UNIVERSITY of WASHINGTON

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

Testing Data Completeness with DQe-c-v2

OHDSI Symposium 2019: Data Quality Workshop
09/17/19

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

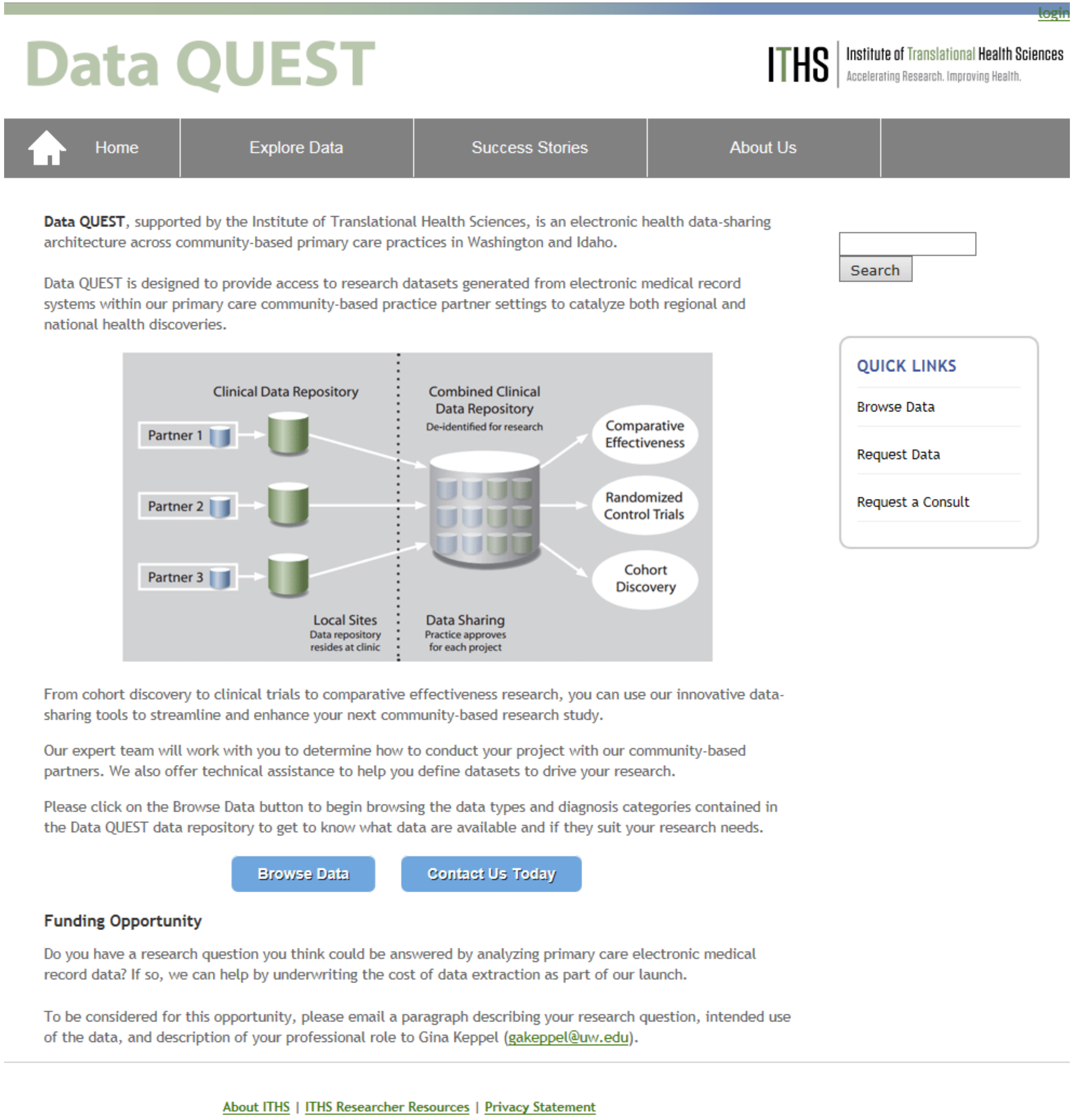
NATIONAL CENTER
FOR DATA TO HEALTH

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
Search
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
Browse Data
Contact Us Today

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.

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.

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).

QUICK LINKS

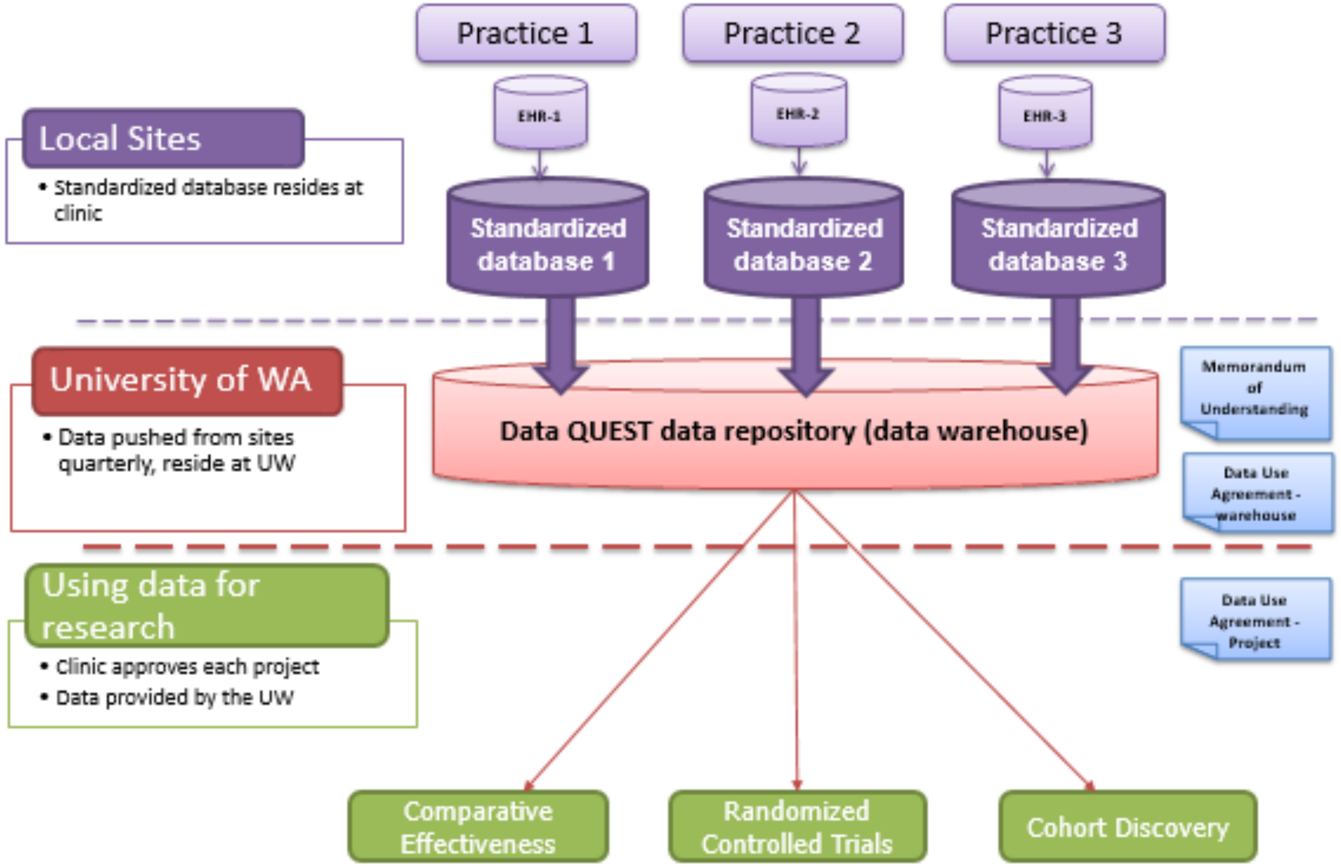
- Browse Data
- Request Data
- Request a Consult

About ITHS | ITHS Researcher Resources | Privacy Statement

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

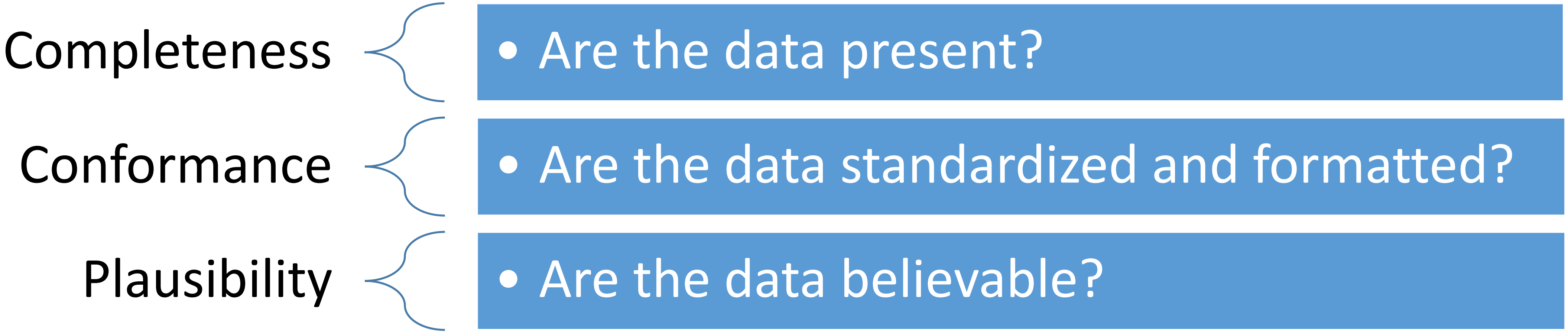
Cohort Discovery

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

WPRN
Collaborative Research. Innovative Care.

Measuring Data Quality Framework

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

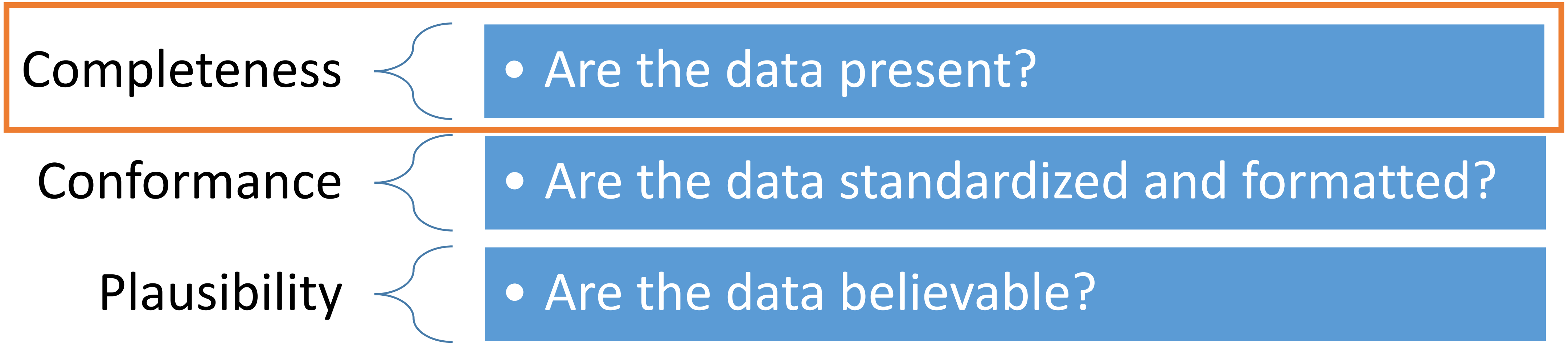


Completeness	• Are the data present?
Conformance	• Are the data standardized and formatted?
Plausibility	• Are the data believable?

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

Measuring Data Quality Framework

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



Completeness
- Are the data present?

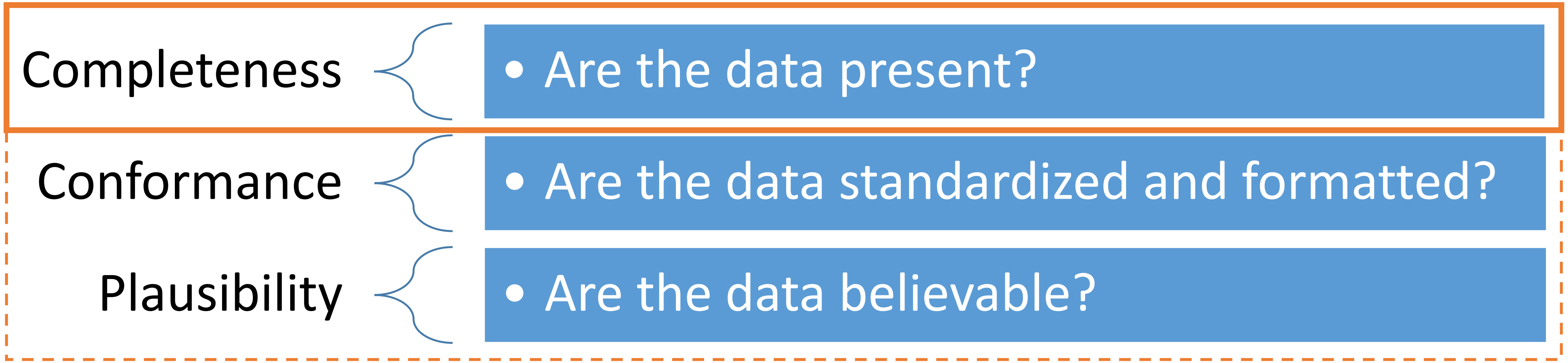
Conformance
- Are the data standardized and formatted?

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

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

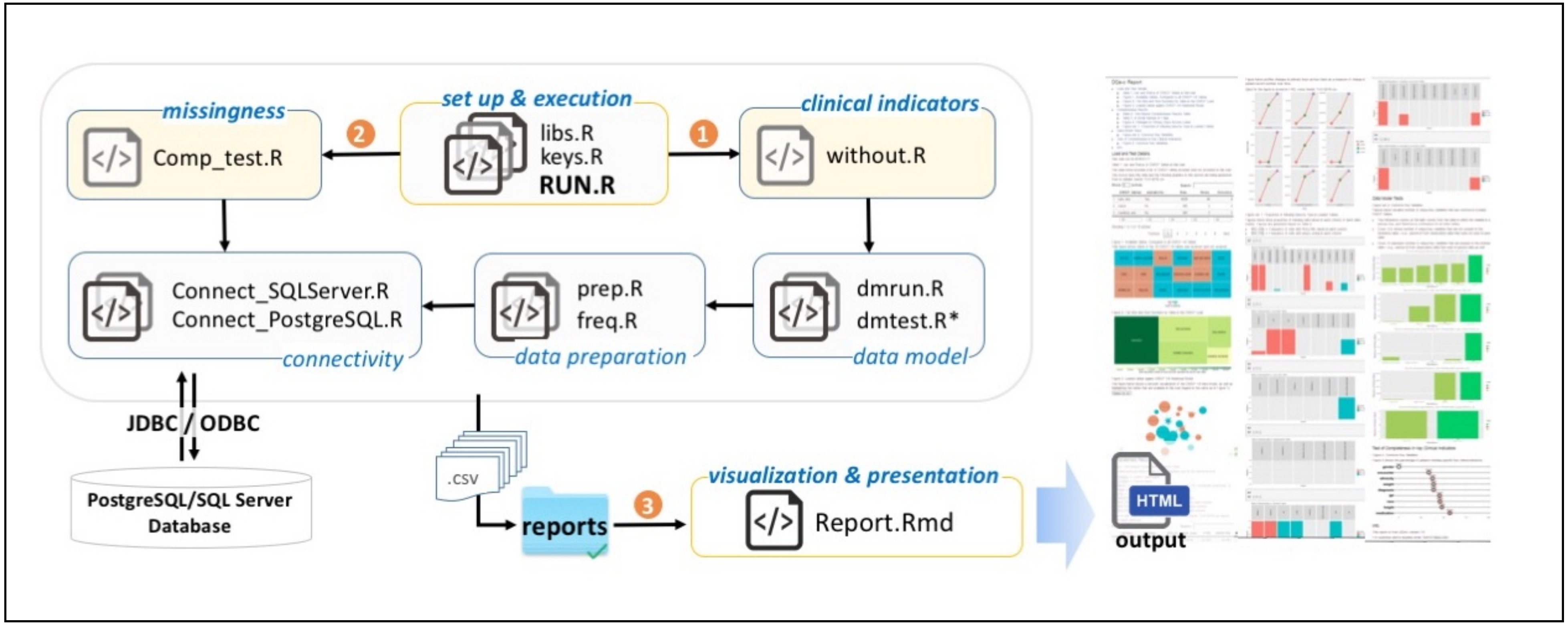
DQ Framework category	TEST
COMPLETENESS	Gender, Visit, Observation completeness (denominator and proportion with valid data)
COMPLETENESS	Key clinical status completeness (denominator and proportion with valid data): Smoking status, alcohol consumption
COMPLETENESS	Measurement completeness (denominator 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)
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

