> **Topics go beyond primary care**
> - **Industry**
> - **Specialty areas**

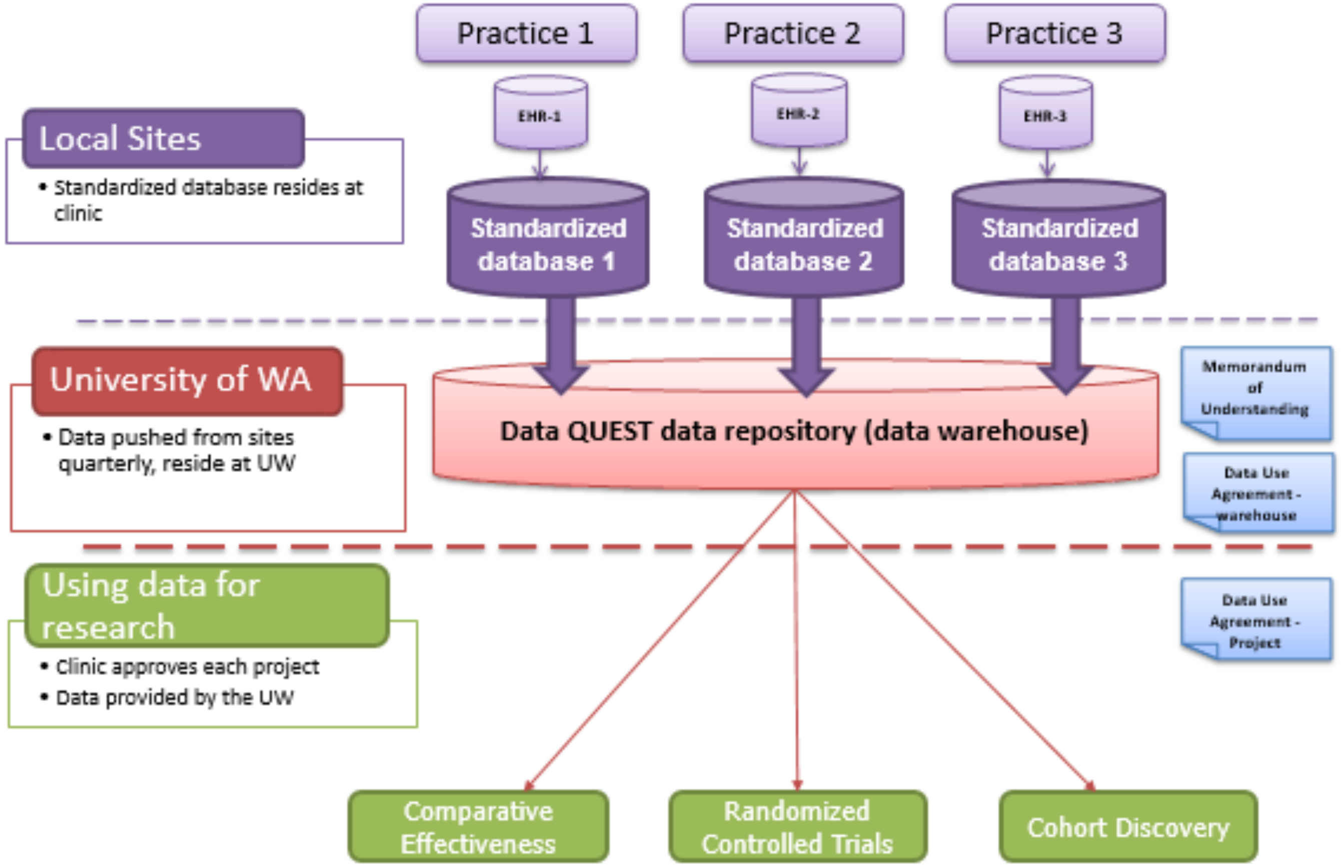
Supporting $100.4M in funded projects addressing:

- prescription opioid management re-design in primary care
- complex patients with multiple chronic diseases
- smoking cessation
- weight loss
- integrated behavioral health in primary care
- pharmacogenomics
- diabetes prevention
- acute pain
- use of handheld ultrasound scans in primary care
- substance use disorders
- practice transformation
- contraceptive guidelines
- drug safety
- antibiotic prescribing

Data QUEST

- 20 data-connected clinics in the WPRN
- Represents over 250,000 patients

An electronic health data-sharing architecture across community-based primary care practices in the WPRN



Practice 1

Practice 2

Practice 3

EHR-1

EHR-2

EHR-3

Standardized database 1

Standardized database 2

Standardized database 3

Local Sites

- Standardized database resides at clinic

University of WA

- Data pushed from sites quarterly, reside at UW

Data QUEST data repository (data warehouse)

Memorandum of Understanding

Data Use Agreement - warehouse

Using data for research

- Clinic approves each project
- Data provided by the UW

Data Use Agreement - Project

Comparative Effectiveness

Randomized Controlled Trials

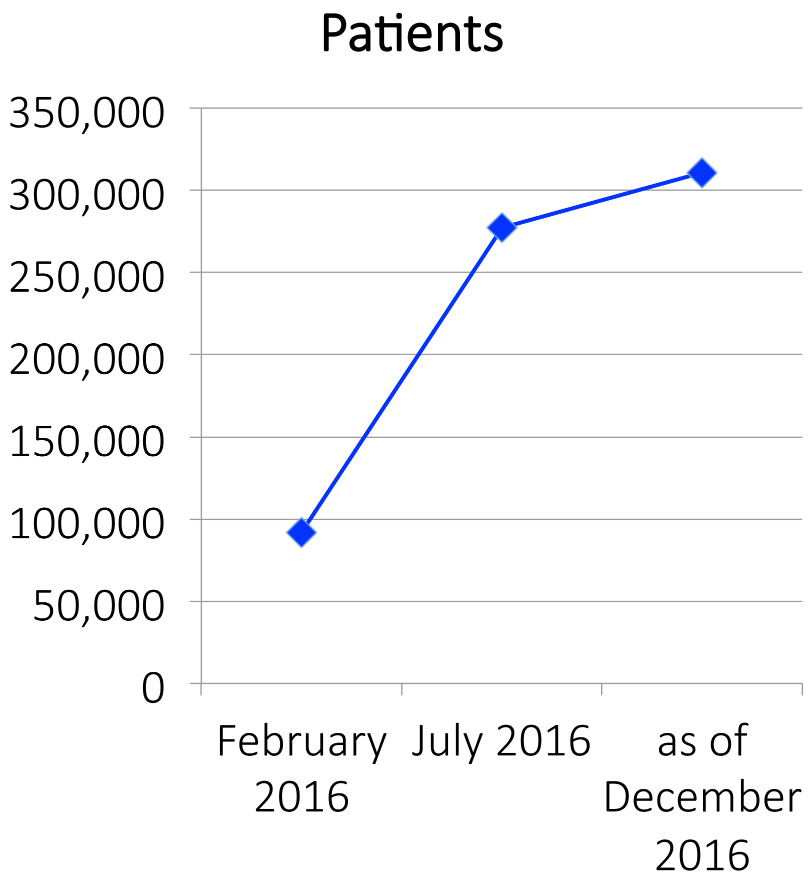
Cohort Discovery

Current UW-hosted Data QUEST Warehouse Patients

310,604 patients in the person table

- 102,330 (33%) at Organization B
- 45,685 (15%) at Organization C
- 27,577 (9%) at Organization N
- 36,001 (12%) at Organization P
- 99,011 (32%) at Organization Y

