A journey toward real-world evidence for regulatory decision-making:

Building confidence in *real-world data*: Data quality reporting

Challenges: Data collection

- **Source data collection**: Health care data are collected to support patient care or to bill payors rather than for research.

Challenges: Data collection

Source data collection: Health care data are collected to support patient care or to bill payors rather than for research.

- The opportunities for errors of omission and distortion are greater than when data are collected for research.

Challenges: Data collection

Source data collection: Health care data are collected to support patient care or to bill payors rather than for research.

- The opportunities for errors of omission and distortion are greater than when data are collected for research.
- The power to standardize and improve data collection methods is less than when it is collected for research.

Challenges & Solutions: Data collection

Source data are heterogeneous: OMOP to the rescue!

- Equivalent codes get mapped to a standard concept.
- Standard representation yields
  - *Semantic interoperability*
  - *Common schema to write code against*
  - *The ability to leverage concept relationships in queries*

Challenges & Solutions: Data normalization

BUT!

Mapping source data to the OMOP CDM is complex!

Challenges & Solutions: Data normalization

BUT!

Mapping source data to the OMOP CDM is complex!

- It is easy to make mistakes when writing ETL code
  - The Rabbit-in-a-Hat tool supports the creation of unit tests – small bits of code that checks whether it functions as intended.

Challenges & Solutions: Data normalization

BUT!

<u>Mapping source data to the OMOP CDM is complex!</u>

- It is easy to make mistakes when writing ETL code
  - The Rabbit-in-a-Hat tool supports the creation of unit tests – small bits of code that checks whether it functions as intended.
  - Various studies have shown that with scrupulous attention, data can be transformed to the CDM with very little information loss.
    - These studies are cited in the Book of OHDSI

Challenges & Solutions: Data normalization

BUT!

<u>Mapping source data to the OMOP CDM is complex!</u>

- Even when coding mistakes are not made, there are many cases where there is more than one defensible way to do the right thing.

Challenges & Solutions: Data normalization

BUT!

<u>Mapping source data to the OMOP CDM is complex!</u>

- Even when mistakes are not made, there are many cases where there are more than one defensible way to do the right thing.
  - THEMIS is an ongoing process of defining and documenting conventions that the OHDSI community has agreed upon.
    - Can be found on the CDM Wiki.

Challenges & Solutions:
Data collection and normalization

- Healthcare data are prone to omissions and distortions
- Mapping source data to CDM is complex
- There are an enormous number of concepts in each domain and datasets are often very large

Kahn harmonized framework for data quality

Kahn and colleagues did an excellent job of synthesizing the terminology and categories used to conceptualize the data quality errors that affect RWD.

- eGEMs (Generating Evidence & Methods to improve patient outcomes), Vol. 4 [2016], Iss. 1, Art. 18

Kahn framework for data quality

Conformance: Do data values adhere to specified standards and formats?

Kahn framework for data quality

Conformance: Do data values adhere to specified standards and formats?

Completeness: Is a particular variable present OR does it contain all recorded values?

Kahn framework for data quality

Conformance: Do data values adhere to specified standards and formats?

Completeness: Is a particular variable present OR does it contain all recorded values?

Plausibility: Are data values believable?

Kahn framework for data quality

Conformance: Adherence to specified standards and formats

- Value
- Relational
- Computation

Completeness: Variable presence OR capture of all recorded values

Plausibility: Values believability

- Uniqueness
- Atemporal
- Temporal

Kahn framework for data quality

- **Verification:** assesses expected values and distributions using resources within the local environment.
- **Validation**: assesses alignment of data values with respect to relevant external benchmarks such as across multiple data sites

Other challenges: Expectations

- People bring the same expectations to healthcare data quality as they do to assessing data collected explicitly for research.
  - The criteria for assessing clinical data warehouse should not be perfection, it should transparency.
  - The goals should be to identify where there might be problems due to collection or ETL coding errors or divergence from conventions and to facilitate actions that address those problems.
- Understanding data provenance completely is desirable, but it might not be necessary for a fulsome assessment of relevant DQ problems when producing RWE.

Goals: Assess whether data are fit for use

FDA's RWE program

Two stage process:

1. Assess the clinical data repository level: I.e. a whole OMOP instance
2. Assess the clinical dataset derived from the repository for the specific purpose of generating evidence

Goals: Assess whether data are fit for use

FDA's RWE program

The Data Quality Dashboard

Where to begin with Data Quality?