Original DQe-c Tool

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

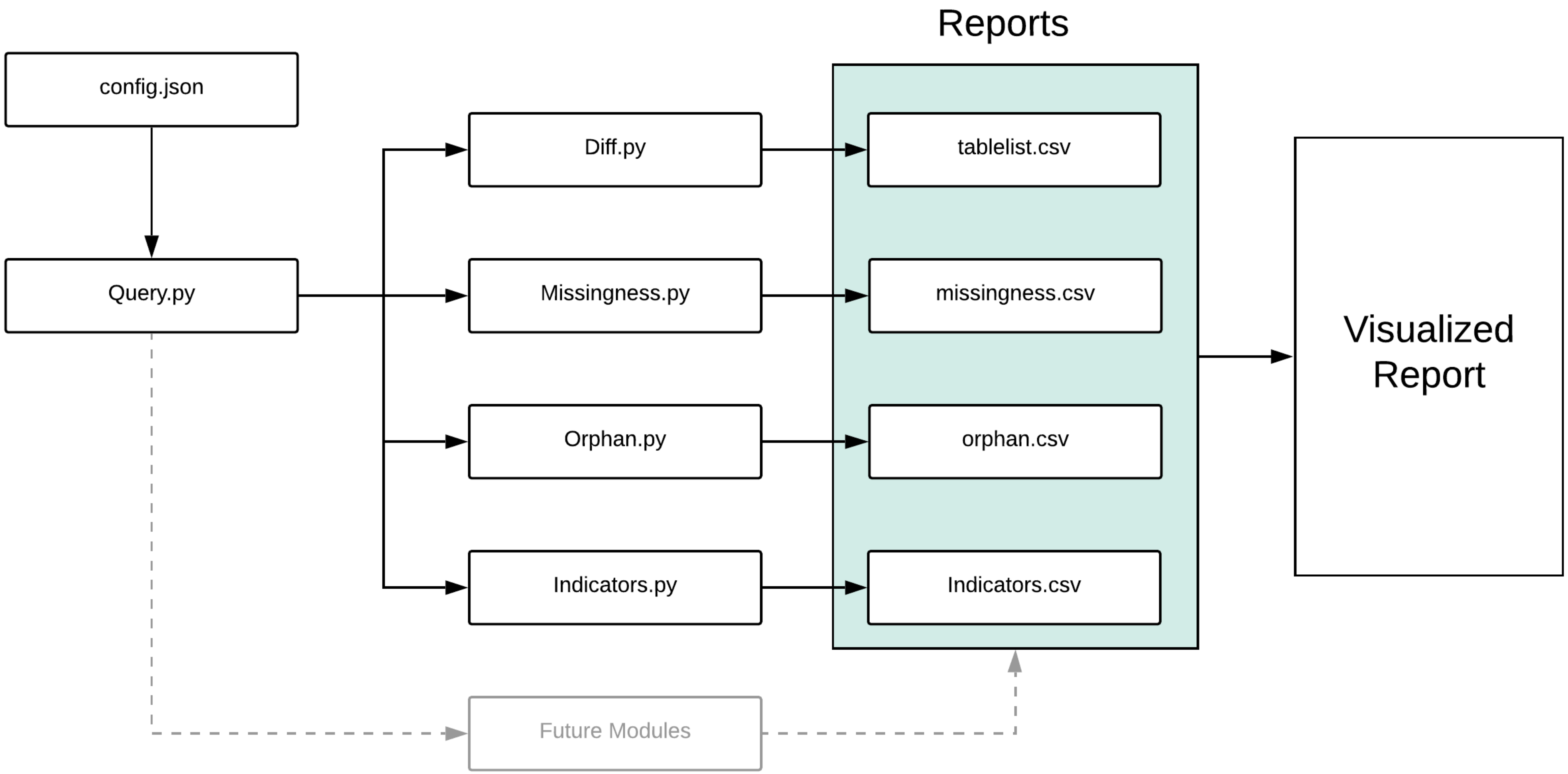
Customization and configuration was difficult

- Hard to add new modules
- Difficult to add new CDMs (or new versions of CDMs)



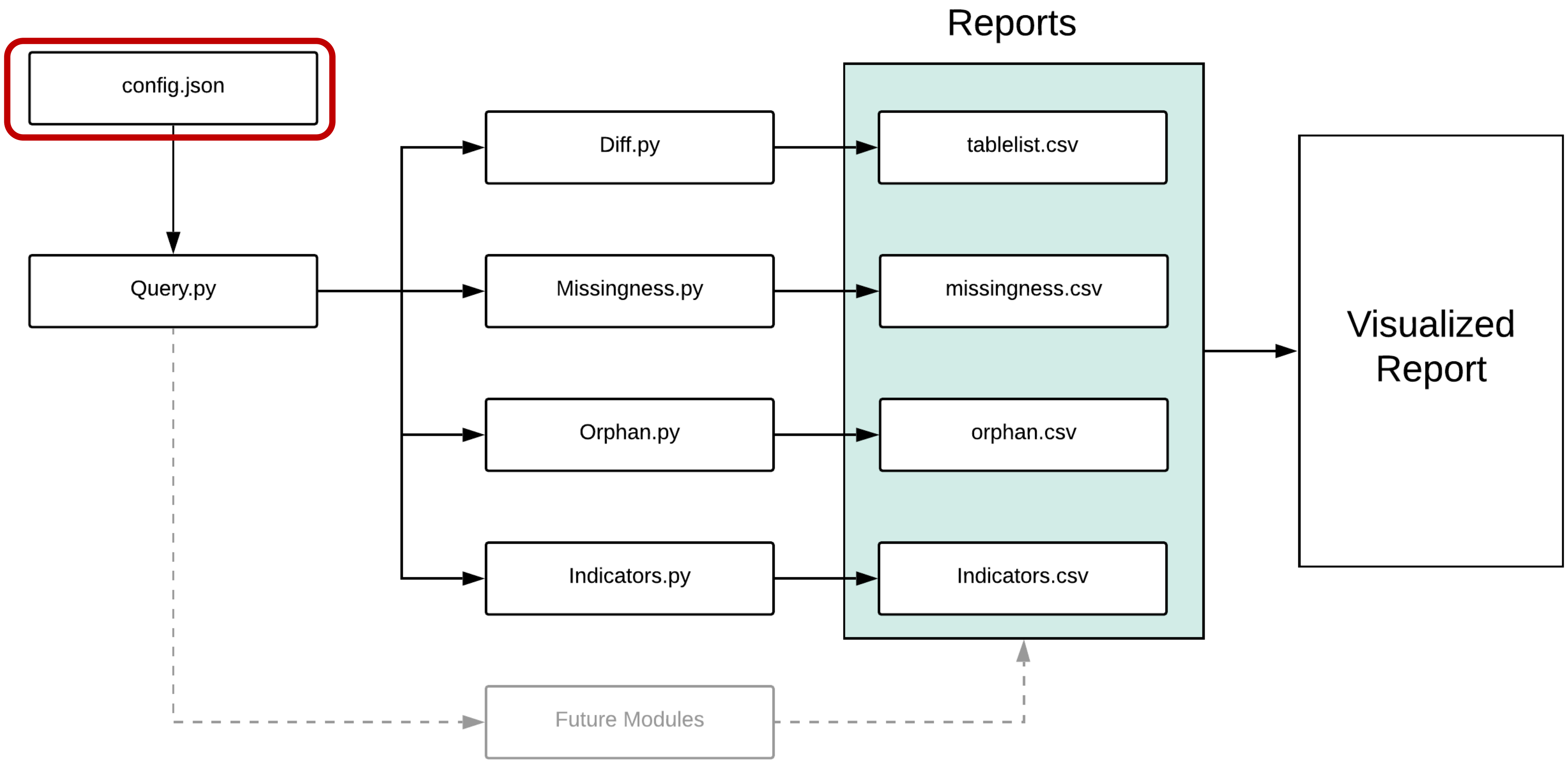
DQe-c-v2 Tool

Modular tool developed in python for assessing **completeness** in EHR data repositories.



DQe-c-v2 Tool

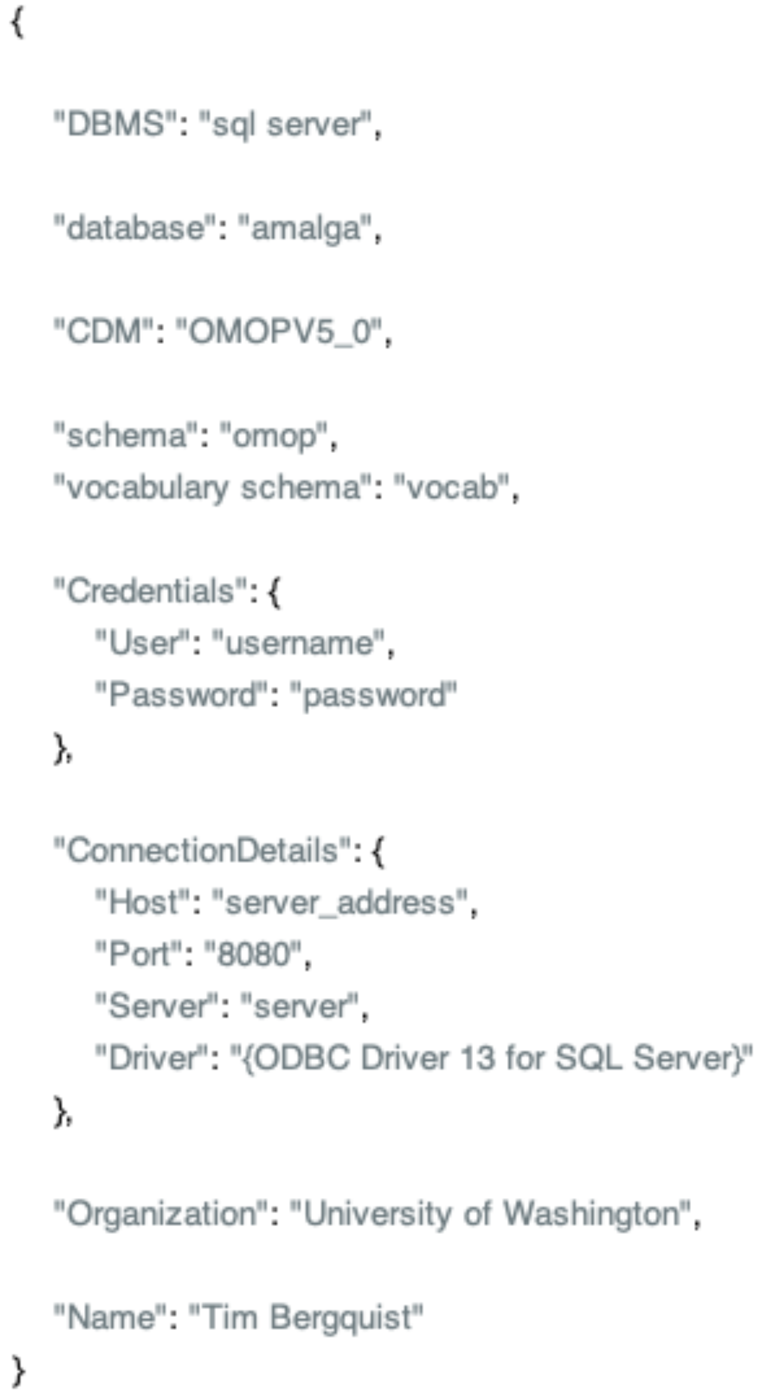
Takes in the database credentials, CDM version, and configurations.



DQe-c-v2 Tool

Takes in the database credentials, CDM version, and configurations.

Simply enter your credentials and configurations into the config.json file.



```
{

  "DBMS": "sql server",

  "database": "amalga",

  "CDM": "OMOPV5_0",

  "schema": "omop",
  "vocabulary schema": "vocab",

  "Credentials": {
    "User": "username",
    "Password": "password"
  },

  "ConnectionDetails": {
    "Host": "server_address",
    "Port": "8080",
    "Server": "server",
    "Driver": "{ODBC Driver 13 for SQL Server}"
  },

  "Organization": "University of Washington",

  "Name": "Tim Bergquist"
}
```

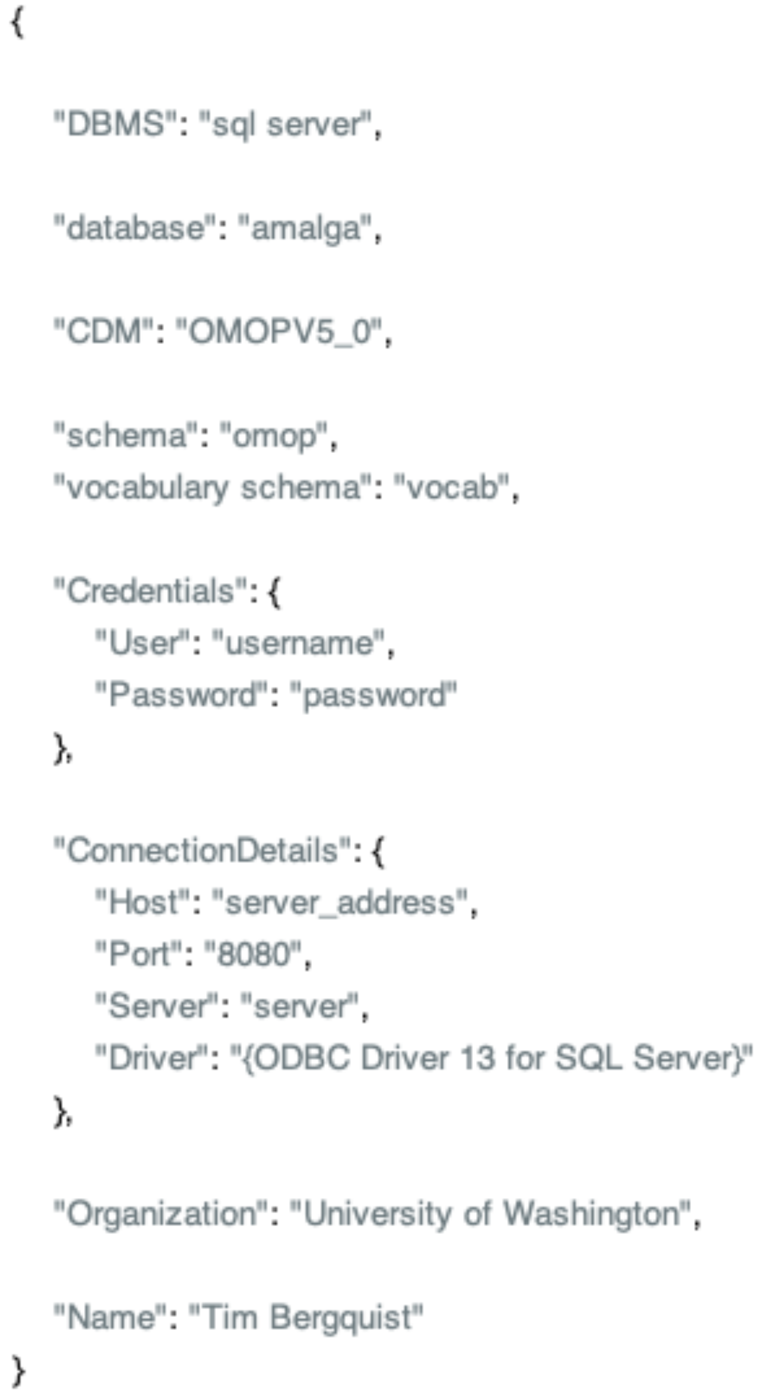
DQe-c-v2 Tool

Takes in the database credentials, CDM version, and configurations.

Simply enter your credentials and configurations into the config.json file.

Run:

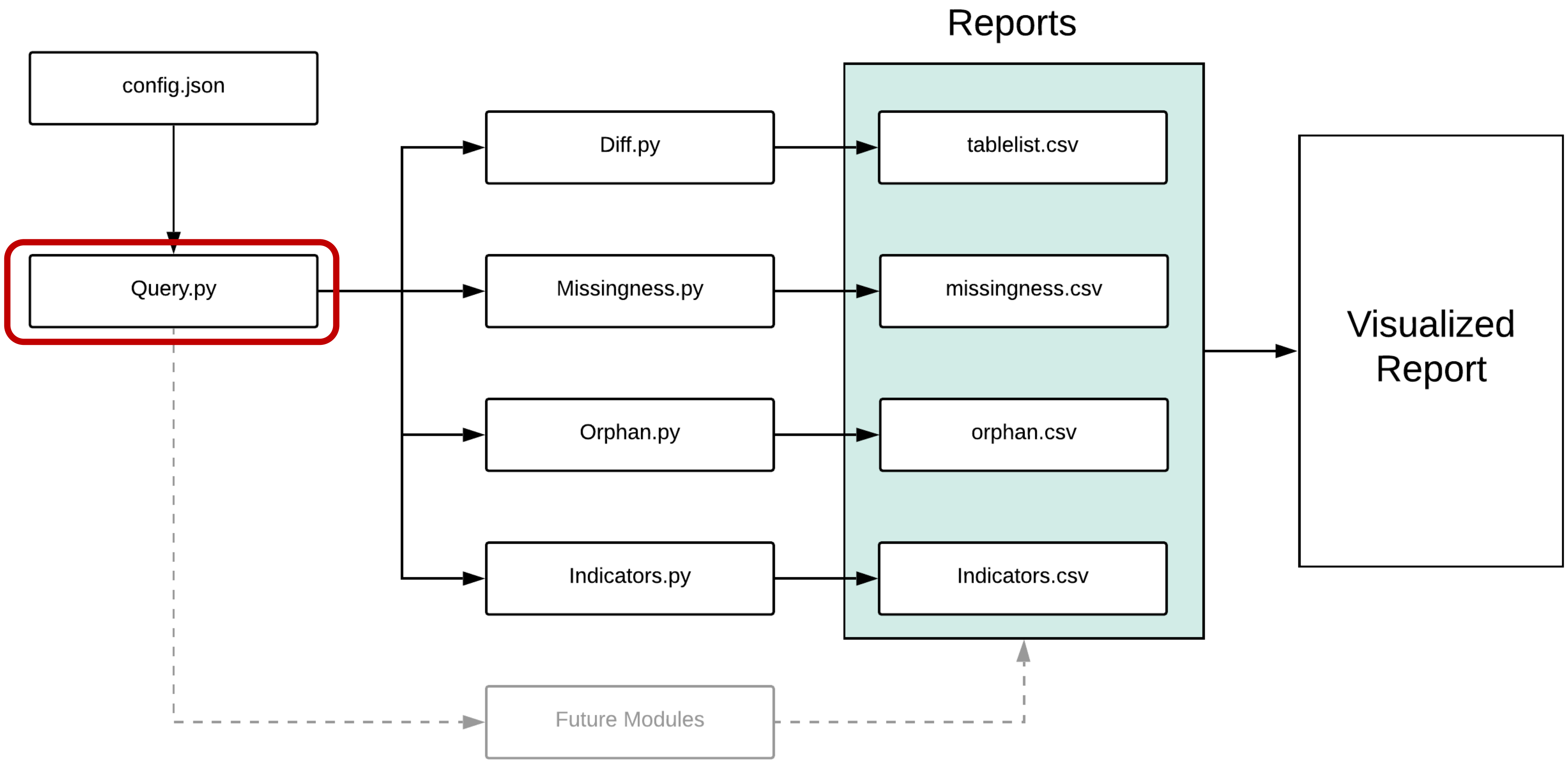
python DQe-c.py –c /path/to/config.json



```
{

    "DBMS": "sql server",

    "database": "amalga",

    "CDM": "OMOPV5_0",

    "schema": "omop",
    "vocabulary schema": "vocab",

    "Credentials": {
        "User": "username",
        "Password": "password"
    },

    "ConnectionDetails": {
        "Host": "server_address",
        "Port": "8080",
        "Server": "server",
        "Driver": "{ODBC Driver 13 for SQL Server}"
    },

    "Organization": "University of Washington",

    "Name": "Tim Bergquist"
}
```

DQe-c-v2 Tool

Sets up the database connection, manages report output, and initiates the CDM files



DQe-c-v2 Tool

Assesses conformance to a Common Data Model. Checks for missing tables and calculates size of tables.