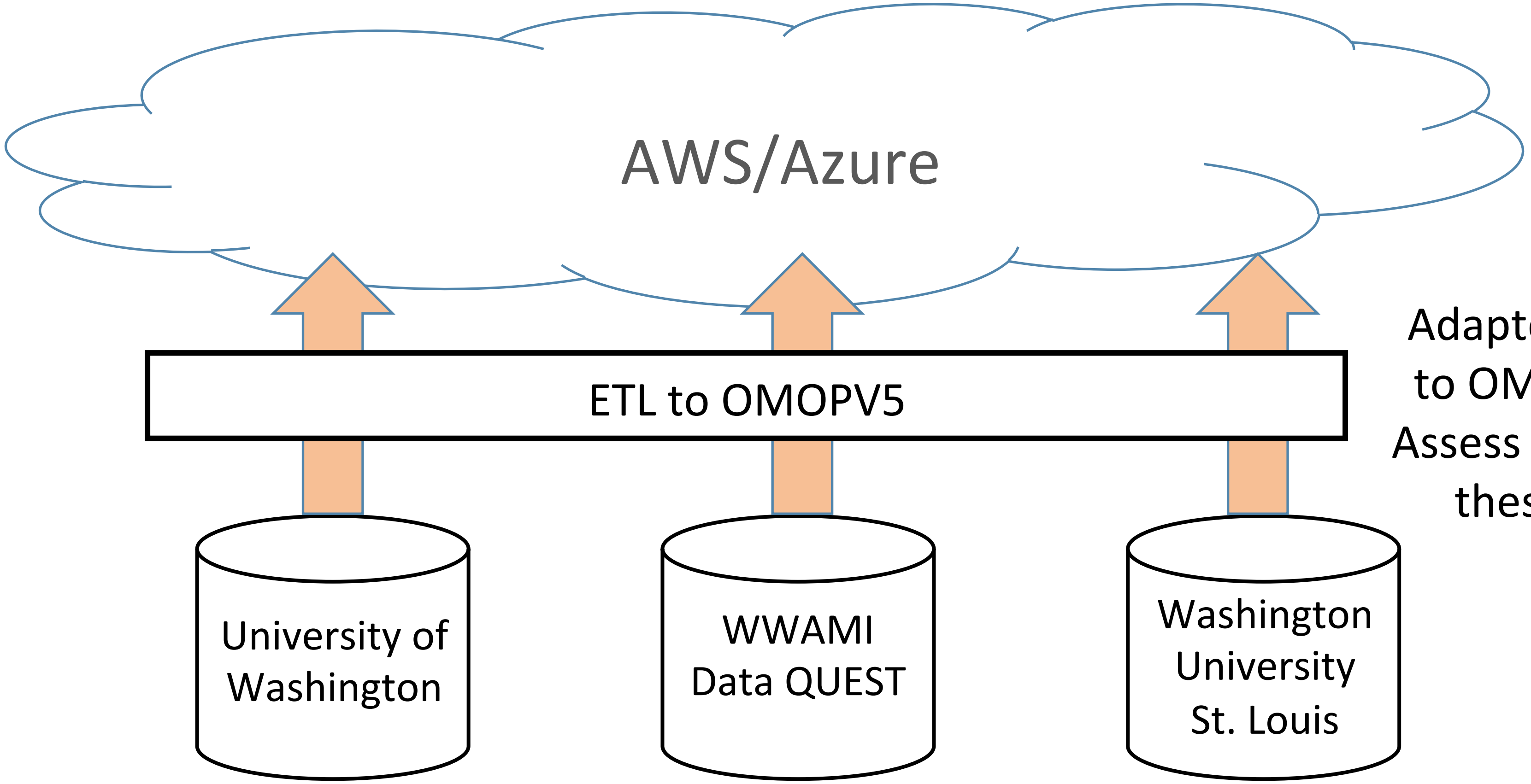
10M encounters



WPRN data needed quality validation

- DQe-c was developed at the University of Washington by Kari Stephens and Hossein Estiri.
- Needed to test and visualize data completeness in the WWAMI network.
- The first iteration worked with OMOP V4 but was mainly run on the client side (not database side)
- Second iteration was improved by Hossein to work with PCORnet CDM. This version was more efficient in processing.

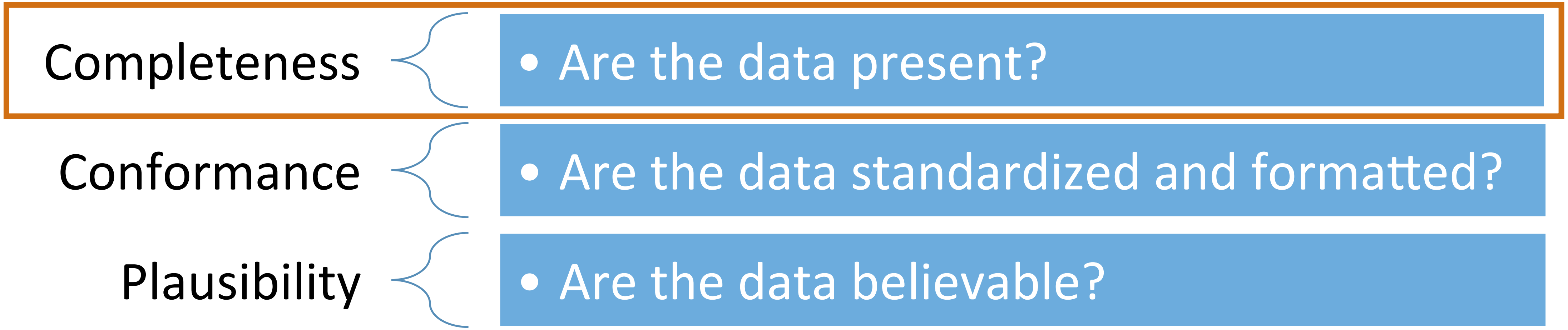
CD2H Multisite Data Integration



Adapted DQe-c to OMOPV5 to Assess Quality of these data

Measuring Data Quality Framework

Operationalizing the framework into: 5 conceptual tests and 17 discrete tests across:



Completeness

- Are the data present?

Conformance

- Are the data standardized and formatted?

Plausibility

- Are the data believable?