CONTEXTS

CATEGORIES

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

Kahn et al., eGEMS 2016

Where to begin with Data Quality?

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

Data Quality Check

An aggregated summary statistic that can be computed from the data
to which a decision threshold can be applied to determine if the statistic meets expectation.

Where to begin with Data Quality?

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

Data Quality Check

An aggregated summary statistic that can be computed from the data

to which a decision threshold can be applied to determine if the statistic meets expectation.

An example data quality check…

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records with a value in the YEAR_OF_BIRTH field of the PERSON table less than 1850.

An example data quality check…

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records with a value in the YEAR_OF_BIRTH field of the PERSON table less than 1850.

…which we can make more generic…

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records with a value in the *CDM field* of the *CDM table* less than *a low value*.

…and apply to a different example.

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records with a value in the DAYS_SUPPLY field of the DRUG_EXPOSURE table less than 0.

What if we add units?

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

For a *measurement with associated unit*,

the number and percent of records with a

value in the *CDM field* of the *CDM table*

less than *a low value*.

What if we add units?

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

For Hemoglobin A1c with unit of percent,

the number and percent of records with a

value in the VALUE_AS_NUMBER field of the

MEASUREMENT table less than 4.

An example completeness check…

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records which are not mapped into a standard concept in the CONDITION_CONCEPT_ID field of the CONDITION_OCCURRENCE table.

An example completeness check…

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records which are not mapped into a standard concept in the CONDITION_CONCEPT_ID field of the CONDITION_OCCURRENCE table.

…which we can make more generic…

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records which are not mapped into a standard concept in the *CDM field* of the *CDM table*.

…and apply to a different example.

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records which are not mapped into a standard concept in the UNIT_CONCEPT_ID field of the MEASUREMENT table.

Data Quality Check *Types*

Check Type	Check Description
Person Completeness	The number and percent of persons in a database that do not have a least one record in the *CDM table.*
Is Required	The number and percent of records with a NULL value in a *CDM field* of a *CDM table* that is considered not nullable.
Is Foreign Key	The number and percent of records that have a value in a foreign key *CDM field* of a *CDM table* that does not exist in the *foreign key table.*
Is Standard Valid Concept	The number and percent of records that do not have a standard, valid concept in the *CDM field* of a *CDM table.*
Plausible Temporal After	The number and percent of records with a value in a *CDM field* of a *CDM table* that occurs prior to a *plausible date.*
...	
Plausible Value Low	For a given *CONCEPT_ID* and *UNIT_CONCEPT_ID* pair, the number and percent of records with a value lower than the *plausible low value.*
Plausible Gender	For a given *CONCEPT_ID*, the number and percent of records associated with persons with an *implausible gender.*

Data Quality Check *Types*

	Verification	Validation
Plausibility	6	1
Conformance	7	1
Completeness	4	1

20 Check *Types*

Data Quality Check *Totals*

	Verification	Validation
Plausibility	1878	287
Conformance	681	104
Completeness	386	15

Total 3,351 Checks

Data Quality Check *Thresholds*

Check Category	Check Type	Check Description	Check Result
Verification - Plausibility	Plausible Value Low	The number and percent of records with a value in the YEAR_OF_BIRTH field of the PERSON table less than 1850.	0%
Verification - Plausibility	Plausible Value Low	The number and percent of records with a value in the DAYS_SUPPLY field of the DRUG_EXPOSURE table less than 0.	0%
Verification - Plausibility	Plausible Value Low	For Hemoglobin A1c percent, the number and percent of records with a value in the VALUE_AS_NUMBER field of the MEASUREMENT table less than 4.	0.01%
Verification - Completeness	Is Standard Valid Concept	The number and percent of records with a value of 0 in the standard concept field CONDITION_CONCEPT_ID in the CONDITION_OCCURRENCE table.	0.02%
Verification - Completeness	Is Standard Valid Concept	The number and percent of records with a value of 0 in the standard concept field UNIT_CONCEPT_ID in the MEASUREMENT table.	93.66%

Data Quality Check *Totals*

	Verification	Validation
Plausibility	1878	287
Conformance		
Completeness	386	15

Total 3,351 Checks

Data Quality Check

An aggregated summary statistic that can be computed from the data

to which a decision threshold can be applied to determine if the statistic meets expectation.