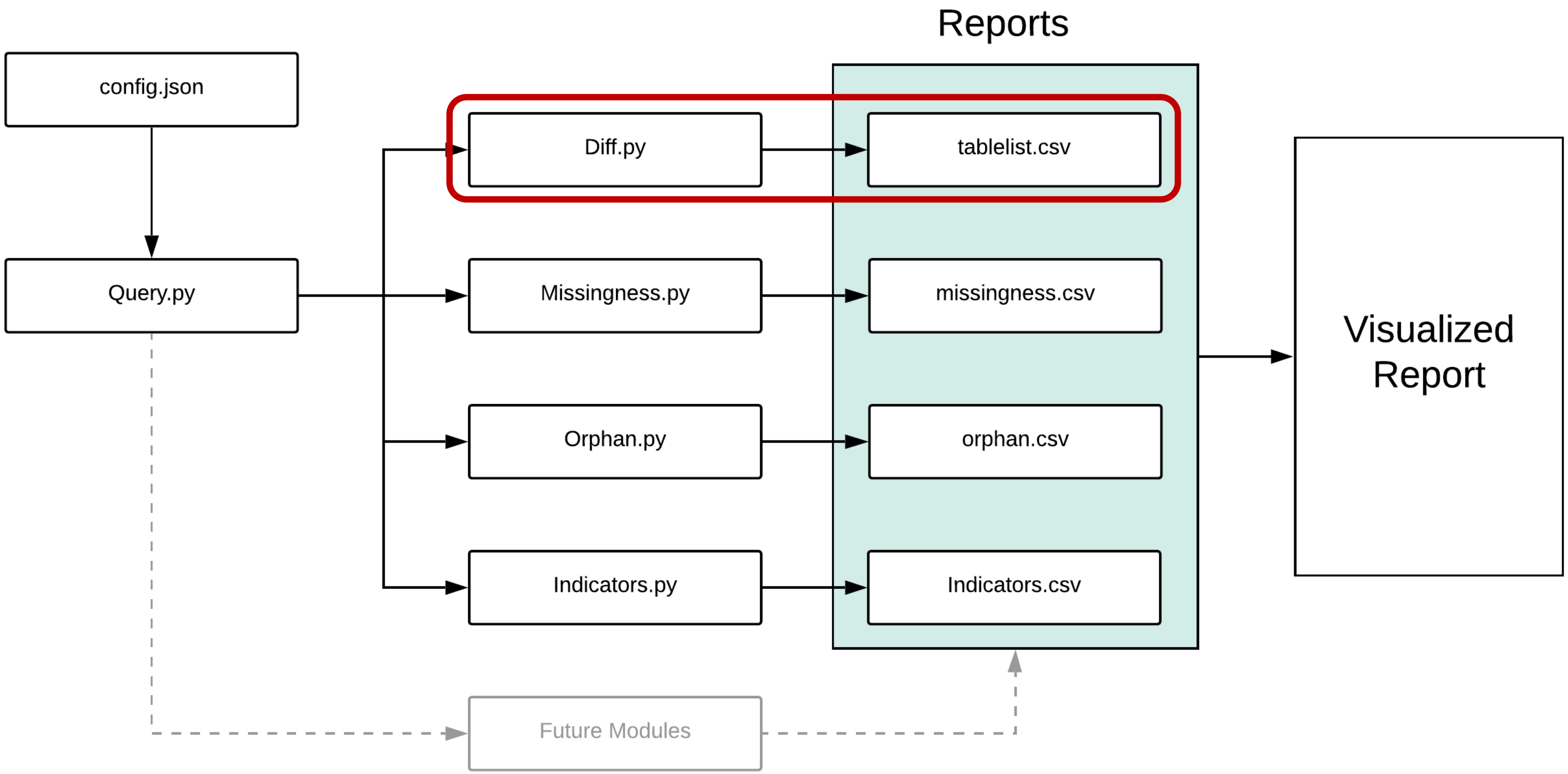


Figure 1. Available Tables, Compared to all CDM (OMOPV5_0) Tables

This figure shows which of the CDM tables are loaded and/or available.

```
## Warning in `[.data.table`(dtfDT, , `:=`("c", fact), with = FALSE):
## with=FALSE ignored, it isn't needed when using :=. See ?':=' for examples.
```

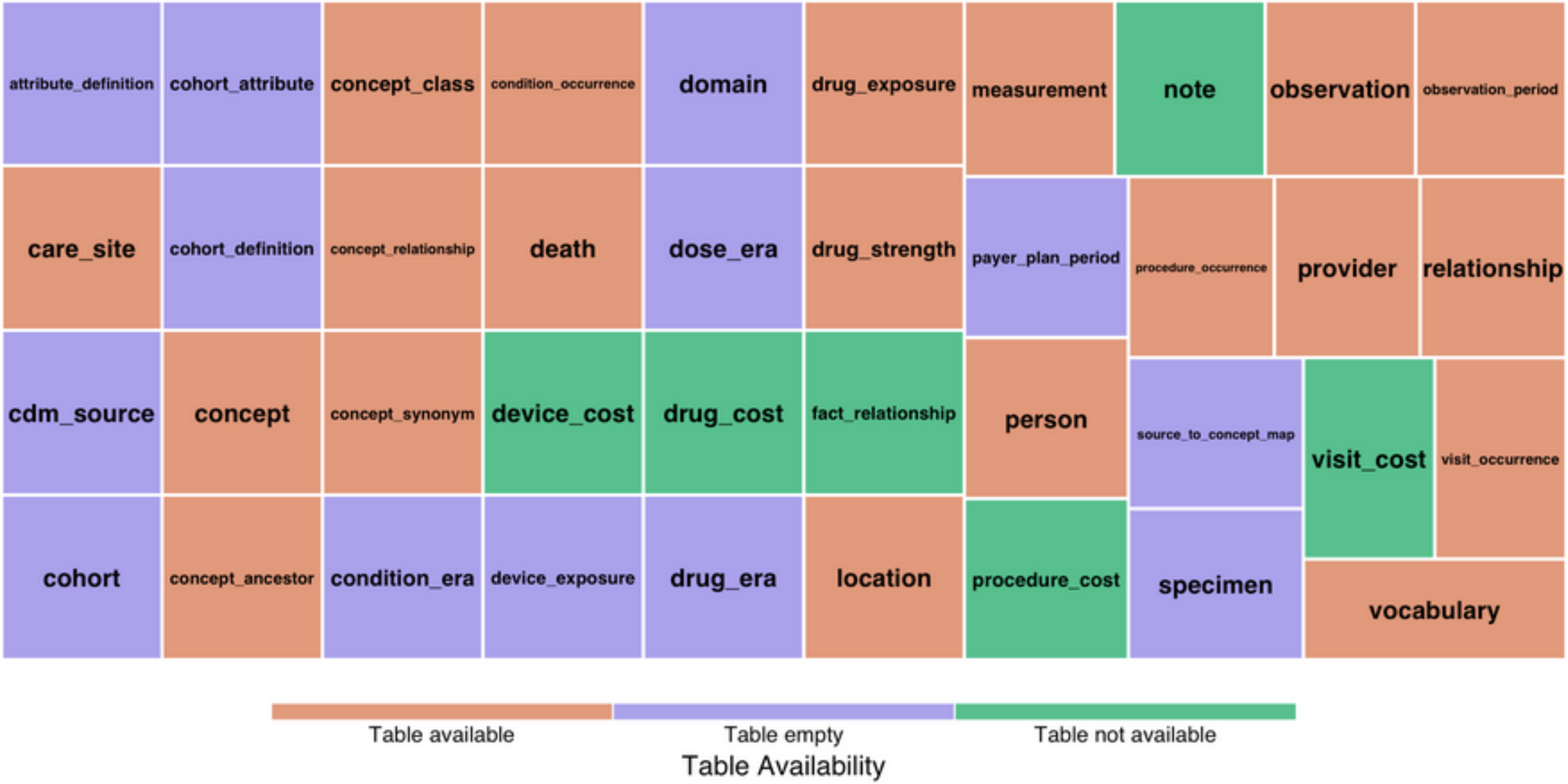
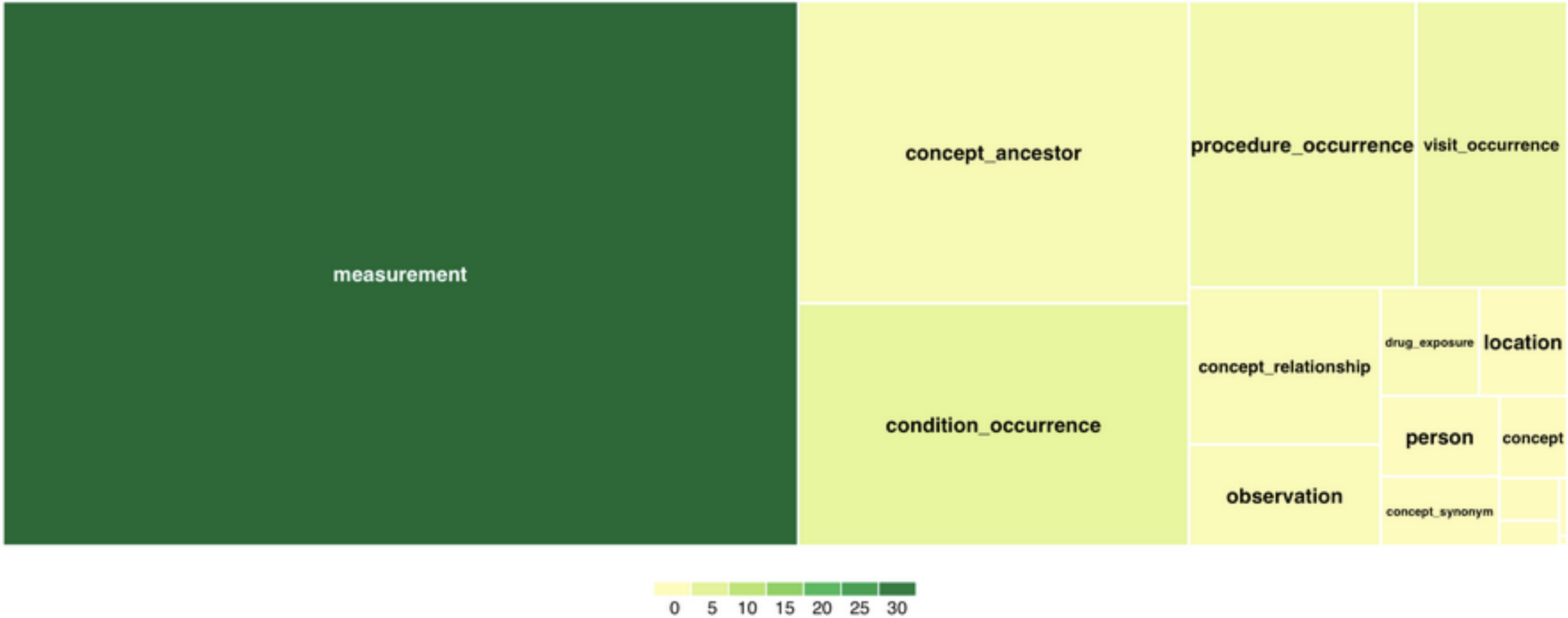
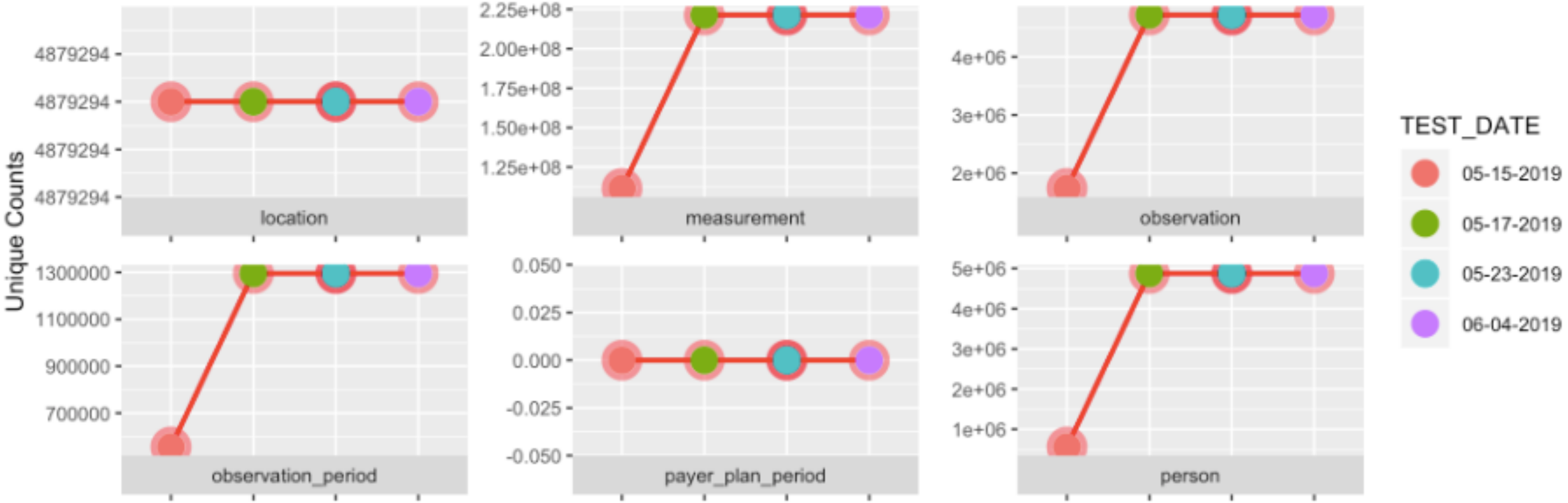


Figure 2. File Size and Row Numbers by Table in the (OMOPV5_0) Load



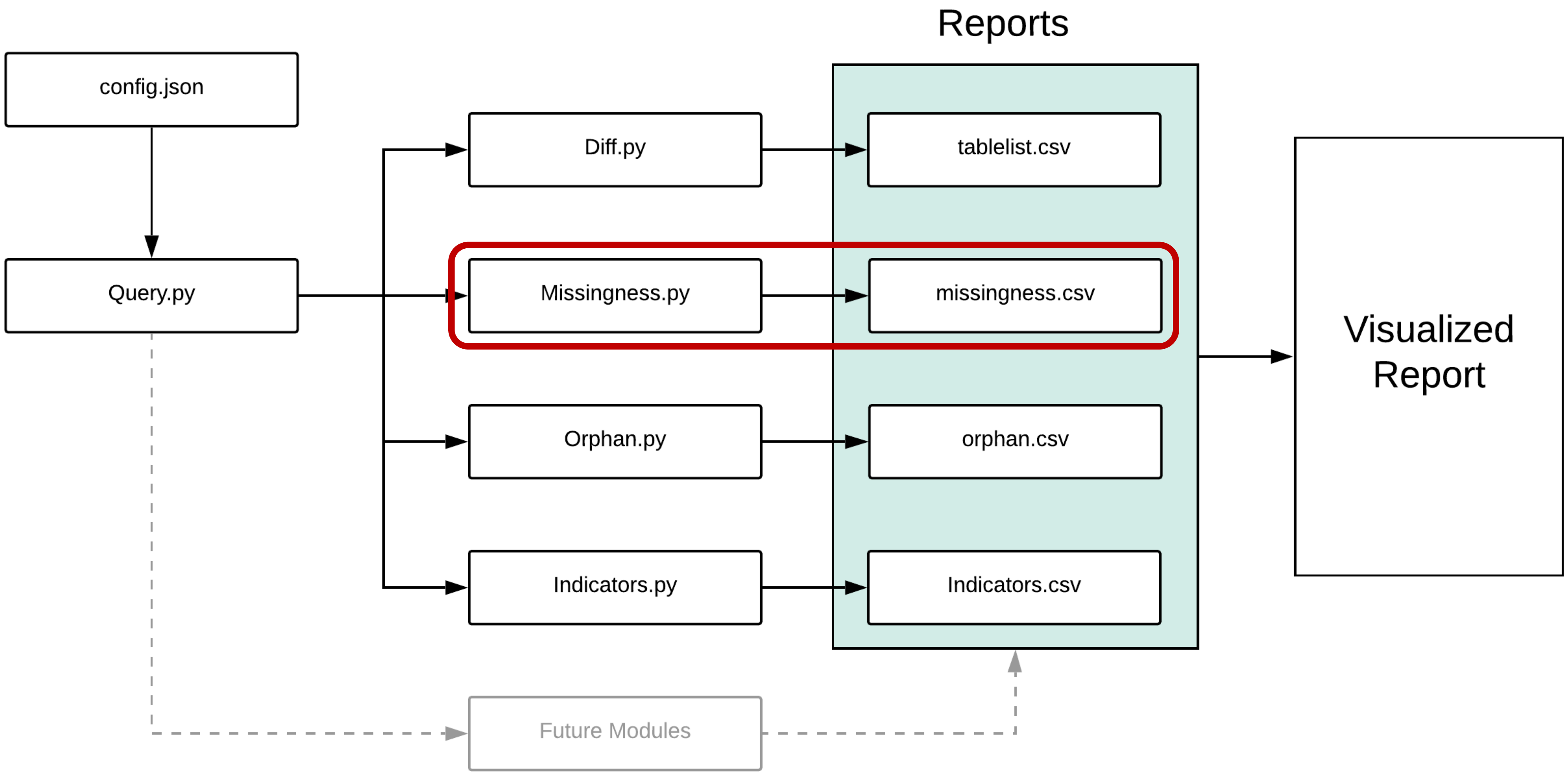
Size represents number of rows and color represent file size (in GB) for each table.