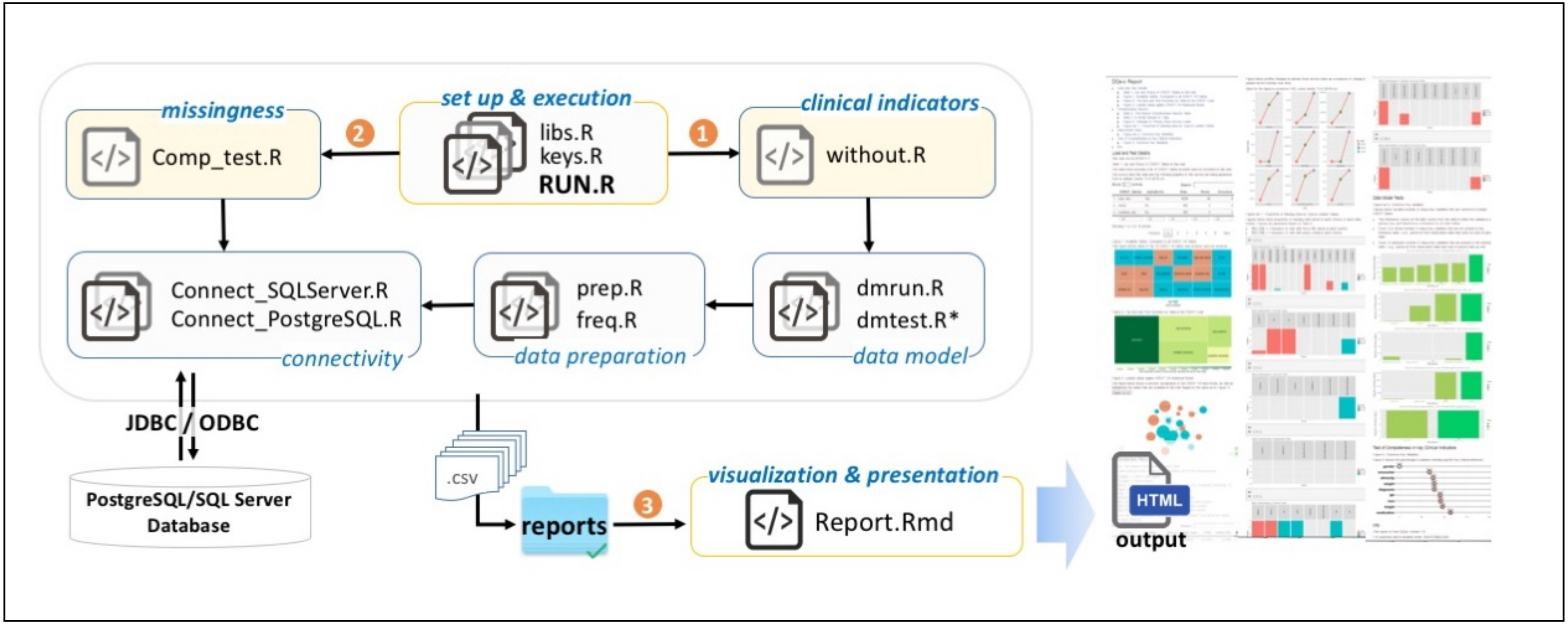
Kahn et al. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMS, 4, 1244. https://www.ncbi.nlm.nih.gov/pubmed/27713905

Data Quality Tests

DQ Framework category	TEST
COMPLETENESS	Number of Tables Received, Number of Observations, Flag Indicator for the table having actual data
COMPLETENESS	GENDER completeness (denominator and proportion with valid data)
COMPLETENESS	Key clinical status completeness (denominator and proportion with valid data): Smoking status, alcohol consumption
COMPLETENESS	VITALS completeness (denominator and proportion with valid data): Height, Weight, SBP, DBP
COMPLETENESS	Cross reference tables that are present in current dataset to expected tables in standard OMOP CDM
COMPLETENESS	Looks for NULL and invalid variable values in each column and visualizes percent missingness
CONFORMANCE	Check that primary and foreign keys relate properly; High Priority: Person_ID, Visit_Occurrence_ID
CONFORMANCE	Checks that orphan don't keys exist (a foreign key is present in a table but no primary key exists in the reference table)
CONFORMANCE	Visualize codes/values entered for DEMOGRAPHICS (Gender, Race, Ethnicity)
PLAUSIBILITY	Comparison of new load to old load (Number of observations, Number of unique patients, Number of tables with rows)
PLAUSIBILITY	Size of tables and rows across the OMOP CDM

DQe-c Tool

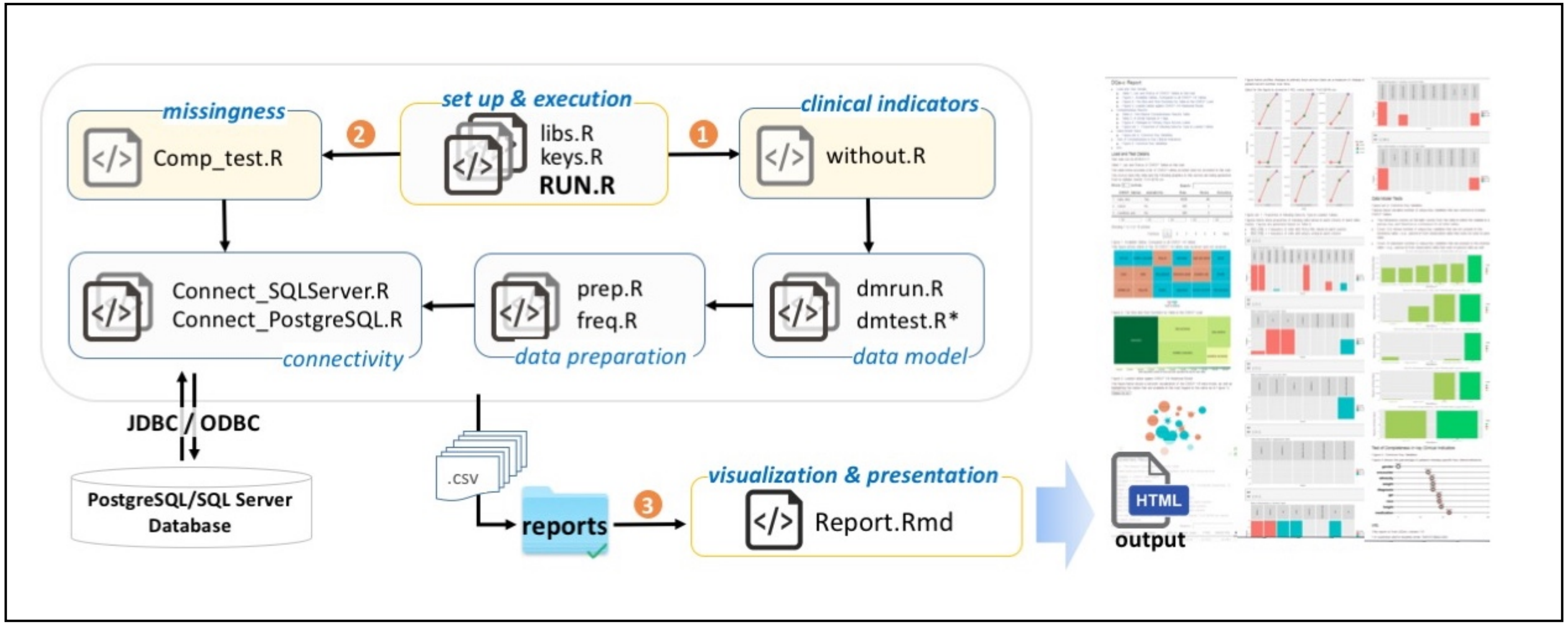
Modular tool developed in R statistical language for assessing **completeness** in EHR data repositories.