Data Quality Check *Thresholds*

Check Category	Check Type	Check Description	Check Result
Verification - Plausibility	Plausible Value Low	The number and percent of records with a value in the YEAR_OF_BIRTH field of the PERSON table less than 1850.	0%
Verification - Plausibility			0%
Verification - Plausibility		MEASUREMENT table less than 4.	0.01%
Verification - Completeness	Is Standard Valid Concept	The number and percent of records with a value of 0 in the standard concept field CONDITION_CONCEPT_ID in the CONDITION_OCCURRENCE table.	0.02%
Verification - Completeness	Is Standard Valid Concept	The number and percent of records with a value of 0 in the standard concept field UNIT_CONCEPT_ID in the MEASUREMENT table.	93.66%

How do we decide if these results are 'good enough'?

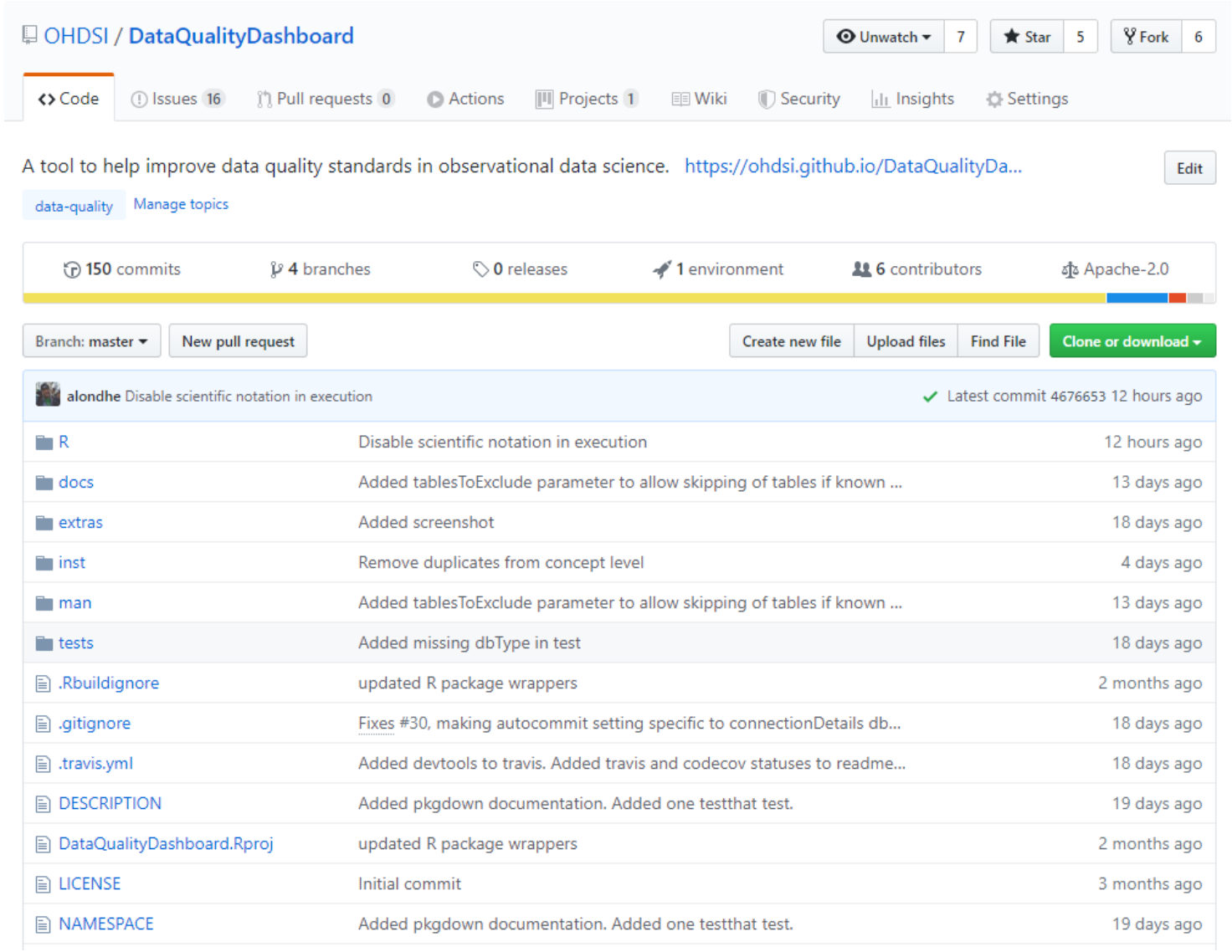
Data Quality Check *Thresholds*

Check Category	Check Type	Check Description	Check Result	Decision Threshold	Pass /Fail
Verification - Plausibility	Plausible Value Low	The number and percent of records with a value in the YEAR_OF_BIRTH field of the PERSON table less than 1850.	0%	0%	PASS
Verification - Plausibility	Plausible Value Low	The number and percent of records with a value in the DAYS_SUPPLY field of the DRUG_EXPOSURE table less than 0.	0%	1%	PASS
Verification - Plausibility	Plausible Value Low	For Hemoglobin A1c percent, the number and percent of records with a value in the VALUE_AS_NUMBER field of the MEASUREMENT table less than 4.	0.01%	5%	PASS
Verification - Completeness	Is Standard Valid Concept	The number and percent of records with a value of 0 in the standard concept field CONDITION_CONCEPT_ID in the CONDITION_OCCURRENCE table.	0.02%	5%	PASS
Verification - Completeness	Is Standard Valid Concept	The number and percent of records with a value of 0 in the standard concept field UNIT_CONCEPT_ID in the MEASUREMENT table.	93.66%	5%	FAIL

Data Quality Check *Thresholds*

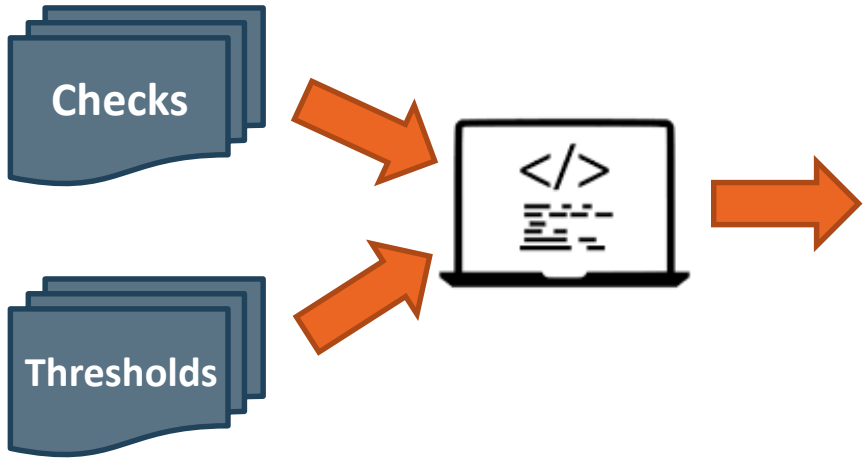
Check Category	Check Type	Check Description	Check Result	Decision Threshold	Pass /Fail