Quickly check that the new data is growing as expected

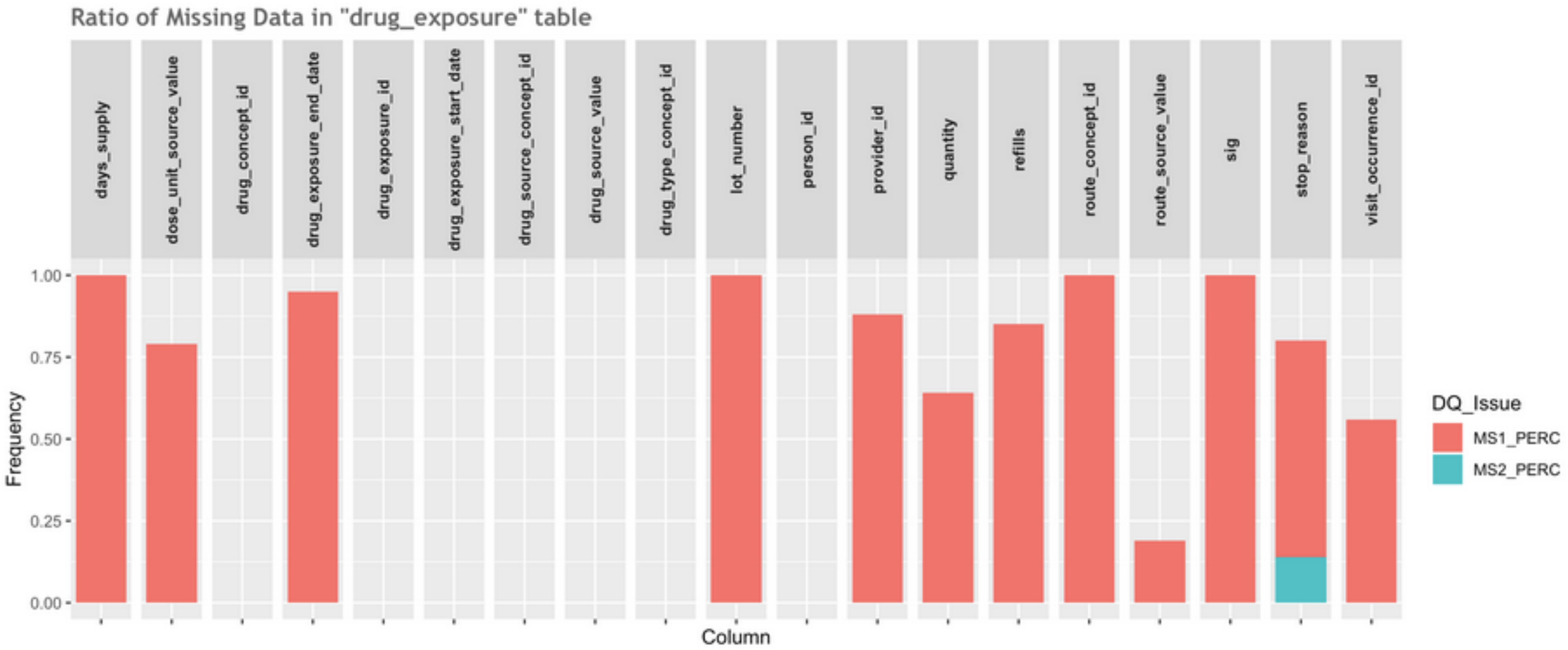


DQe-c-v2 Tool

Assesses **completeness** of all columns in the available tables in the database.
Checks for null and nonsense values.

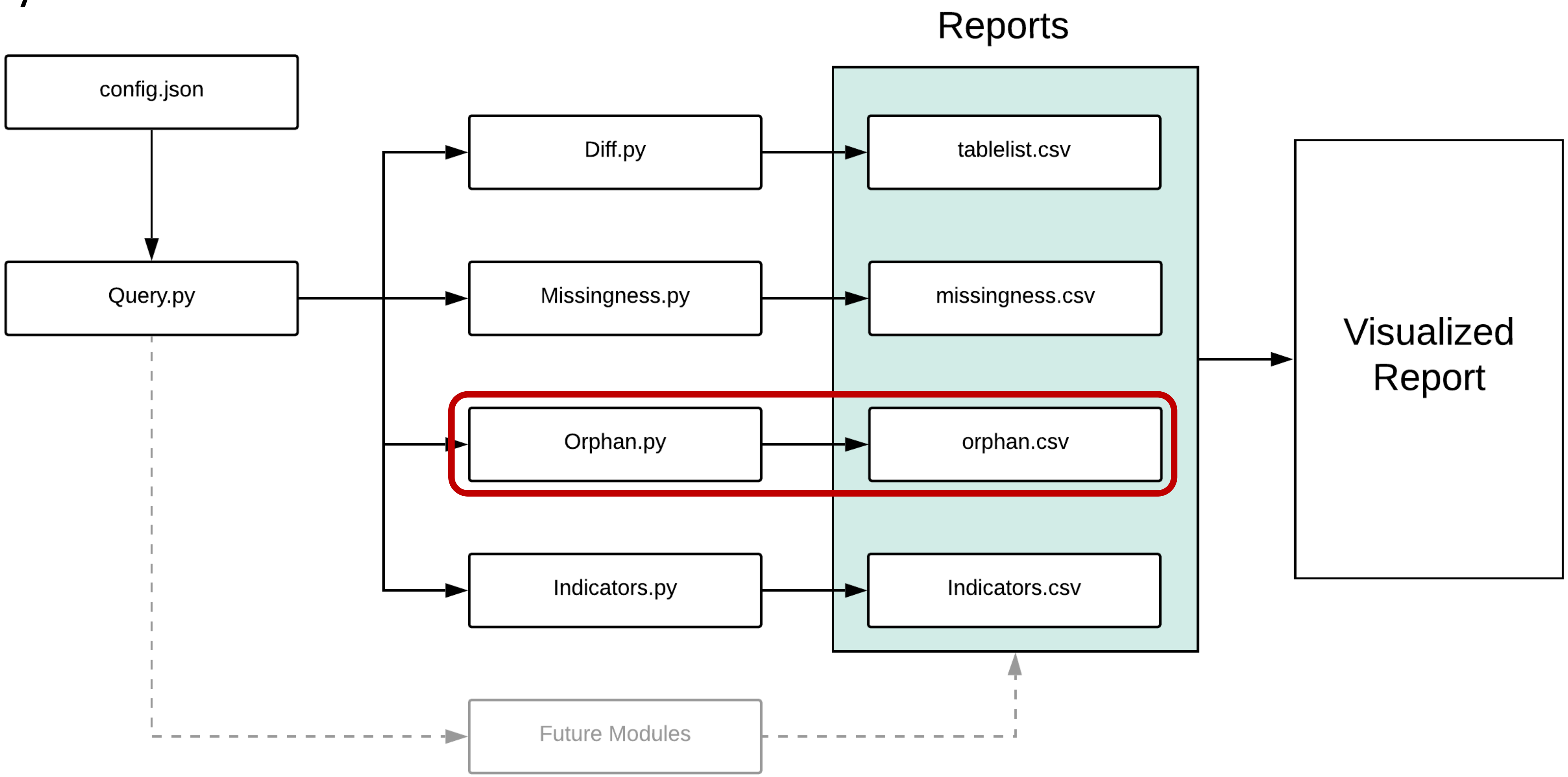


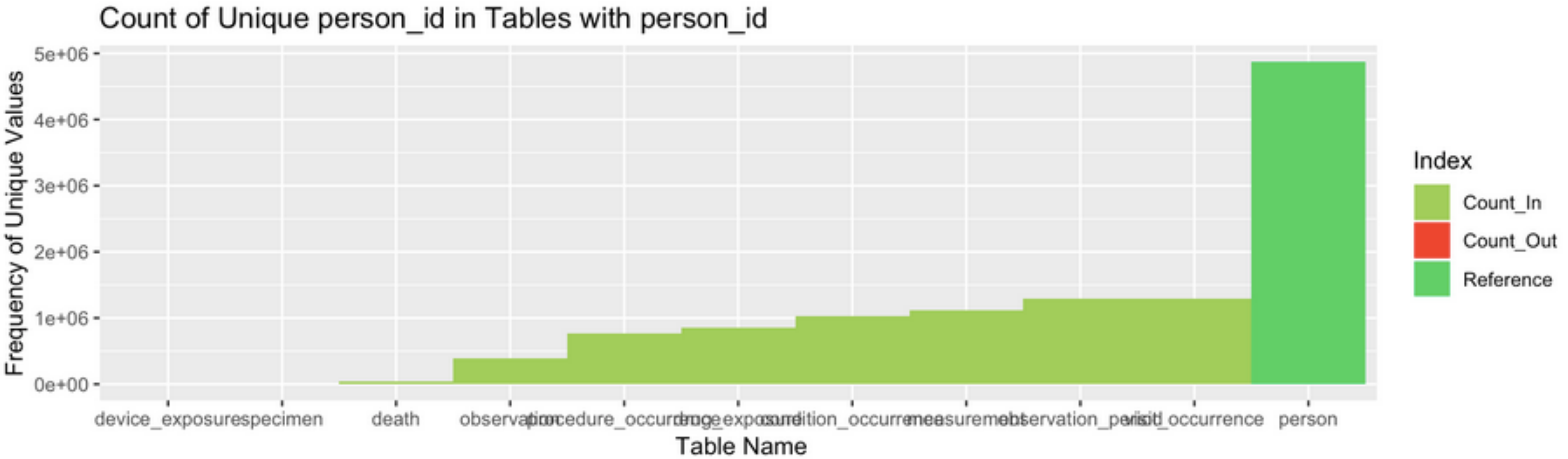
Identify empty or useful columns in each of your OMOP tables.

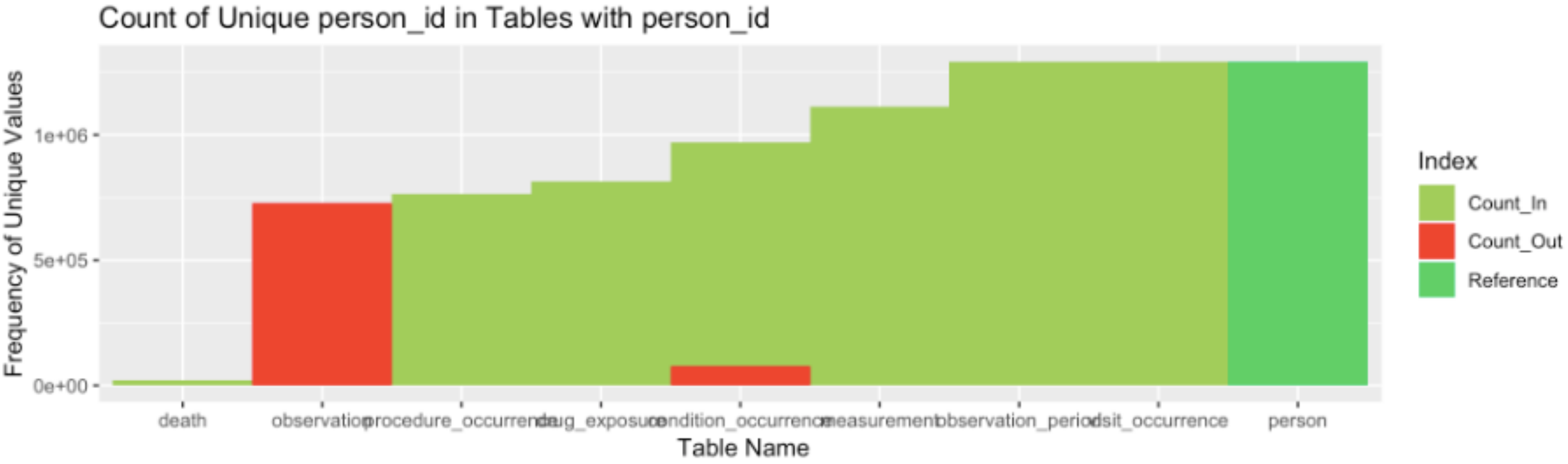


DQe-c-v2 Tool

Checks for orphan keys, foreign keys not present in the primary table.




Count of Unique person_id in Tables with person_id
Frequency of Unique Values
5e+06
4e+06
3e+06
2e+06
1e+06
0e+00
device_exposure
specimen
death
observation
procedure_occurrence
drug_exposure
condition_occurrence
measurement
observation_period
visit_occurrence
person
Table Name
Index
Count_In
Count_Out
Reference


Count of Unique person_id in Tables with person_id
Frequency of Unique Values
1e+06
5e+05
0e+00
death
observation
procedure_occurrence
drug_exposure
condition_occurrence
measurement
observation_period
visit_occurrence
person
Table Name
Index
Count_In
Count_Out
Reference

DQe-c-v2 Tool

Checks for missingness in clinical indicators. (What percent of patients have a heart rate measure, blood pressure measurement, etc.)