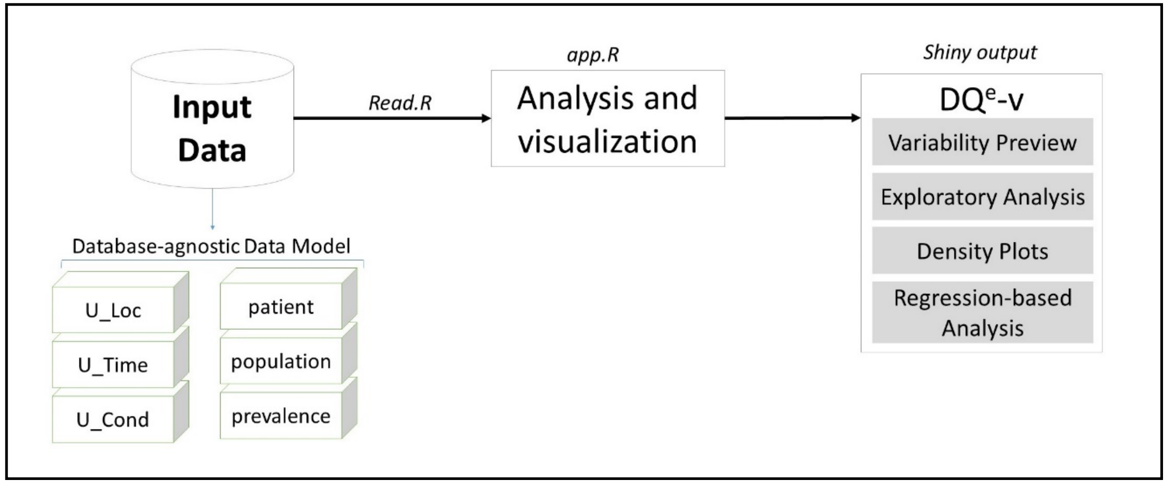


DQe Tool Architecture

DQe-c

modular tool developed in R statistical language for assessing completeness in EHR data repositories





DQe-v

interactive interface powered by the shiny package version 0.13.0 in R

DQe-c Tool

Clinical Indicators

Checks for common clinical variables and reports percent missing.

Example: What percentage of patients have a blood pressure reading

Missingness

Checks that all tables in the reference CDM are present, and reports missing tables.

Checks all columns in the CDM and reports the percentage of rows that are missing valid data.

Data Model

Checks for orphan keys in foreign tables.

Data Preparation

Gathers necessary data to run calculations.

Builds data frames and reports table and row sizes

Visualization and Presentation

Builds an HTML report of all the tests

Operationalizing use of DQe tools for data quality testing

* Data QUEST
* DARTNet Institute
* CD2H

DQe-c/DQe-v Reports Standard Operating Procedure (SOP)

Version 2 December 2016

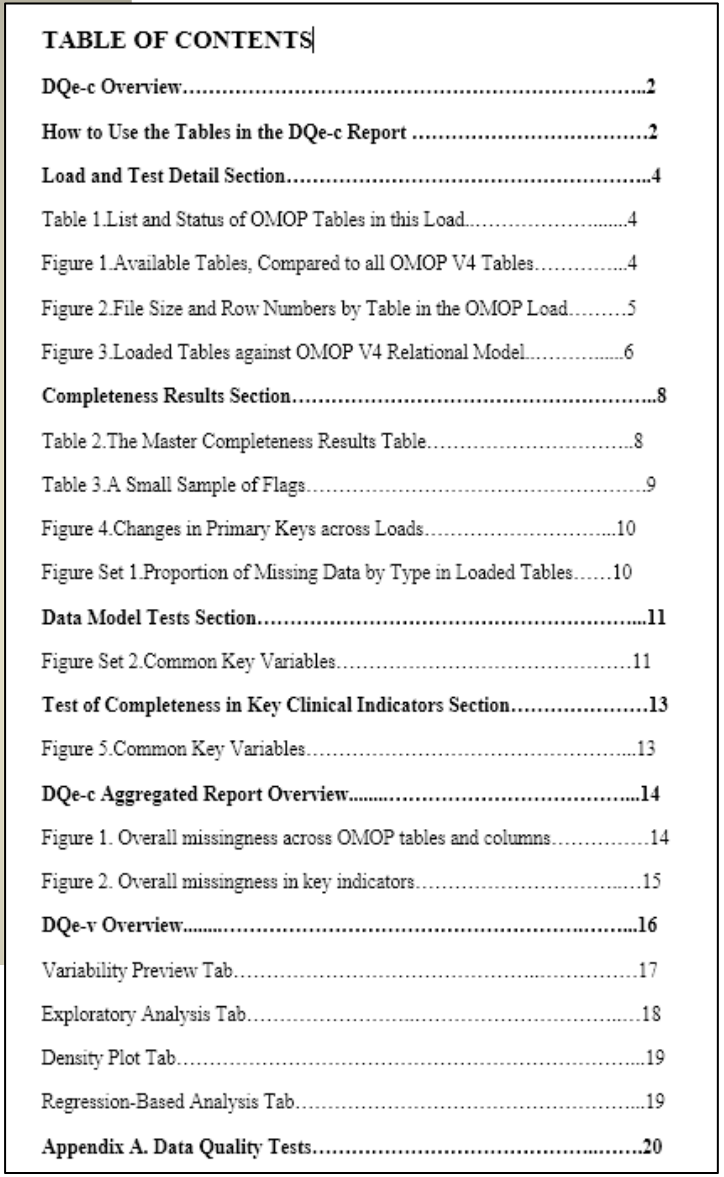
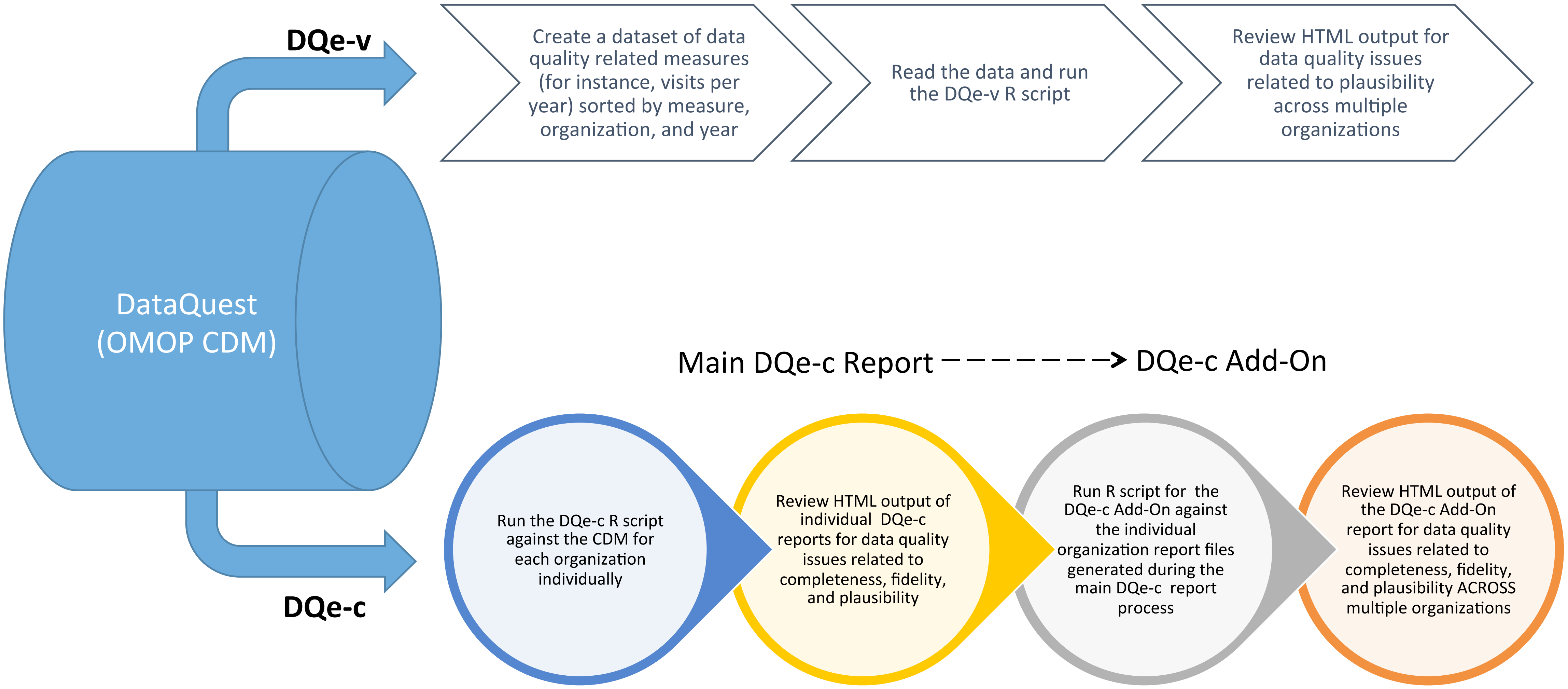