Verification - Plausibility	Plausible Value Low	The number and percent of records with a value in the YEAR_OF_BIRTH field of the PERSON table less than 1850.	0%	0%	PASS
Verification - Plausibility	Plausible Value Low	The number and percent of records with a value in the DAYS_SUPPLY field of the DRUG_EXPOSURE table less than 0.	0%	1%	PASS
Verification - Plausibility	Plausible Value Low	For Hemoglobin A1c percent, the number and percent of records with a value in the VALUE_AS_NUMBER field of the MEASUREMENT table less than 4.	0.01%	5%	PASS
Verification - Completeness	Is Standard Valid Concept	The number and percent of records with a value of 0 in the standard concept field CONDITION_CONCEPT_ID in the CONDITION_OCCURRENCE table.	0.02%	5%	PASS
Verification - Completeness	Is Standard Valid Concept	The number and percent of records with a value of 0 in the standard concept field UNIT_CONCEPT_ID in the MEASUREMENT table.	93.66%	**95%**	**PASS**

Data Quality Dashboard



https://github.com/OHDSI/DataQualityDashboard

Data Quality Dashboard



IBM MARKETSCAN COMMERCIAL CLAIMS AND ENCOUNTERS DATABASE

OVERVIEW

METADATA

RESULTS

ABOUT

RESULTS

IBM MARKETSCAN COMMERCIAL CLAIMS AND ENCOUNTERS DATABASE

Results generated at 2019-09-06 22:20:12 in 7 hours

Column visibility | CSV

Show 5 entries | Search:

STATUS	CONTEXT	CATEGORY	SUBCATEGORY	LEVEL	DESCRIPTION	% RECORDS
PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the range_high of the MEASUREMENT. (Threshold=100%).	82.14%
PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the visit_detail_id of the MEASUREMENT. (Threshold=100%).	80.90%
PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the value_source_value of the MEASUREMENT. (Threshold=100%).	79.89%
PASS	Validation	Completeness	None	TABLE	The number and percent of persons in the CDM that do not have at least one record in the DEVICE_EXPOSURE table (Threshold=100%).	76.70%
FAIL	Verification	Plausibility	Atemporal	CONCEPT	For the combination of CONCEPT_ID 3016049 (Testosterone Free [Mass/volume] in Serum or Plasma) and UNIT_CONCEPT_ID 8845 (picogram per milliliter), the number and percent of records that have a value less than 5.00e+00. (Threshold=1%).	72.43%