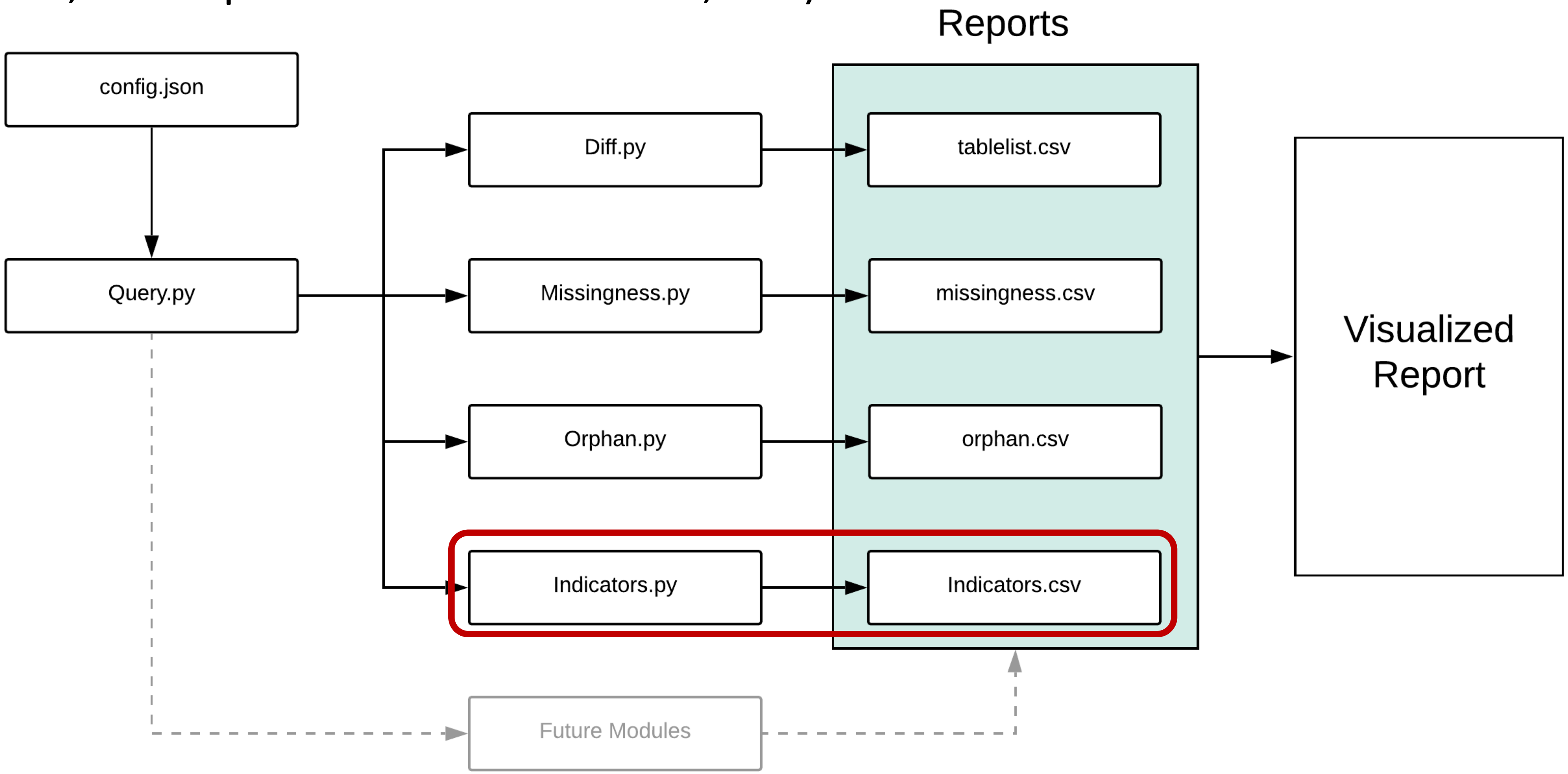
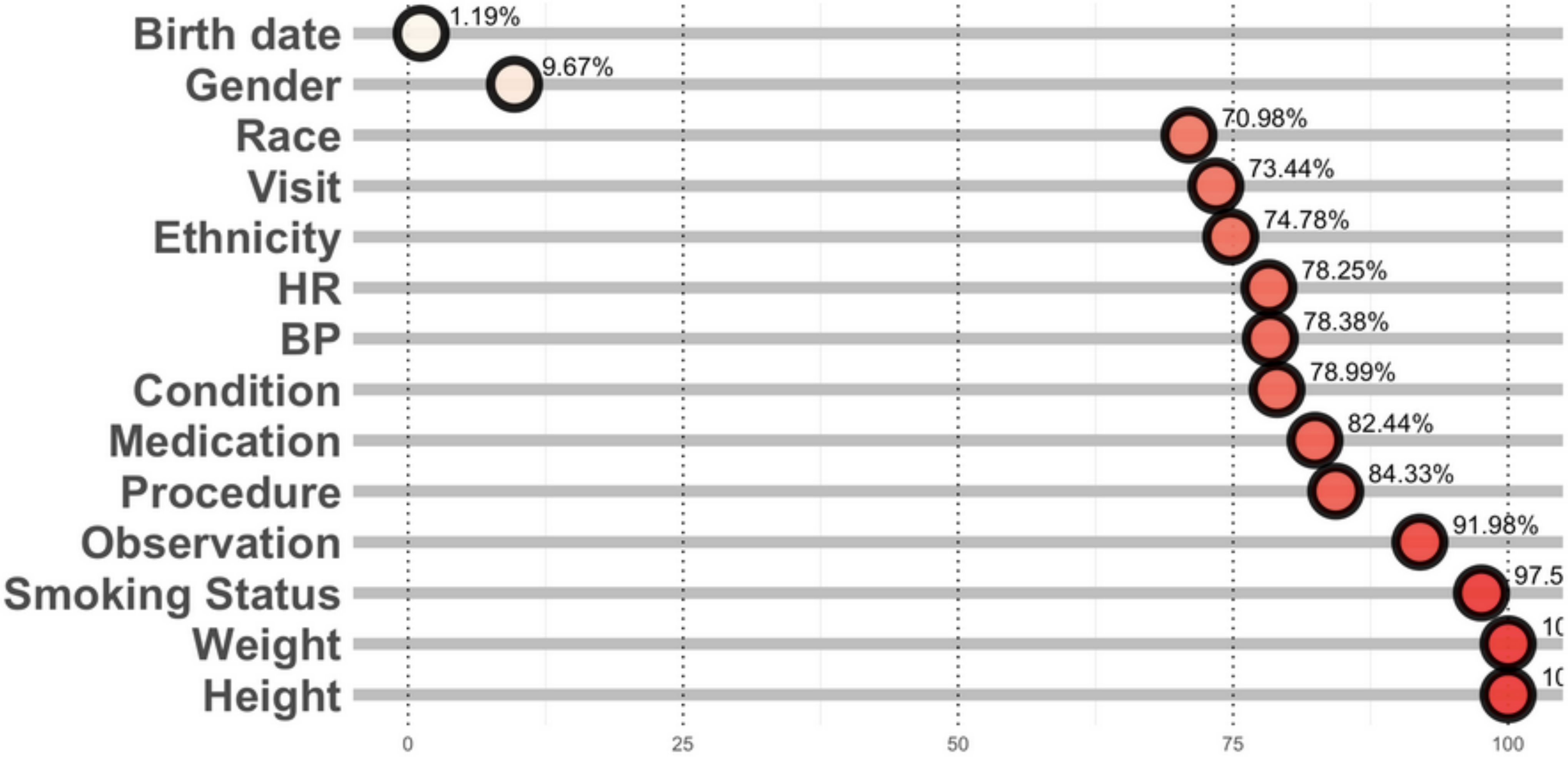


Figure 5. Common Key Variables

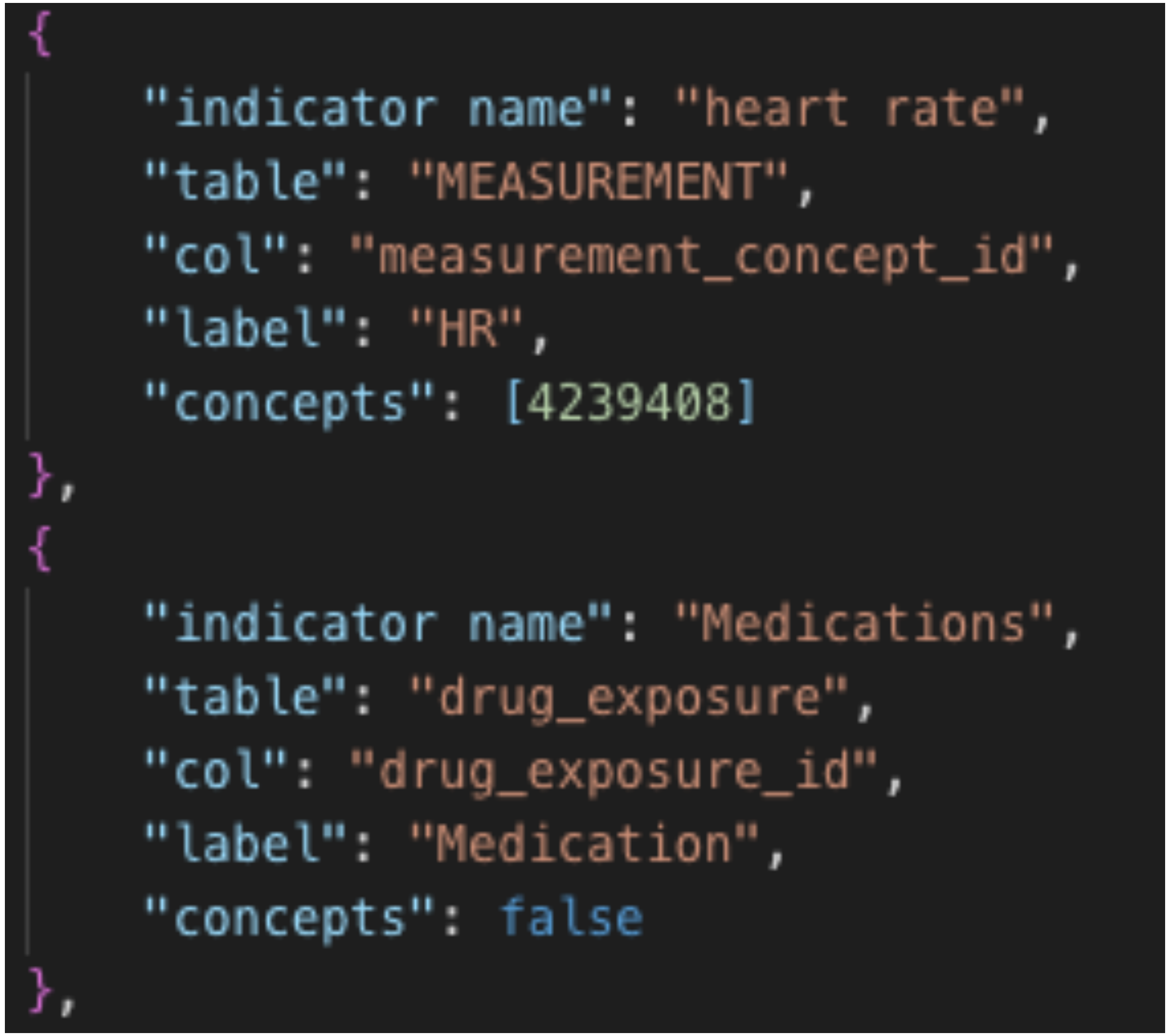
Figure 5 shows the percentage of patients missing specific key clinical indicators.



Adding a new indicator test is straight forward!

Completion as the presence of a concept. Calculates what percentage of patients have the identified concept(s).

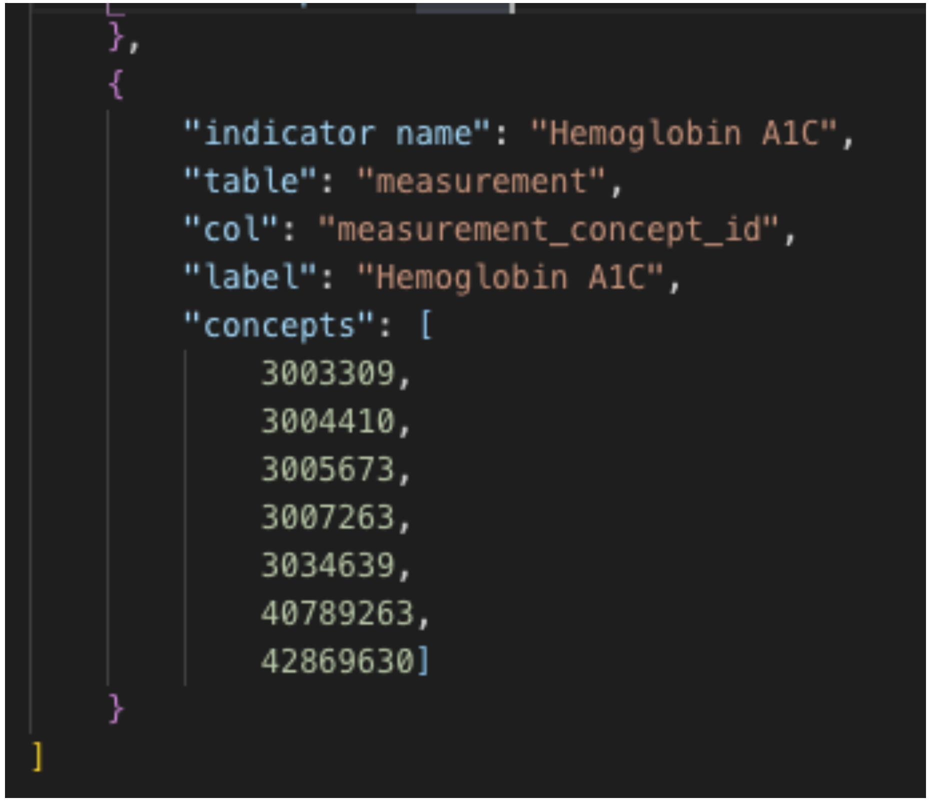
Completion as the presence of a non-null. Calculates what percentage of patients have a non-null value in the identified table-column.



```
{
    "indicator name": "heart rate",
    "table": "MEASUREMENT",
    "col": "measurement_concept_id",
    "label": "HR",
    "concepts": [4239408]
},
{
    "indicator name": "Medications",
    "table": "drug_exposure",
    "col": "drug_exposure_id",
    "label": "Medication",
    "concepts": false
},
```

We can add a new indictor test by just adding five new fields.

Adding testing for A1C Hemoglobin.
Calculates what percentage of patients have a hemoglobin A1C measurement.

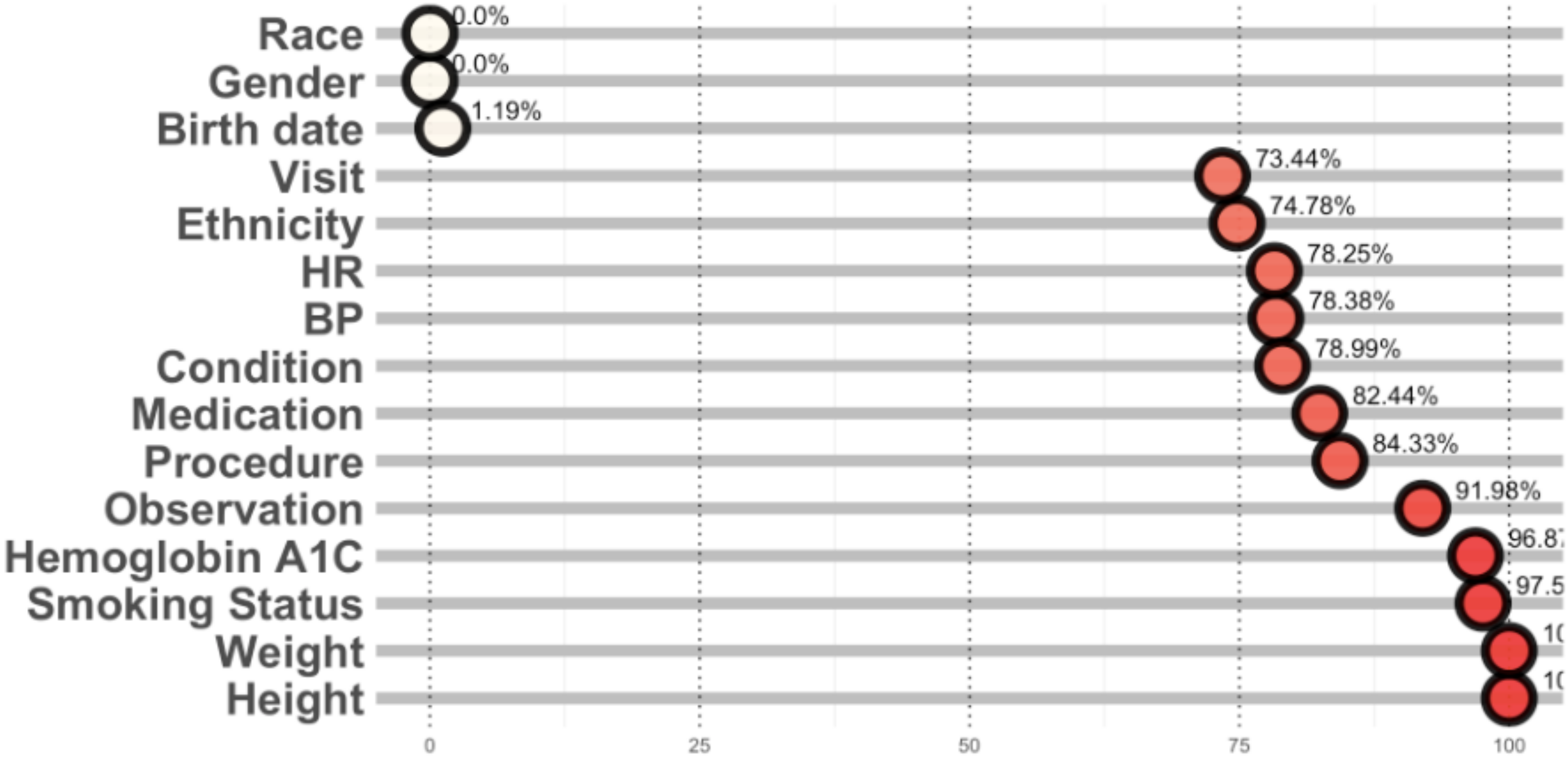


```
    },
    {
        "indicator name": "Hemoglobin A1C",
        "table": "measurement",
        "col": "measurement_concept_id",
        "label": "Hemoglobin A1C",
        "concepts": [
            3003309,
            3004410,
            3005673,
            3007263,
            3034639,
            40789263,
            42869630]
    }
]
```

Test of Completeness in Key Clinical Indicators

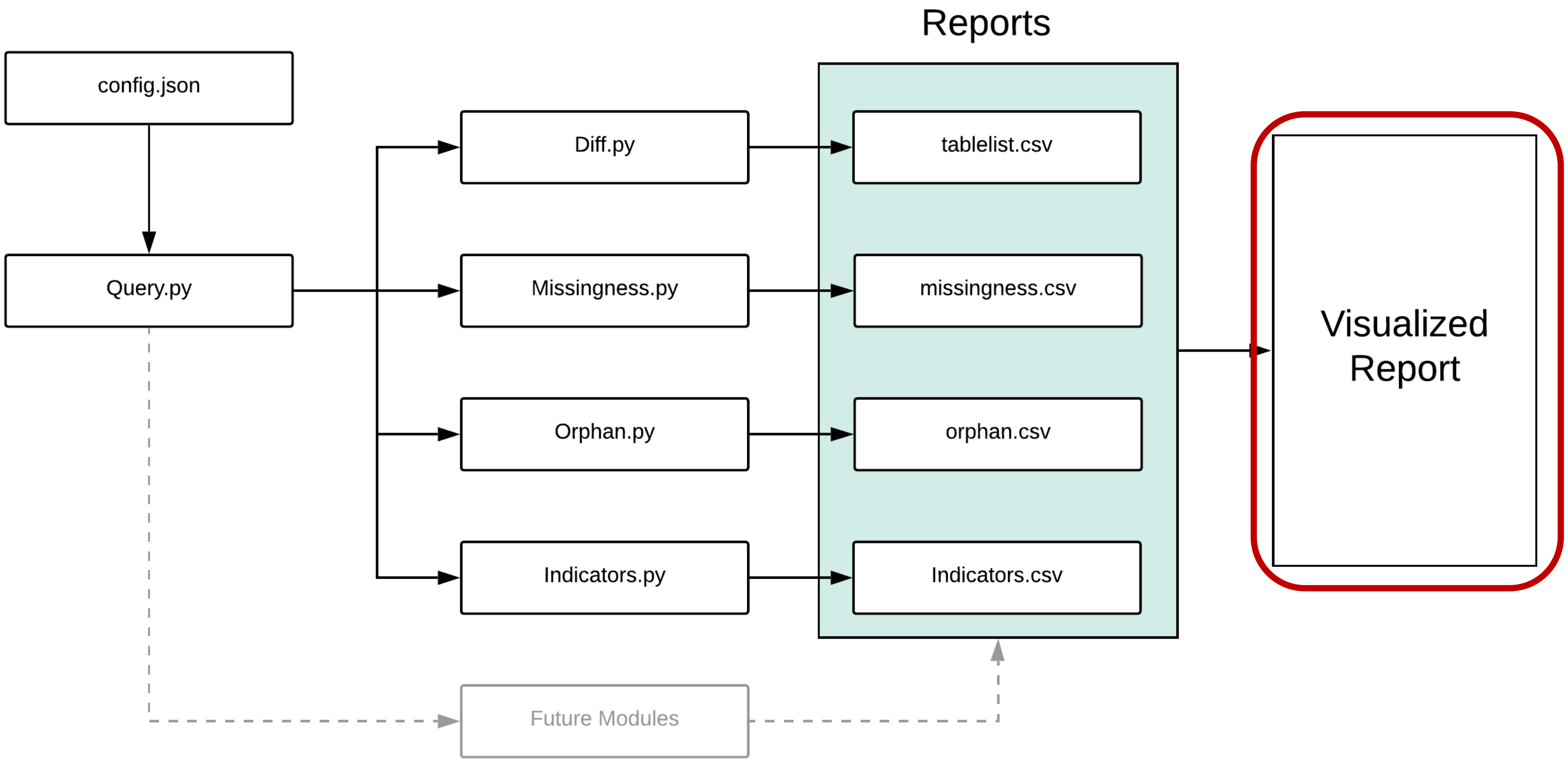
Figure 5. Common Key Variables

Figure 5 shows the percentage of patients missing specific key clinical indicators.