TABLE OF CONTENTS

DQe-c and DQe-v Report Flows



DQe-v

Create a dataset of data quality related measures (for instance, visits per year) sorted by measure, organization, and year

Read the data and run the DQe-v R script

Review HTML output for data quality issues related to plausibility across multiple organizations

DataQuest (OMOP CDM)

Main DQe-c Report ------→ DQe-c Add-On

DQe-c

Run the DQe-c R script against the CDM for each organization individually

Review HTML output of individual DQe-c reports for data quality issues related to completeness, fidelity, and plausibility

Run R script for the DQe-c Add-On against the individual organization report files generated during the main DQe-c report process

Review HTML output of the DQe-c Add-On report for data quality issues related to completeness, fidelity, and plausibility ACROSS multiple organizations

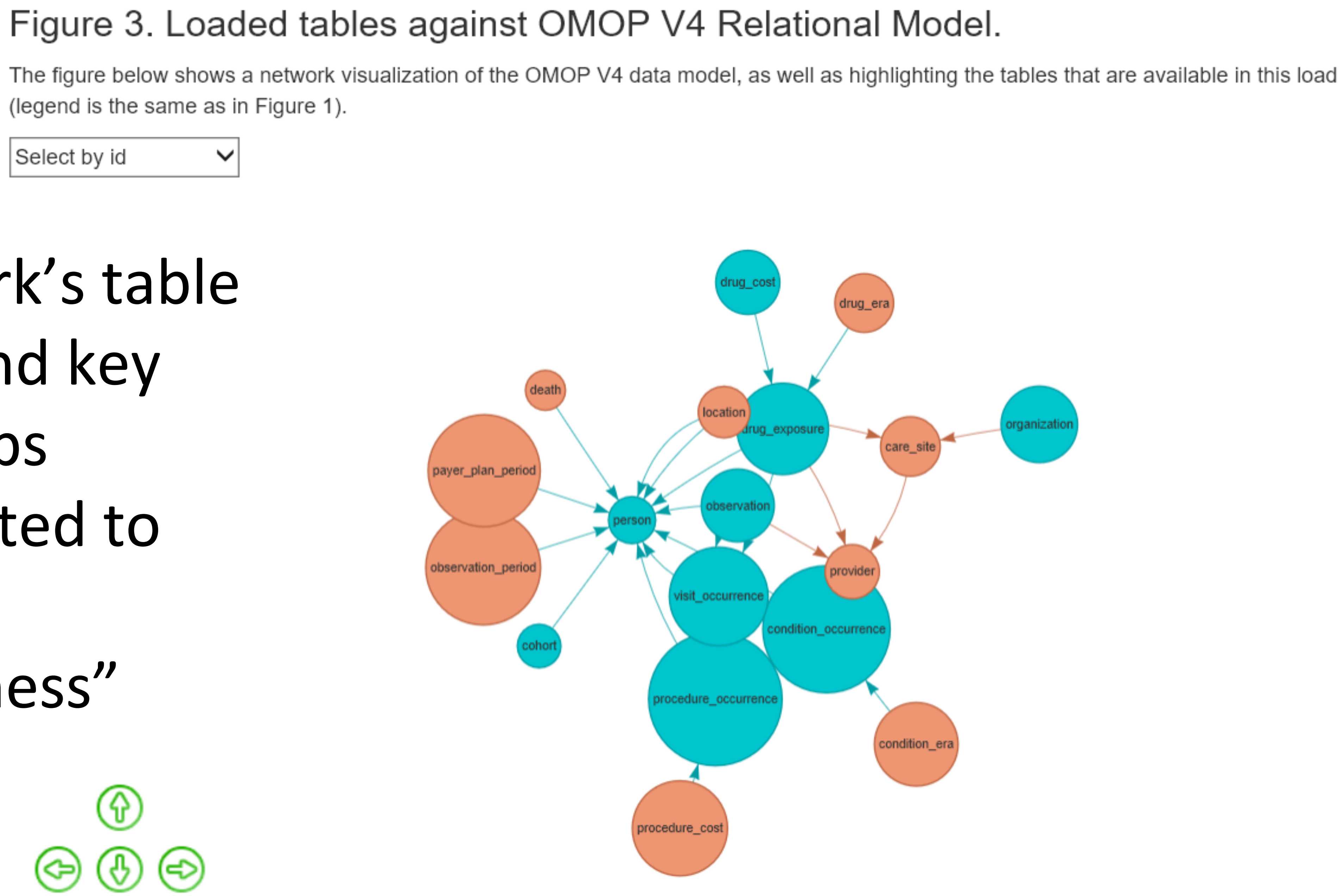


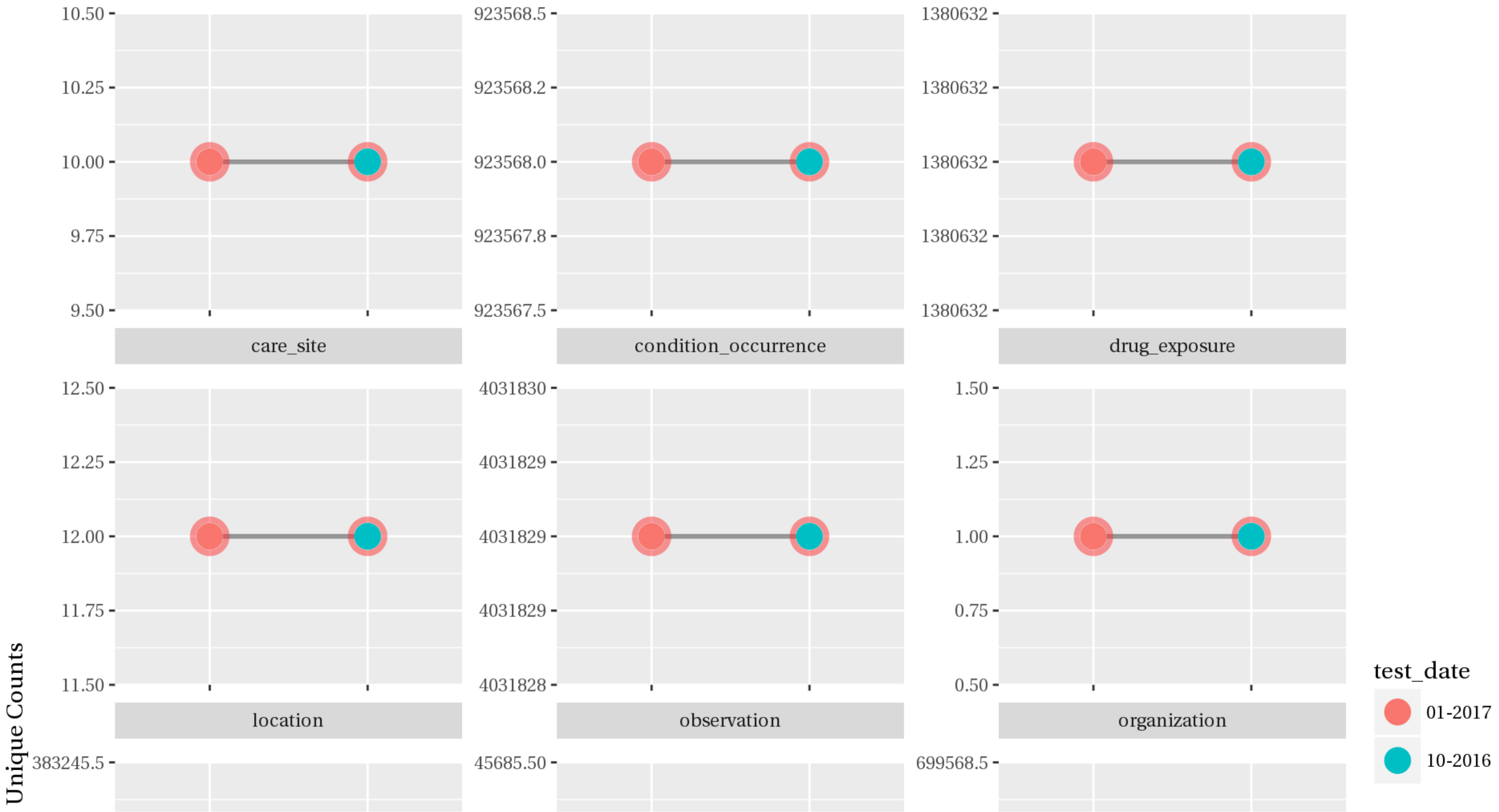
Figure 3. Loaded tables against OMOP V4 Relational Model.

The figure below shows a network visualization of the OMOP V4 data model, as well as highlighting the tables that are available in this load (legend is the same as in Figure 1).

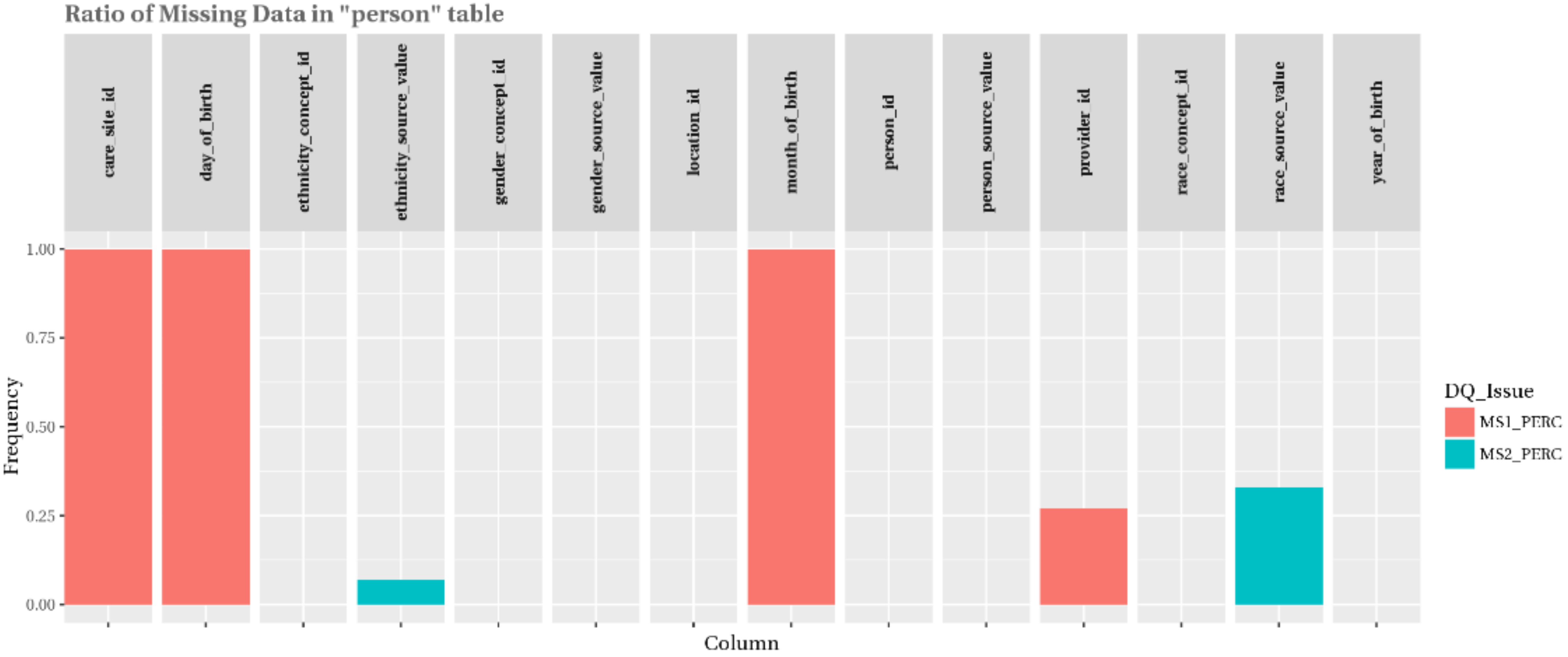
The network’s table schemas and key relationships

- Color coated to display “missingness”