Showing 126 to 130 of 3,351 entries

Previous 1 ... 25 26 27 ... 671 Next

https://github.com/OHDSI/DataQualityDashboard

Data Quality Dashboard – Korea

DATA QUALITY ASSESSMENT

THE NATIONAL HEALTH INSURANCE SERVICE?NATIONAL SAMPLE COHORT

Results generated at 2019-08-28 14:14:40 in 2 hours

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	173	7	180	96%	211	72	283	75%	384	79	463	83%
Conformance	631	40	671	94%	104	0	104	100%	735	40	775	95%
Completeness	378	8	386	98%	2	13	15	13%	380	21	401	95%
Total	1182	55	1237	96%	317	85	402	79%	1499	140	1639	**91%**

Data Quality Dashboard – IBM CCAE

IBM MARKETSCAN COMMERCIAL CLAIMS AND ENCOUNTERS DATABASE

OVERVIEW

METADATA

RESULTS

ABOUT

RESULTS

IBM MARKETSCAN COMMERCIAL CLAIMS AND ENCOUNTERS DATABASE

Results generated at 2019-09-06 22:20:12 in 7 hours

Column visibility | CSV

Show 5 entries

Search:

STATUS	CONTEXT	CATEGORY	SUBCATEGORY	LEVEL	DESCRIPTION	% RECORDS
PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the value_as_string of the OBSERVATION. (Threshold=100%).	94.51%
FAIL	Verification	Completeness	None	FIELD	The number and percent of records with a value of 0 in the standard concept field unit_concept_id in the MEASUREMENT table. (Threshold=5%).	93.66%
FAIL	Verification	Plausibility	Atemporal	CONCEPT	For the combination of CONCEPT_ID 3007359 (Bilirubin.indirect [Mass/volume] in Serum or Plasma) and UNIT_CONCEPT_ID 8840 (milligram per deciliter), the number and percent of records that have a value less than 1.00e+00. (Threshold=1%).	92.90%
PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the visit_detail_id of the OBSERVATION. (Threshold=100%).	92.75%
FAIL	Verification	Plausibility	Atemporal	CONCEPT	For the combination of CONCEPT_ID 3010340 (Triiodothyronine (T3) [Mass/volume] in Serum or Plasma) and UNIT_CONCEPT_ID 8842 (nanogram per milliliter), the number and percent of records that have a value less than 6.00e+01. (Threshold=1%).	92.60%

Showing 106 to 110 of 3,351 entries

Previous 1 ... 21 22 23 ... 671 Next

Data Quality Dashboard – IBM CCAE

IBM MARKETSCAN COMMERCIAL CLAIMS AND ENCOUNTERS DATABASE

OVERVIEW

METADATA

RESULTS

	STATUS	CONTEXT	CATEGORY	SUBCATEGORY	LEVEL	DESCRIPTION	% RECORDS
⊞	PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the value_as_string of the OBSERVATION. (Threshold=100%).	94.51%
⊞	FAIL	Verification	Completeness	None	FIELD	The number and percent of records with a value of 0 in the standard concept field unit_concept_id in the MEASUREMENT table. (Threshold=5%).	93.66%
⊞	FAIL	Verification	Plausibility	Atemporal	CONCEPT	For the combination of CONCEPT_ID 3007359 (Bilirubin.indirect [Mass/volume] in Serum or Plasma) and UNIT_CONCEPT_ID 8840 (milligram per deciliter), the number and percent of records that have a value less than 1.00e+00. (Threshold=1%).	92.90%
⊟	PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the visit_detail_id of the OBSERVATION. (Threshold=100%).	92.75%

Name:	measureValueCompleteness
Description:	The number and percent of records with a NULL value in the visit_detail_id of the OBSERVATION. (Threshold=100%).
Level:	FIELD
# Rows Violated:	5989330347
% Rows Violated:	92.75%
Execution Time:	15.447462 secs