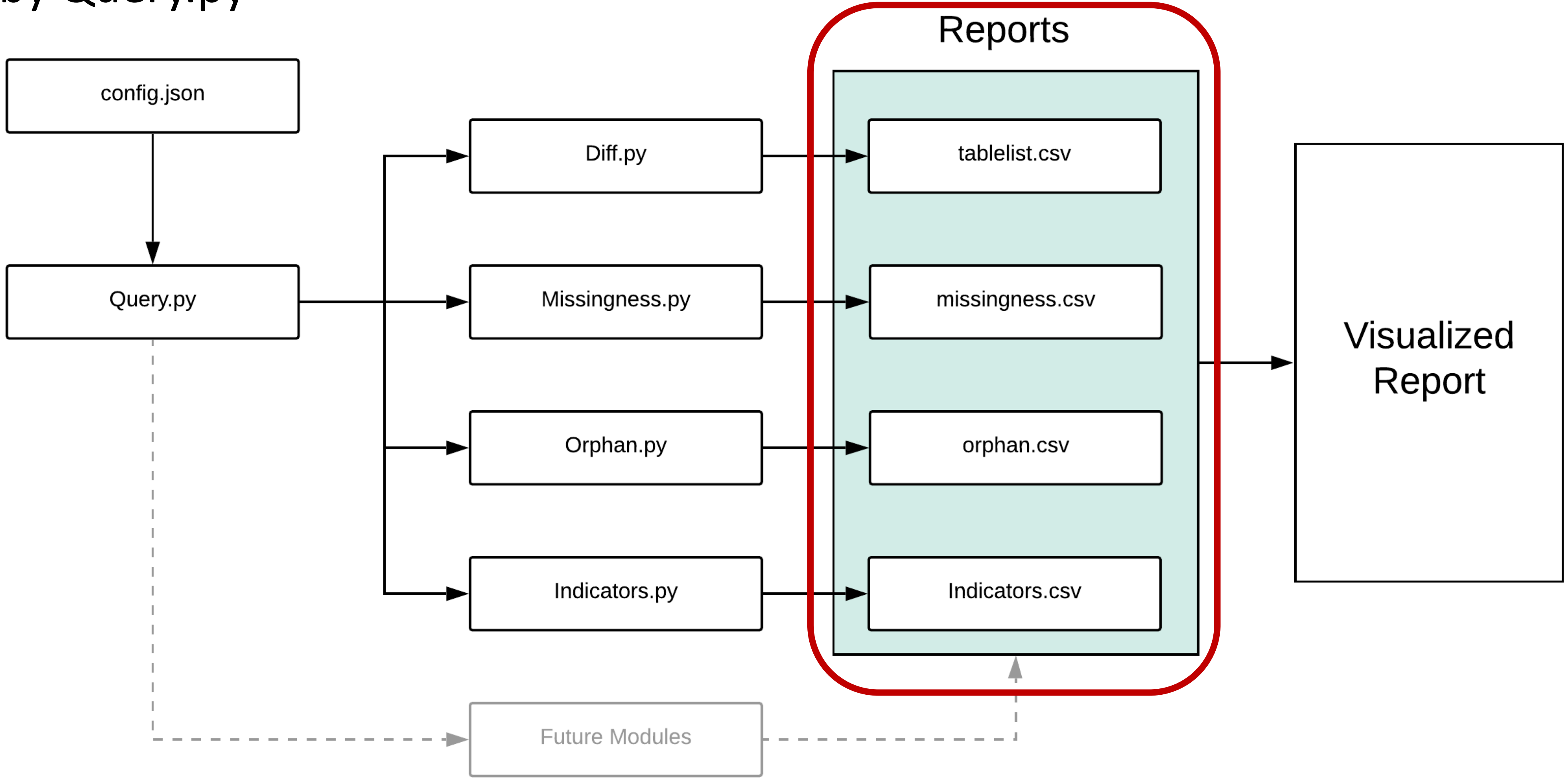
DQe-c-v2 Tool

All reports are combined into a visualization dashboard



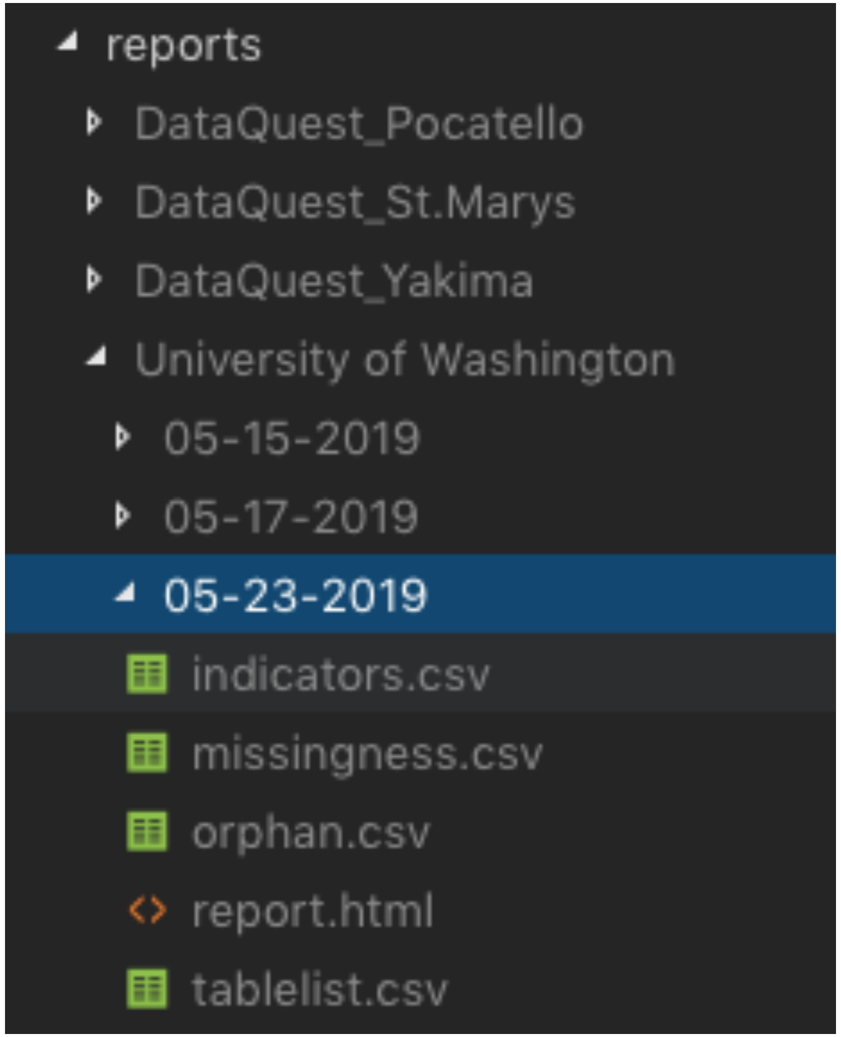
DQe-c-v2 Tool

All these modules output csv reports. The output folders are managed by Query.py



DQe-c-v2 Tool

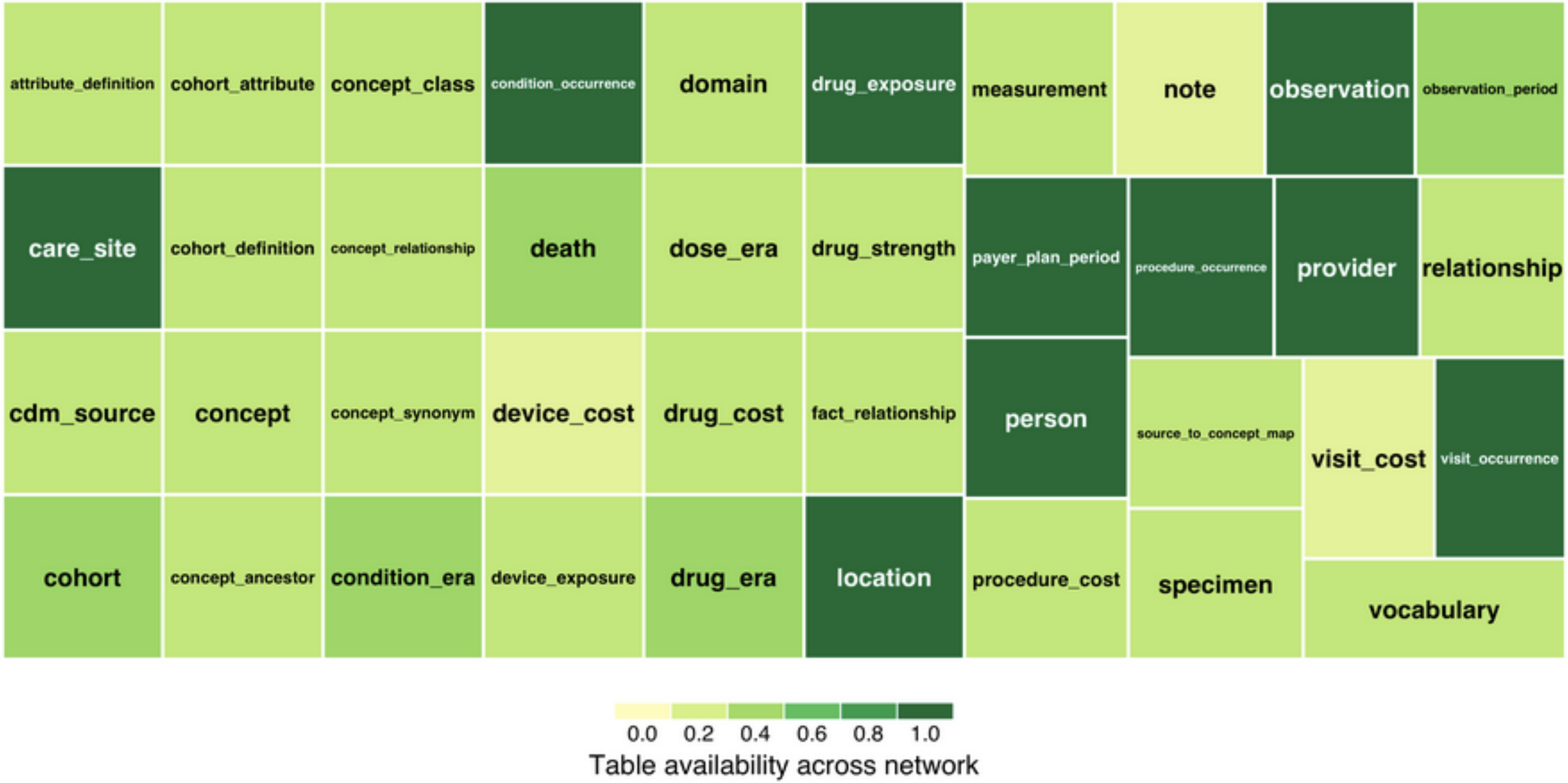
All these modules output csv reports. The output folders are managed by Query.py to account for different test dates and organizations.



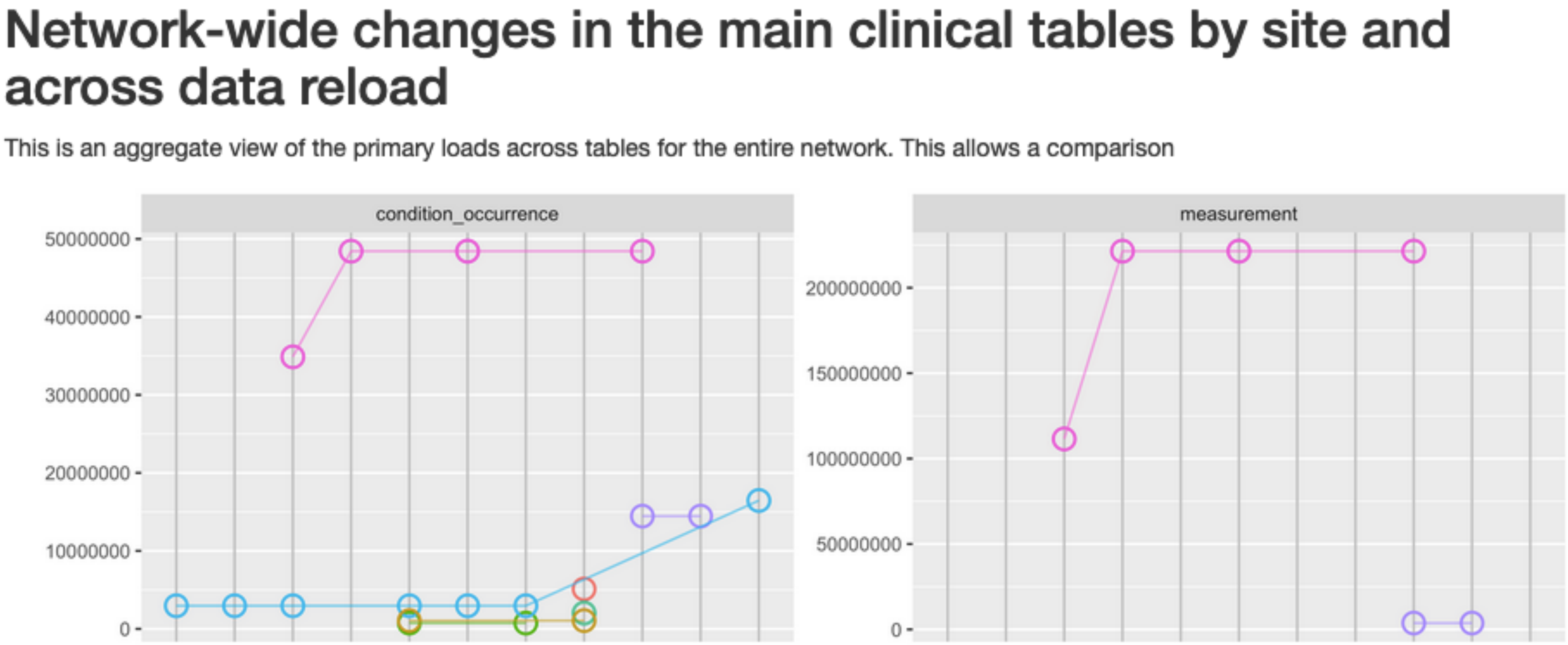
```
◢ reports
  ▸ DataQuest_Pocatello
  ▸ DataQuest_St.Marys
  ▸ DataQuest_Yakima
  ◢ University of Washington
    ▸ 05-15-2019
    ▸ 05-17-2019
    ◢ 05-23-2019
      indicators.csv
      missingness.csv
      orphan.csv
      report.html
      tablelist.csv
```