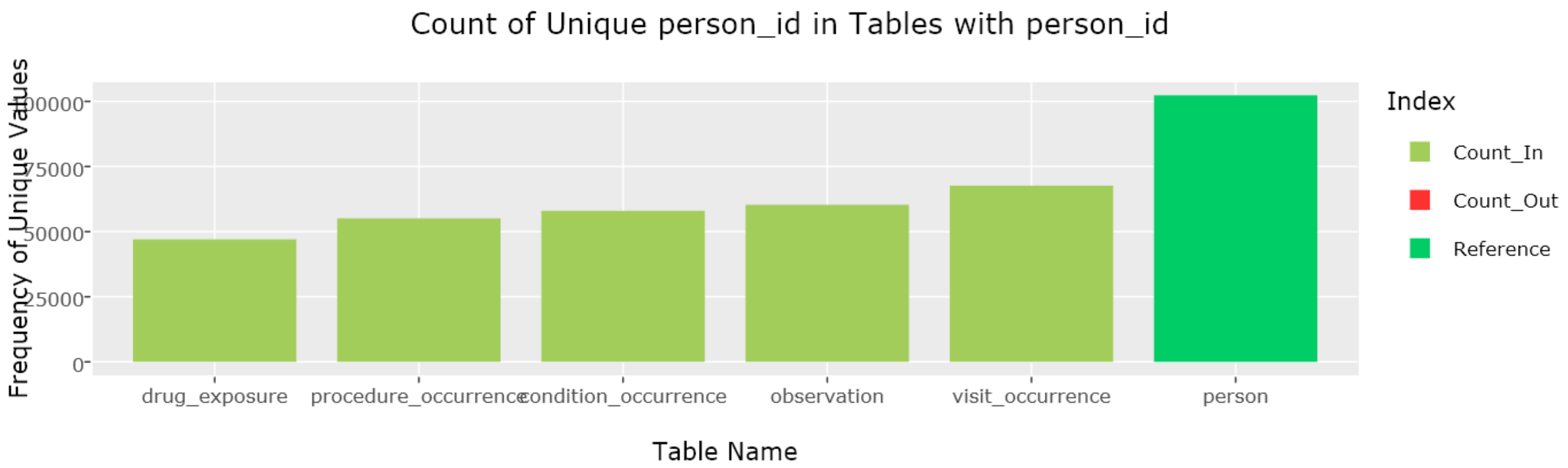
Completeness example:
Number of primary keys for available tables over time



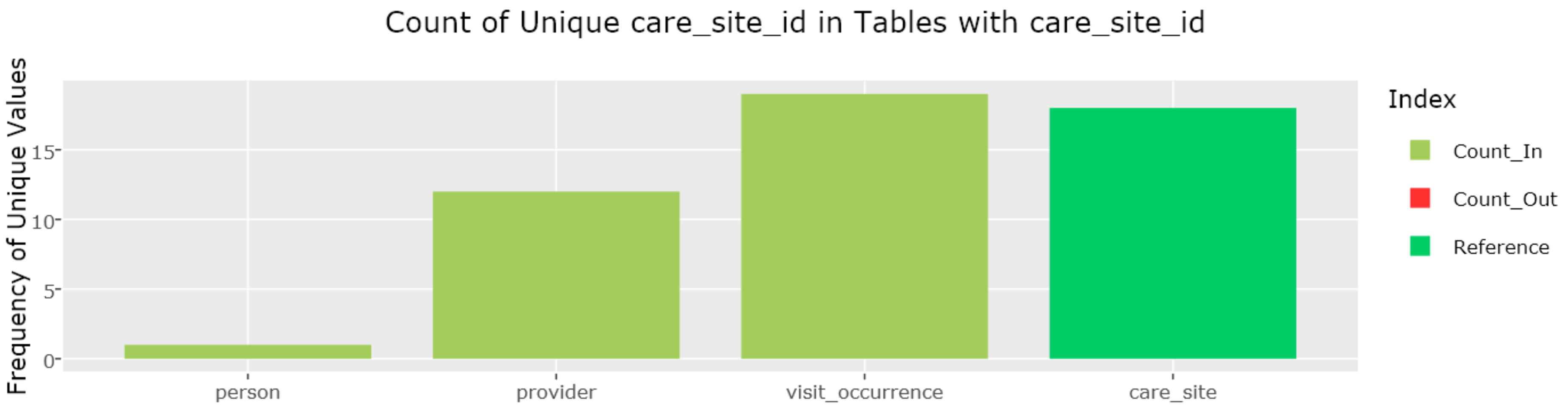
Completeness example:
Detailing columns with proportion of missingness (null vs. blank)