Data Quality Dashboard – IBM CCAE

IBM MARKETSCAN COMMERCIAL CLAIMS AND ENCOUNTERS DATABASE

OVERVIEW

METADATA

RESULTS

ABOUT

Name:	measureValueCompleteness
Description:	The number and percent of records with a NULL value in the visit_detail_id of the OBSERVATION. (Threshold=100%).
Level:	FIELD
# Rows Violated:	5989330347
% Rows Violated:	92.75%
Execution Time:	15.447462 secs

SQL Query:

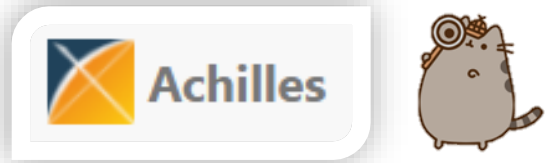
```
/*********
MEASURE_VALUE_COMPLETENESS
Computing number of null values and the proportion to total records per field

Parameters used in this template:
cdmDatabaseSchema = cdm_ibm_ccae_v1022
cdmTableName = OBSERVATION
cdmFieldName = visit_detail_id
**********/


SELECT num_violated_rows, CASE WHEN denominator.num_rows = 0 THEN 0 ELSE 1.0*num_violated_rows/denominator.num_rows END  AS pct_violated_rows
FROM
(
	SELECT COUNT(violated_rows.violating_field) AS num_violated_rows
	FROM
	(
		SELECT 'OBSERVATION.visit_detail_id' AS violating_field, OBSERVATION.*
		FROM cdm_ibm_ccae_v1022.OBSERVATION
		WHERE cdm_ibm_ccae_v1022.OBSERVATION.visit_detail_id IS NULL
	) violated_rows
) violated_row_count,
(
	SELECT COUNT(*) AS num_rows
	FROM cdm_ibm_ccae_v1022.OBSERVATION
) denominator
;
```

Data Quality Dashboard – Final Thoughts

- The Data Quality Dashboard (DQD) takes a set of data quality check types and systematically applies them to all relevant tables and fields
- It then evaluates the checks based on an a priori set of decision thresholds
- These checks and threshold results are then communicated through a JSON object and the Shiny app

https://github.com/OHDSI/DataQualityDashboard

Achilles



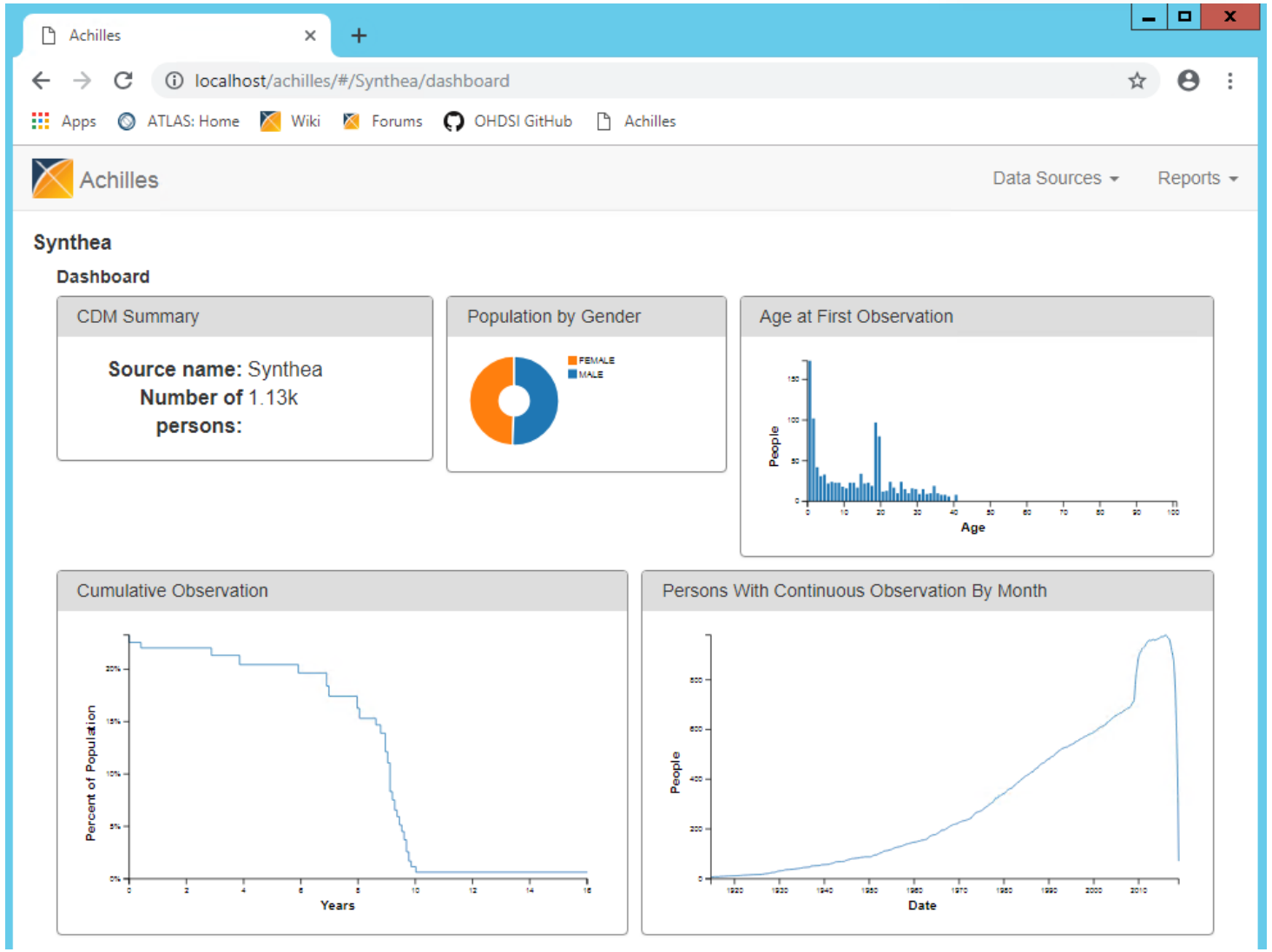
Achilles is a data characterization and quality tool available for download here:

https://github.com/OHDSI/Achilles

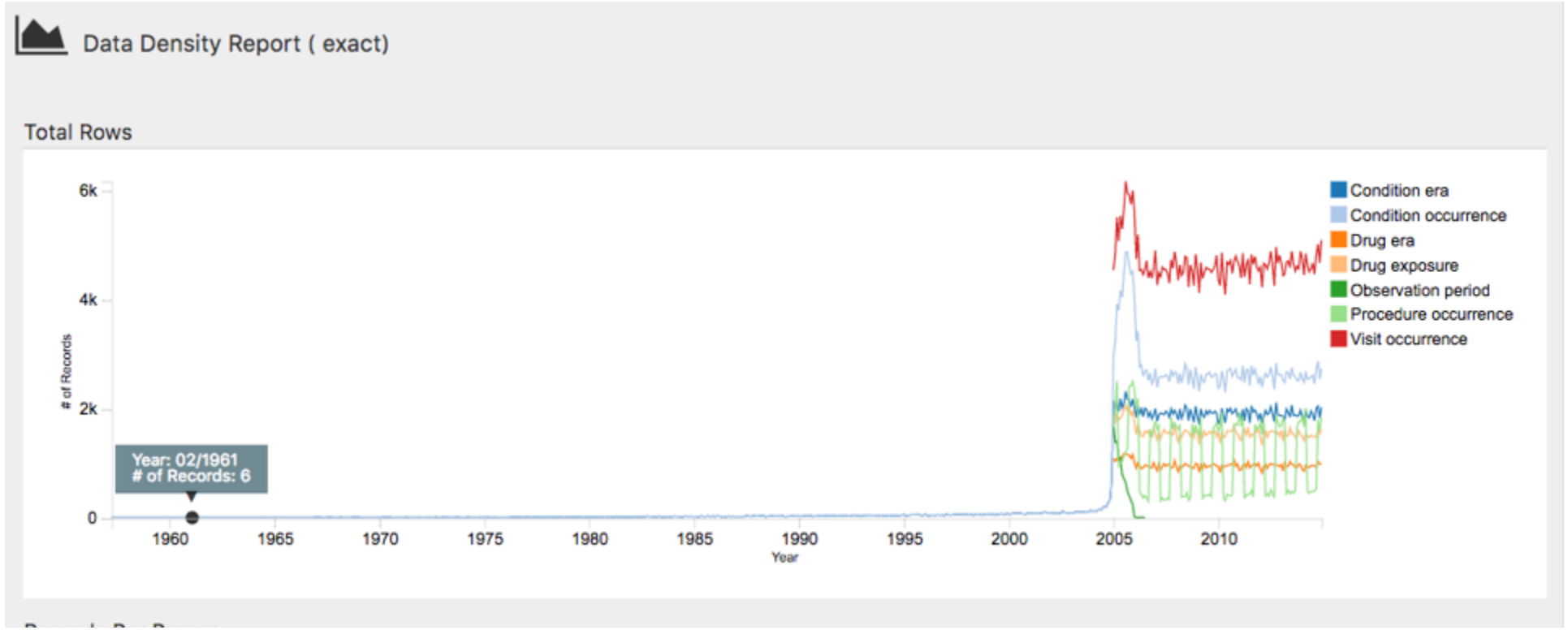
For an example of how it was run for on sample data, that R script is located here:

https://github.com/OHDSI/Tutorial-ETL/blob/master/materials/Achilles/achillesRun.R