DQe-c-v2 Network Aggregation Tool

A network-level preview of table availability



DQe-c-v2 Network Aggregation Tool

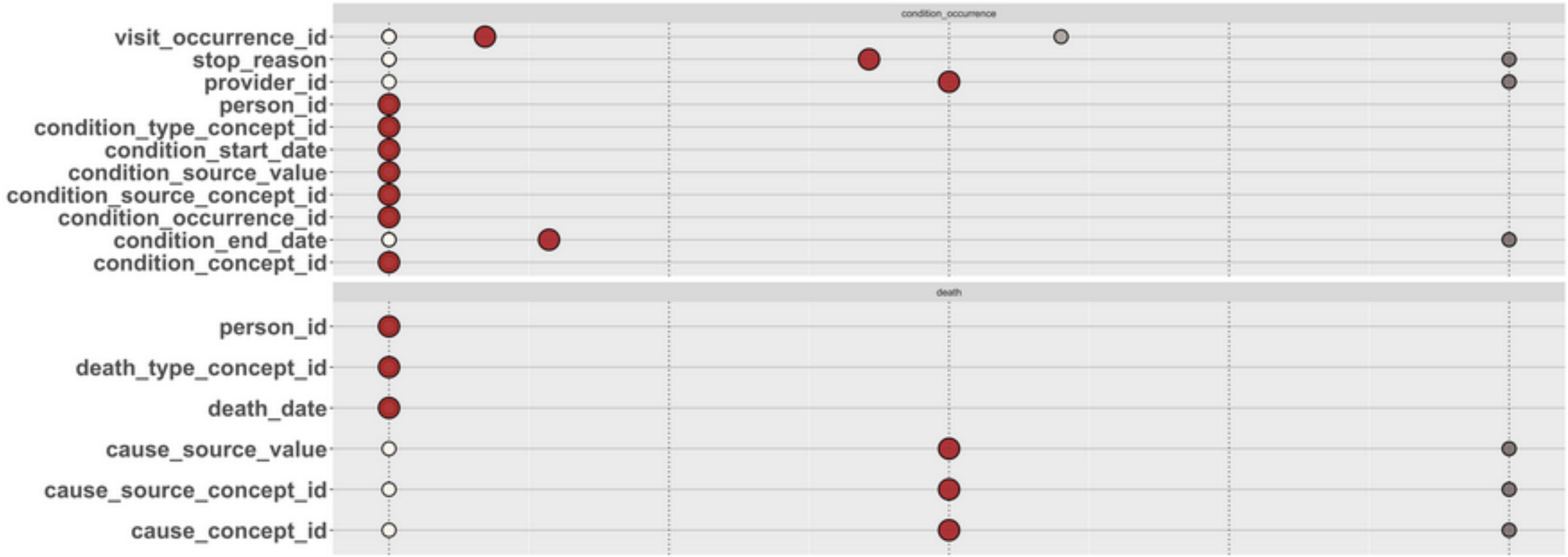


Network-wide changes in the main clinical tables by site and across data reload

This is an aggregate view of the primary loads across tables for the entire network. This allows a comparison

DQe-c-v2 Network Aggregation Tool

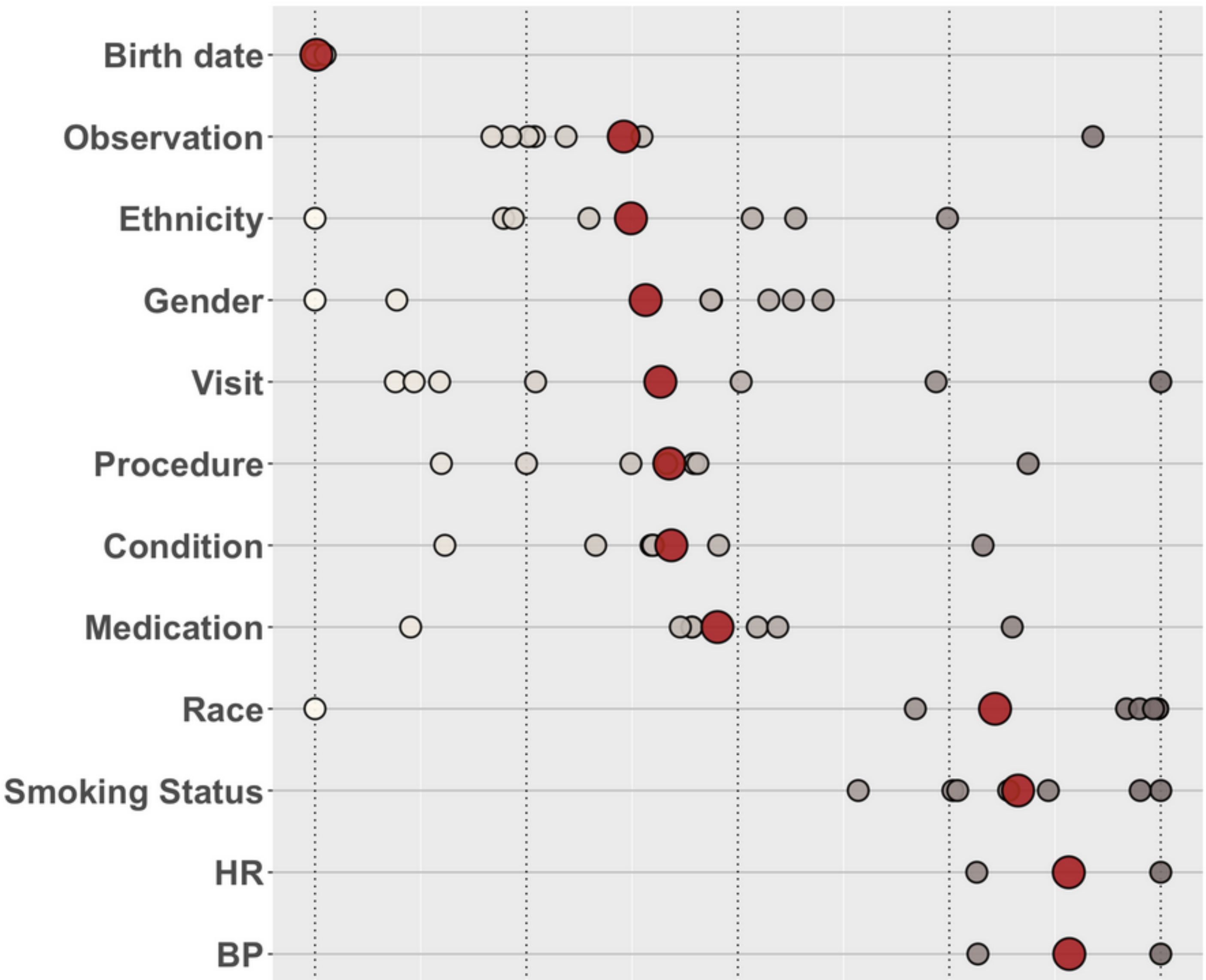
Network-wide missingness in available tables.



DQe-c-v2 Network Aggregation Tool

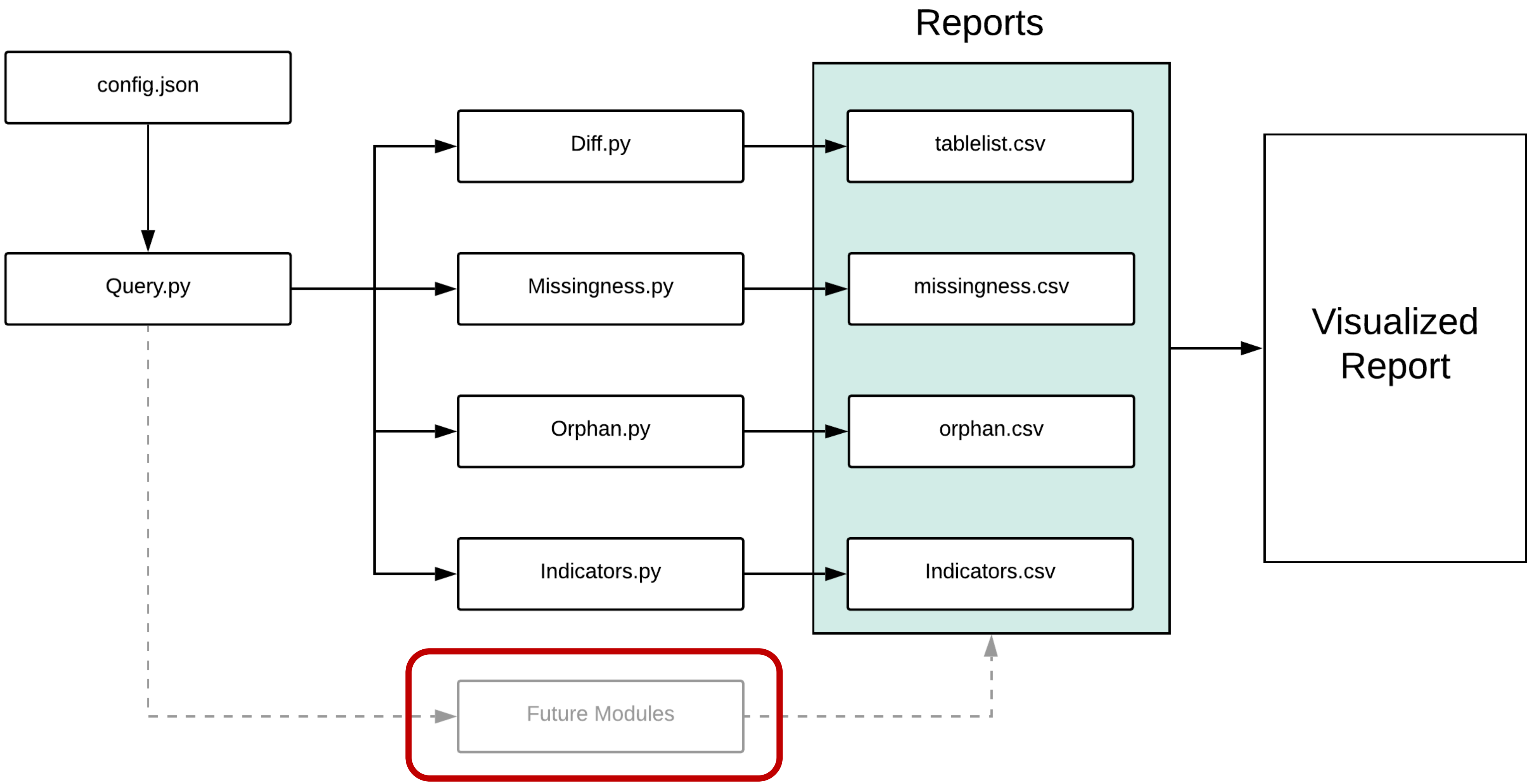
Figure 9. Indicator differences across the network

Figure 9 shows the different indicator measurements from across the network.

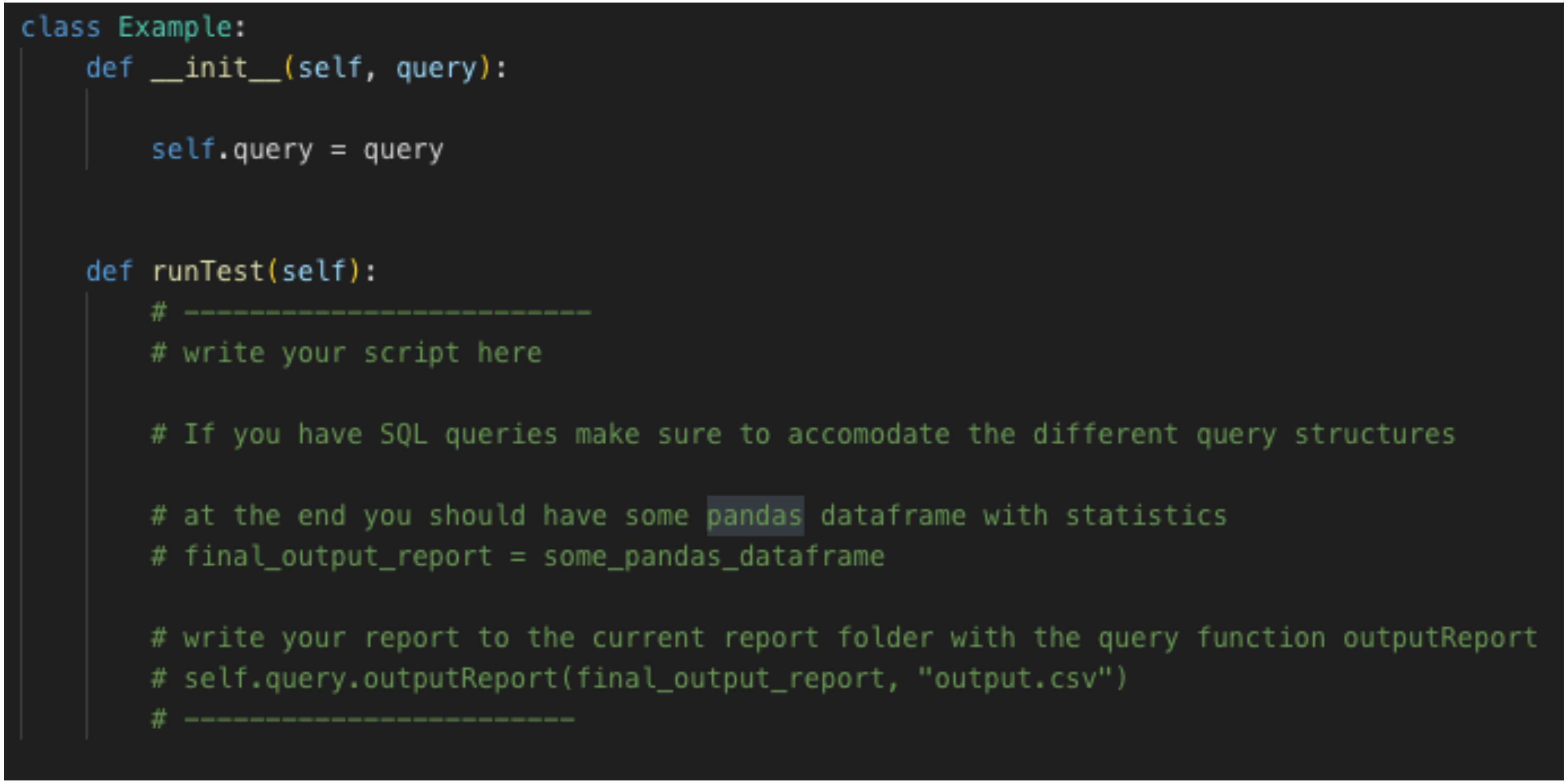


DQe-c-v2 Tool

Reports are visualized into an HTML file. Easy to embed into a website



Adding New Modules



```
class Example:
    def __init__(self, query):

        self.query = query


    def runTest(self):
        # -------------------------
        # write your script here

        # If you have SQL queries make sure to accomodate the different query structures

        # at the end you should have some pandas dataframe with statistics
        # final_output_report = some_pandas_dataframe

        # write your report to the current report folder with the query function outputReport
        # self.query.outputReport(final_output_report, "output.csv")
        # ------------------------
```

Vocabulary Summary

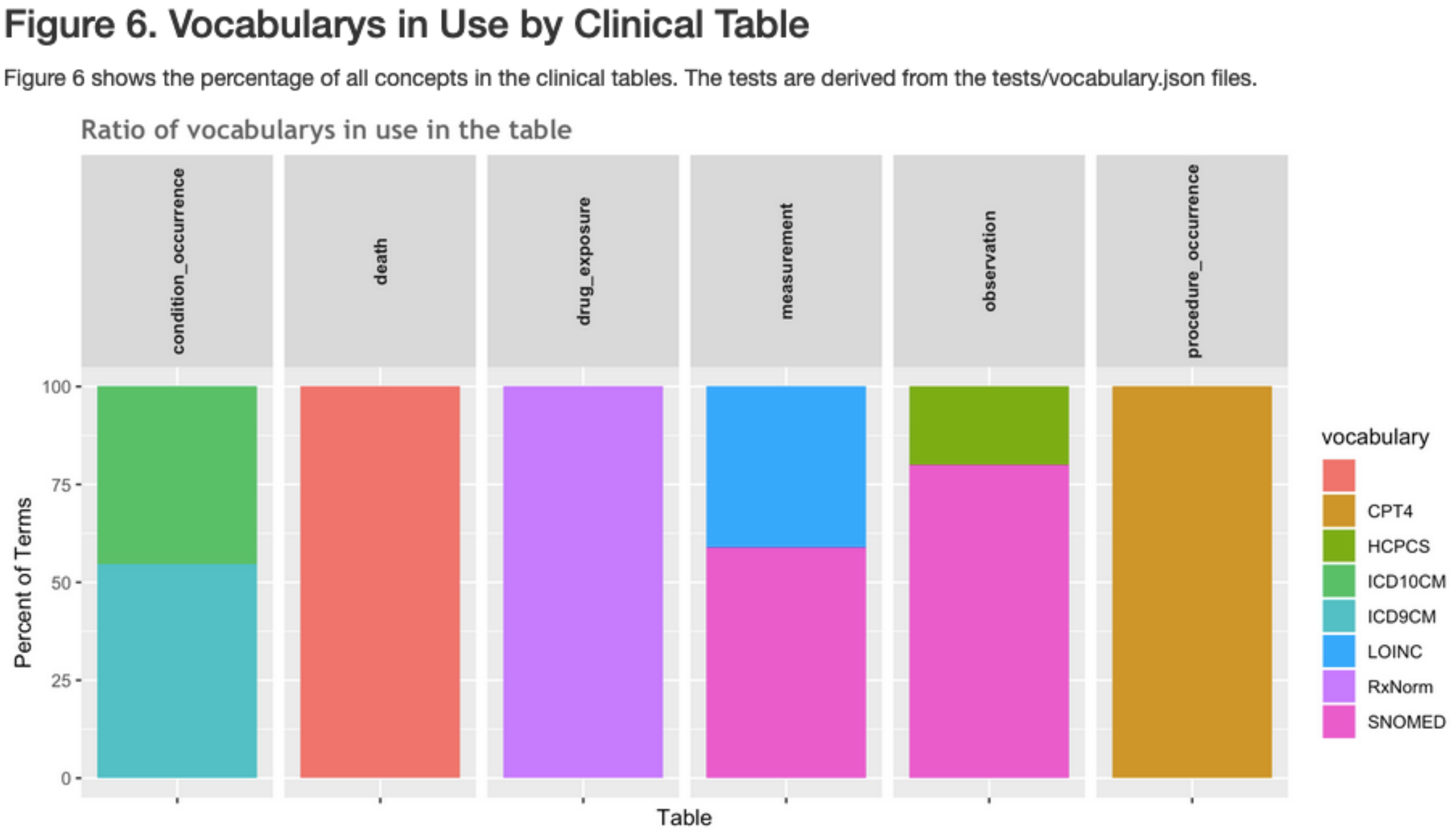
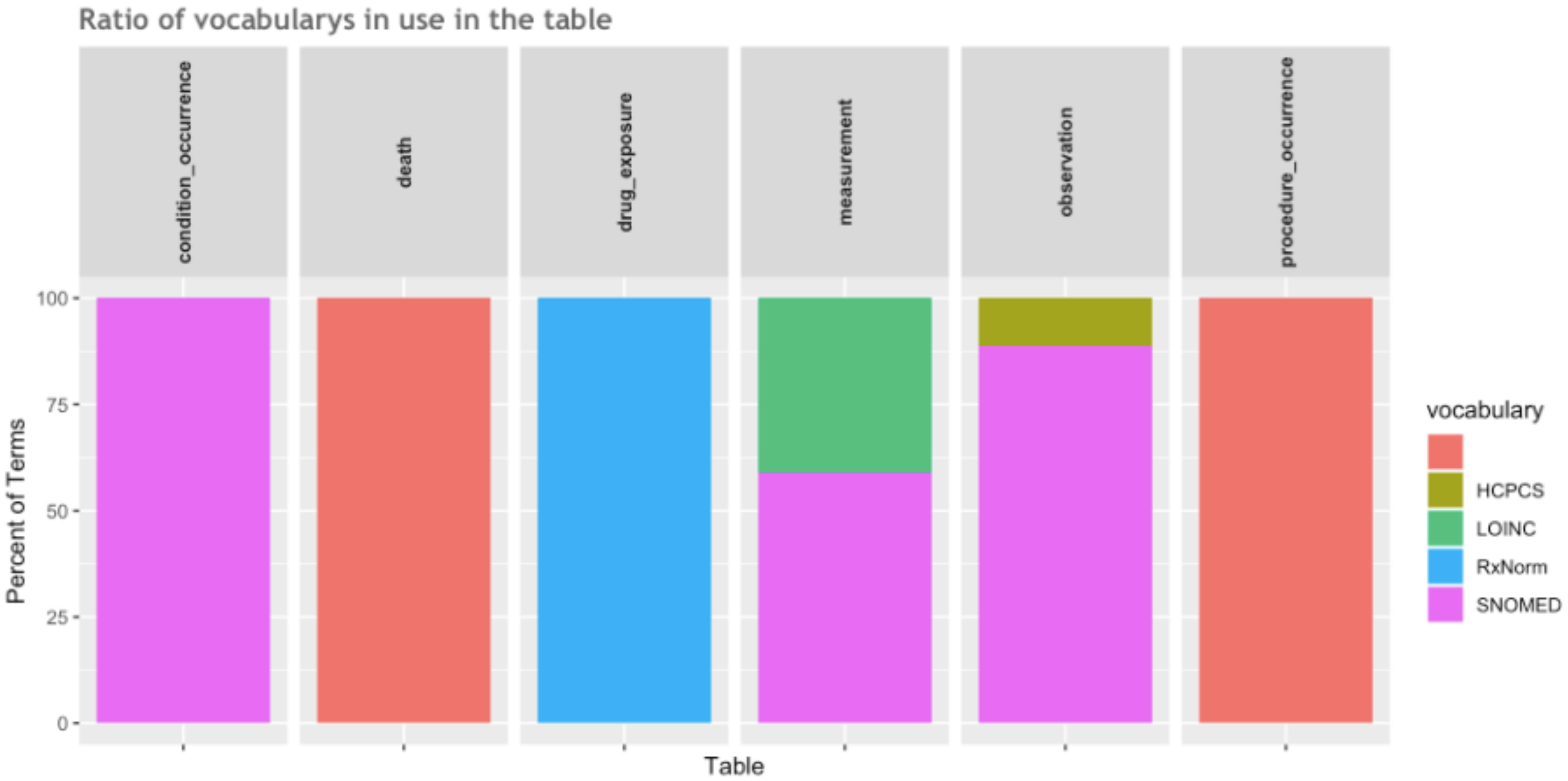