Fidelity example:
Detailing totals of key overlap across core tables

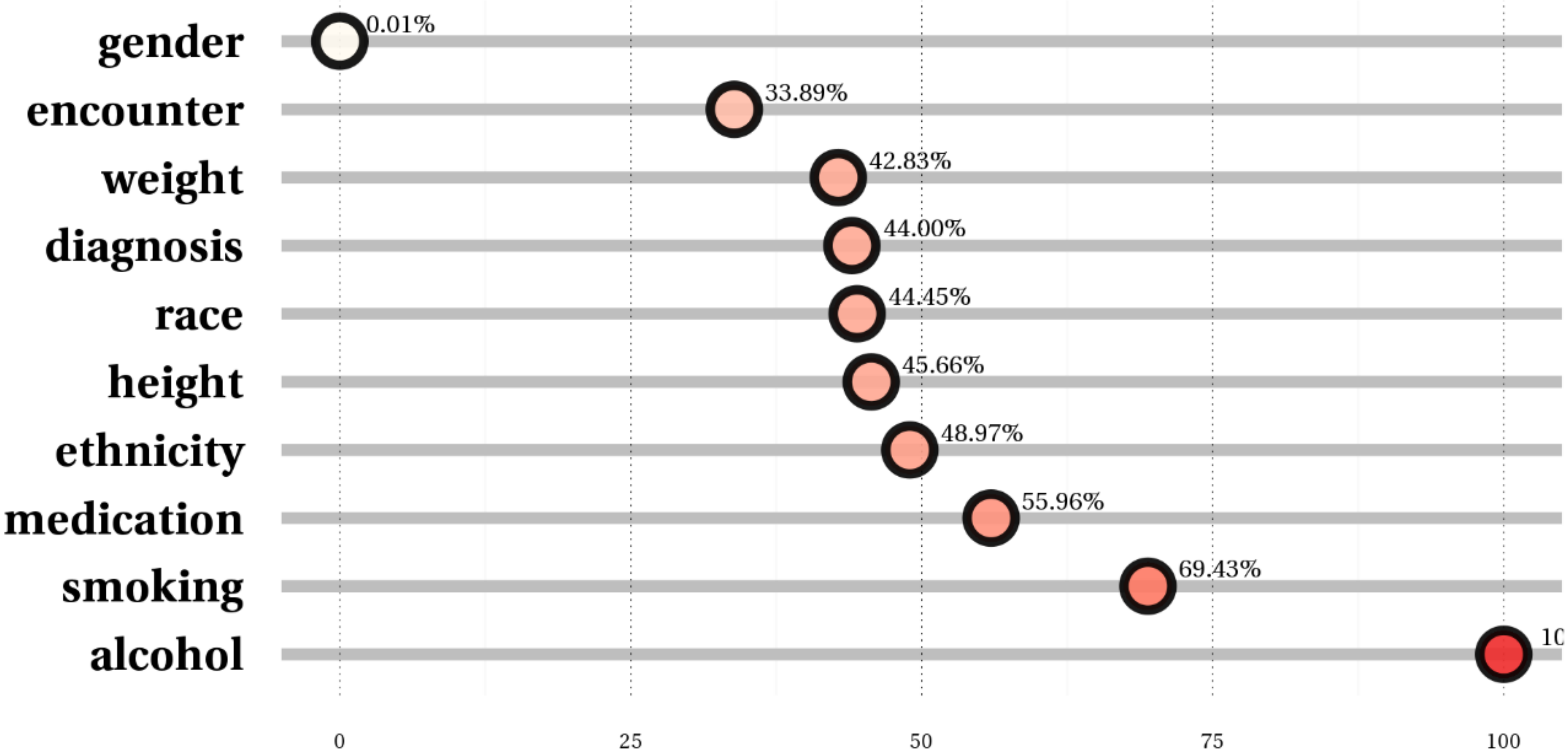


Count of Unique person_id in Tables with person_id



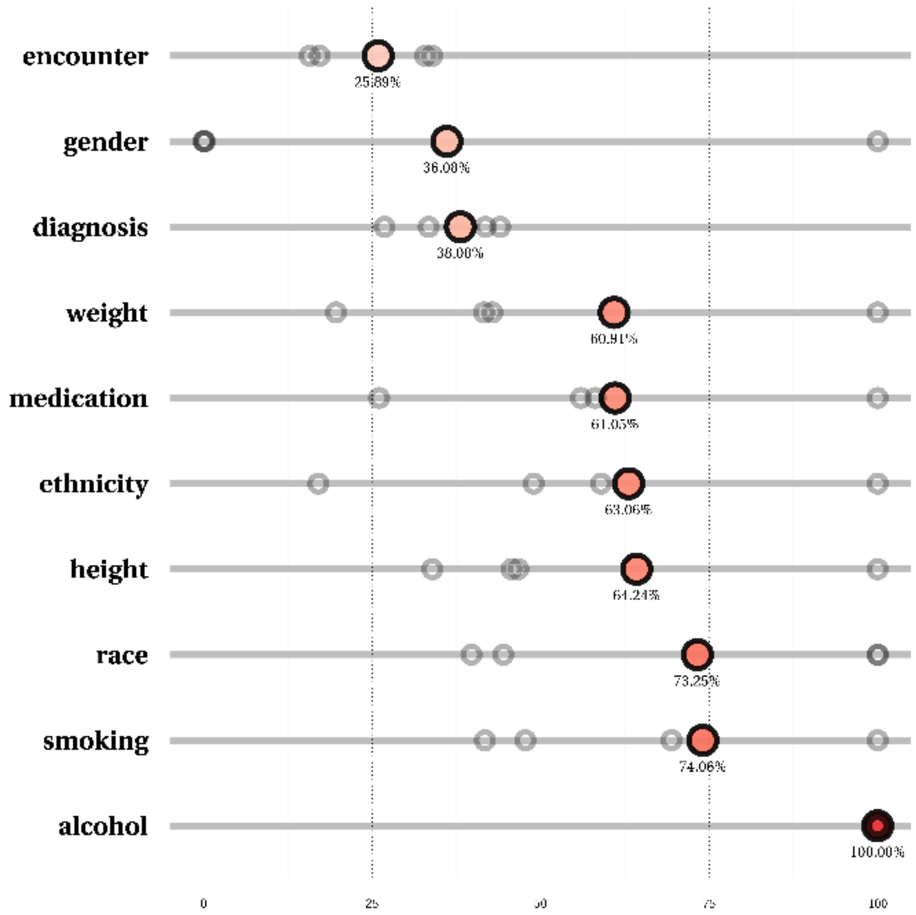
Count of Unique care_site_id in Tables with care_site_id

Completeness/Fidelity example: Percent of patients missing specific key clinical indicators

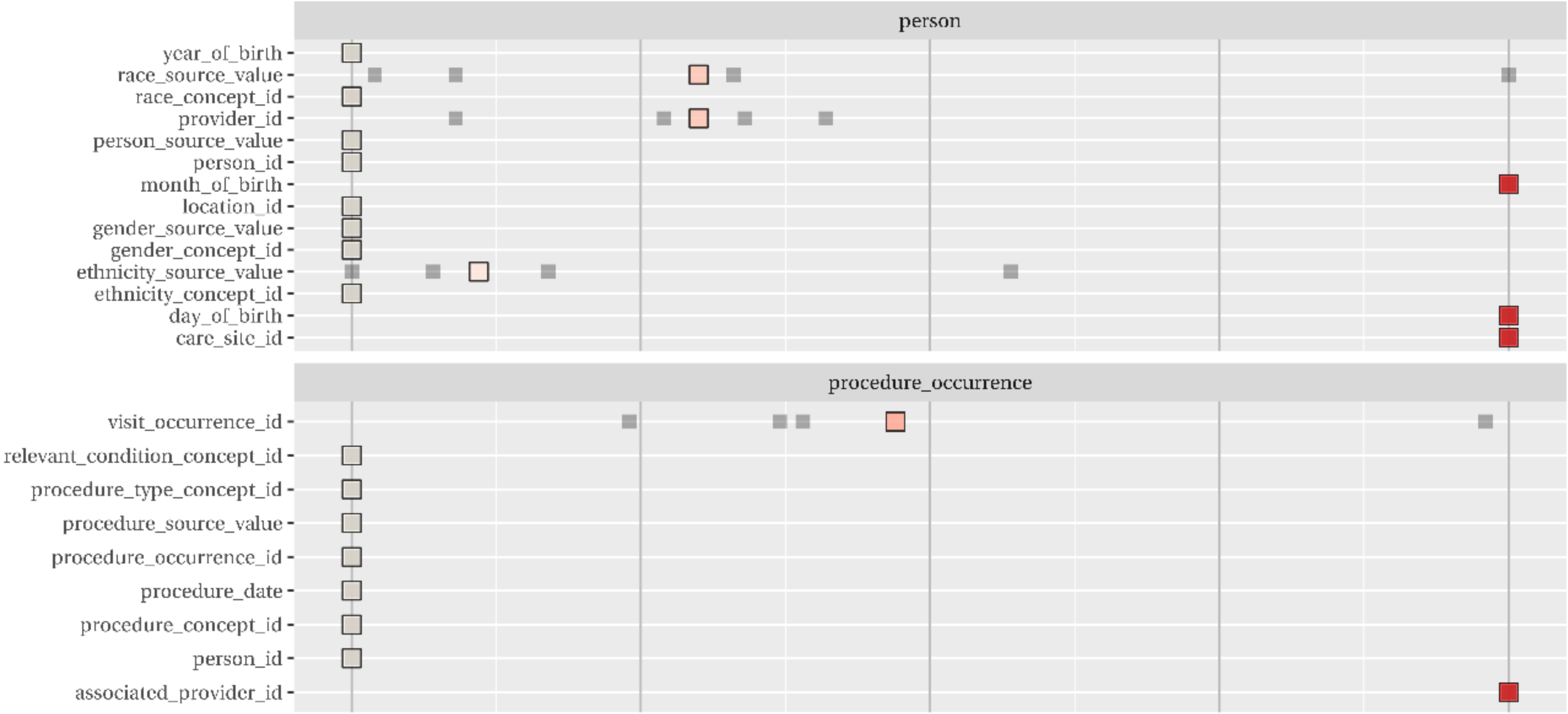


Completeness/Fidelity example across sites: Percent of patients missing specific key clinical indicators

Figure 2. Overall missingness in key indicators



Completeness example across sites/clinics: Percent of patients missing in columns across sites



Next Steps

- Make DQe-c compatible with PostgreSQL and ORACLE
- Add new tests as needed…

Thank you!

Contact: Tim Bergquist
trberg@uw.edu

https://dataquest.iths.org/

https://github.com/WWAMI-DataQuest