Figure 6. Vocabularys in Use by Clinical Table

Figure 6 shows the percentage of all concepts in the clinical tables. The tests are derived from the tests/vocabulary.json files.

Vocabulary Summary



Vocabulary Summary

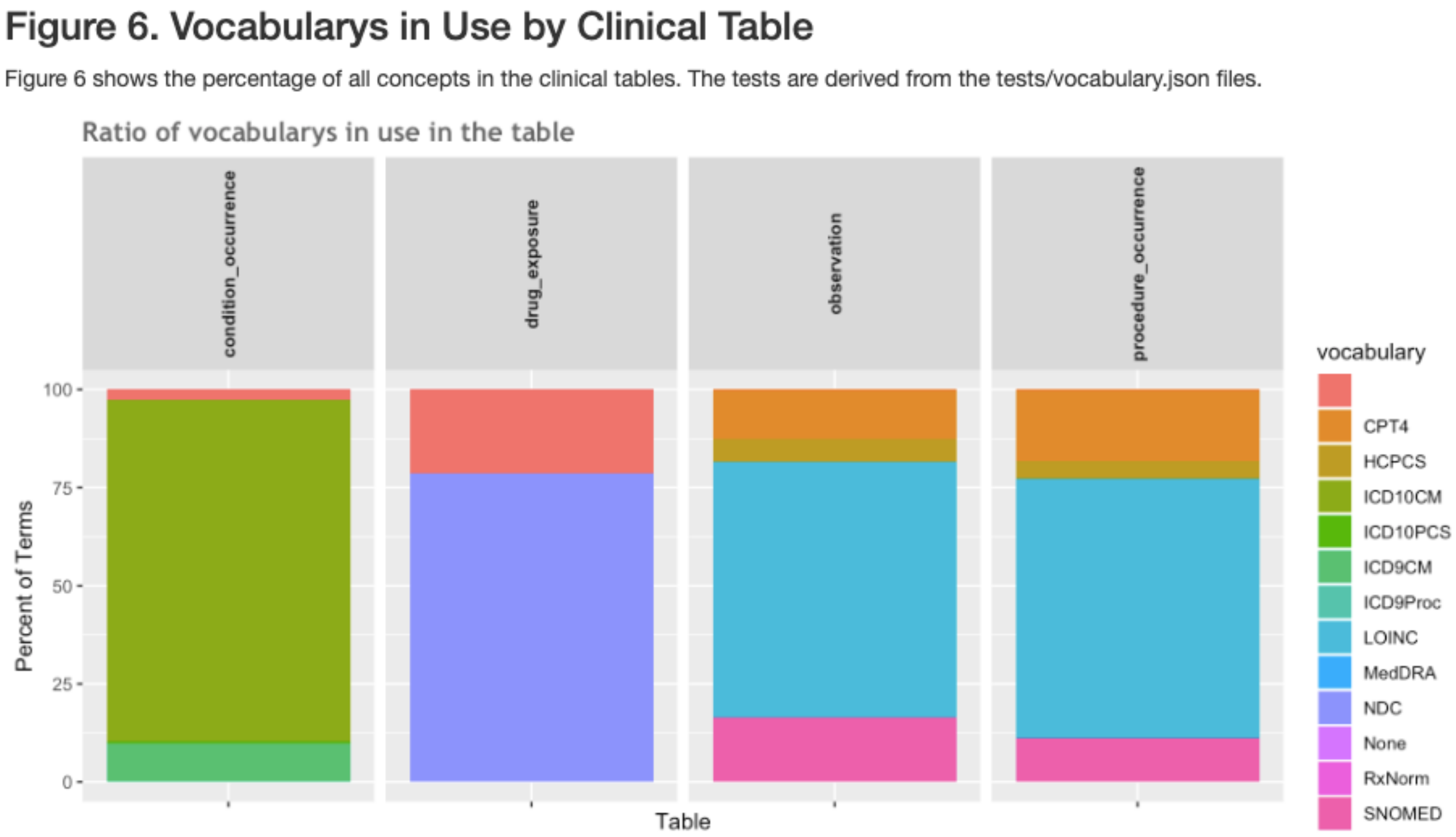


Figure 6. Vocabularys in Use by Clinical Table

Figure 6 shows the percentage of all concepts in the clinical tables. The tests are derived from the tests/vocabulary.json files.

Temporal Plausibility

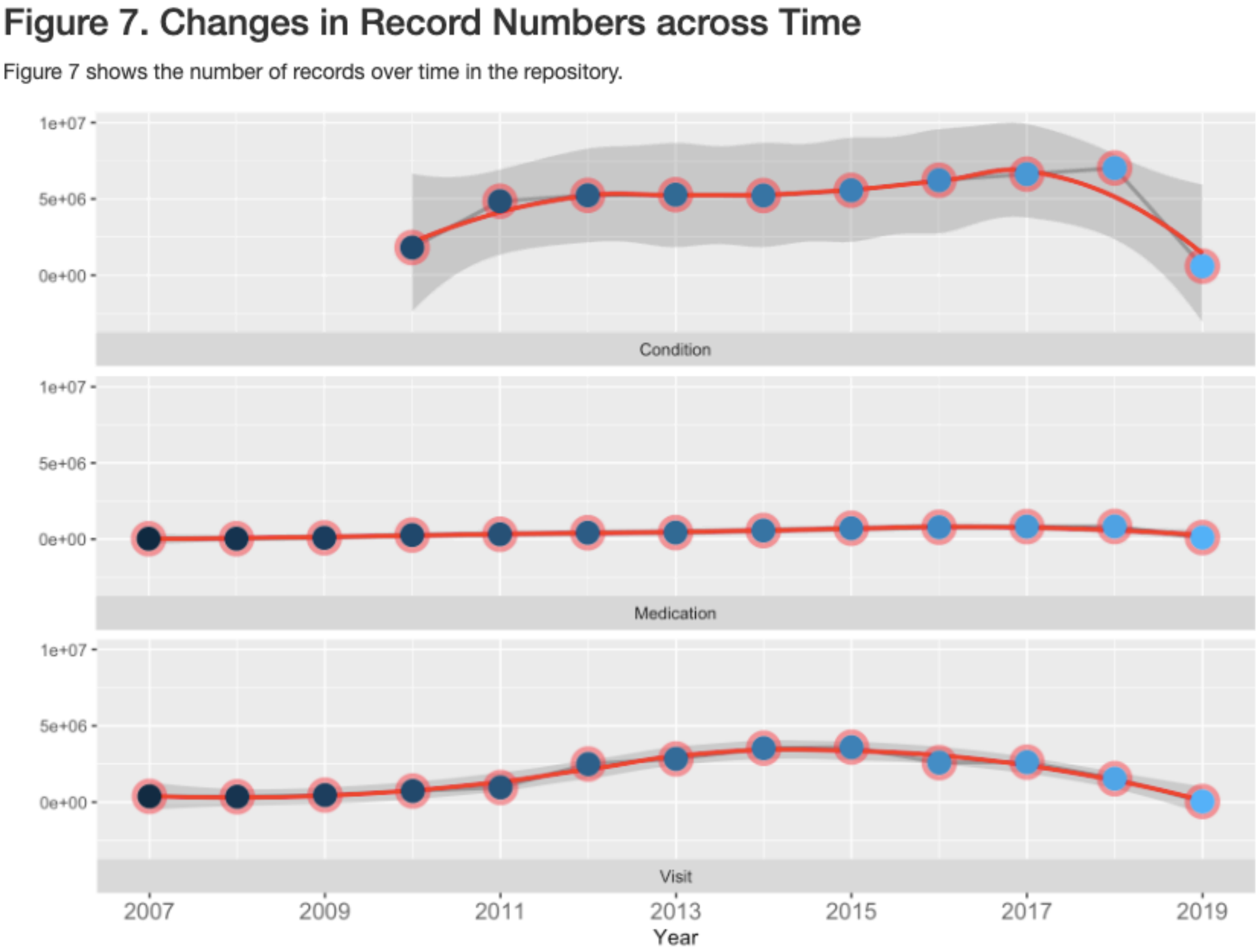
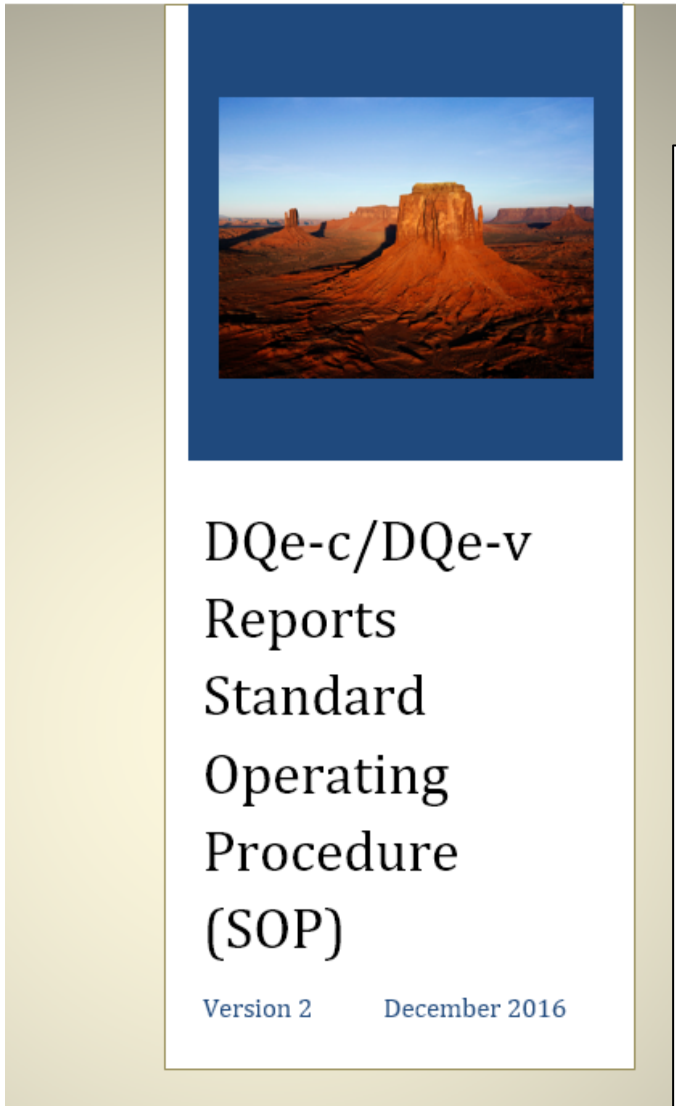


Figure 7. Changes in Record Numbers across Time

Figure 7 shows the number of records over time in the repository.

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

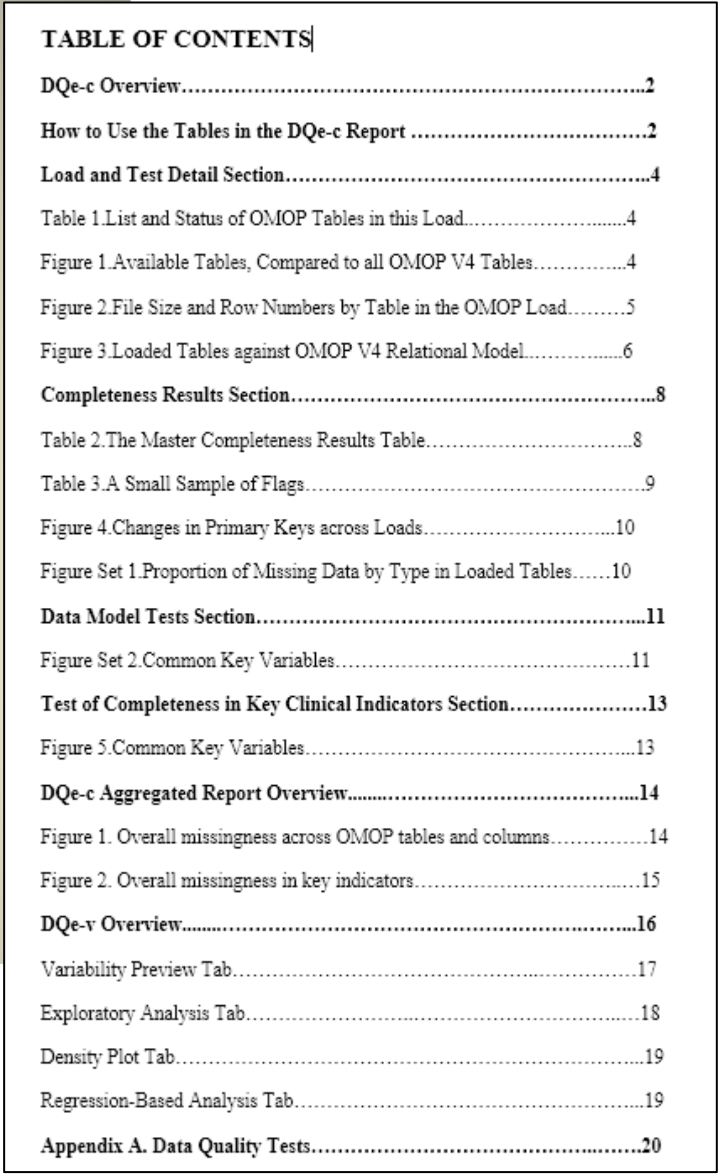


TABLE OF CONTENTS

https://github.com/WWAMI-DataQuest/DQe-c_OMOPv4/tree/master/docs

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

Questions?

- We are looking for collaborators and contributors!
- Contact me if you need help getting the tool up and running.
- We are always looking for feedback.

Thanks to Kari Stephens, Hossein Estiri, WPRN, ITHS, and CD2H!

Contact: Tim Bergquist trberg@uw.edu

https://dataquest.iths.org/

https://ctsa.ncats.nih.gov/cd2h/

https://github.com/data2health/DQe-c-v2

CD2H Data Quality Project

https://ctsa.ncats.nih.gov/cd2h/data-quality-methods-and-tools-to-support-ctsa-hub-data-sharing/

NATIONAL CENTER FOR DATA TO HEALTH