Achilles

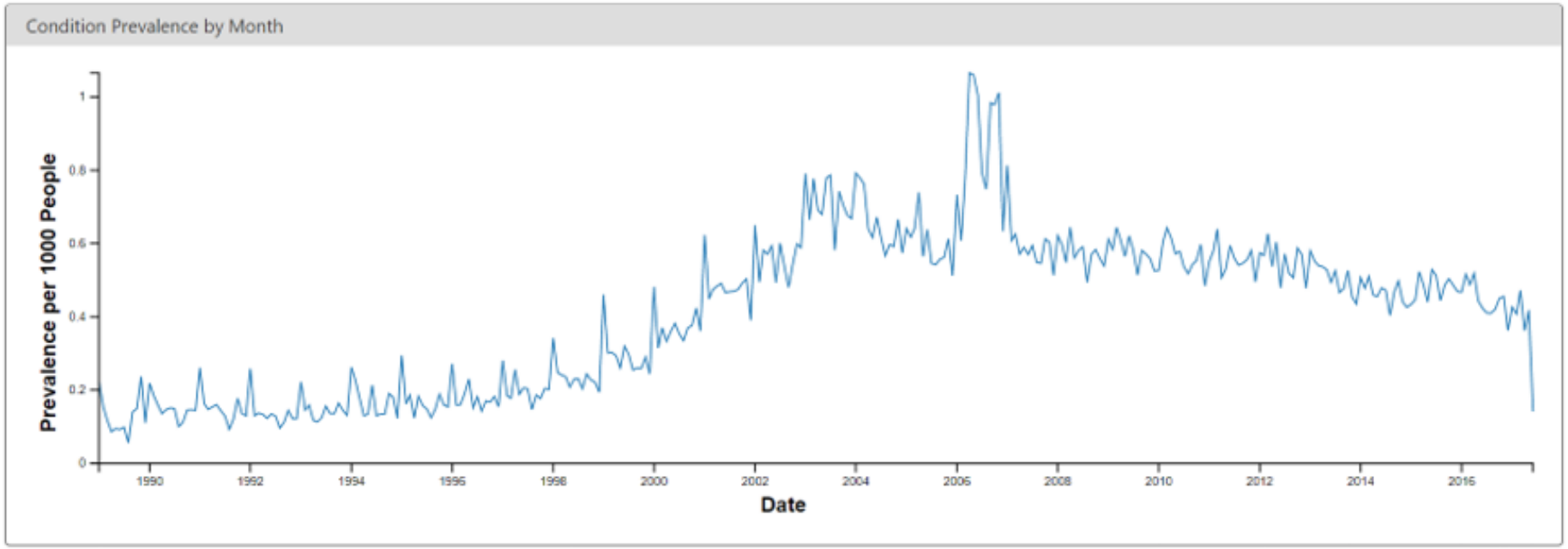

Achilles
localhost/achilles/#/Synthea/dashboard
Apps
ATLAS: Home
Wiki
Forums
OHDSI GitHub
Achilles
Achilles
Data Sources
Reports
Synthea
Dashboard
CDM Summary
Source name: Synthea
Number of persons: 1.13k
Population by Gender
FEMALE
MALE
Age at First Observation
People
Age
Cumulative Observation
Percent of Population
Years
Persons With Continuous Observation By Month
People
Date

Achilles



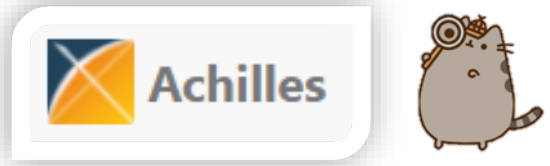
This plot shows that the bulk of the data starts in 2005. However, there also appear to be a few records from around 1961, which is likely an error in the data.

Achilles



This change coincides with changes in the reimbursement rules in this specific country, leading to more diagnoses but probably not a true increase in prevalence in the underlying population.

Achilles Heel



Achilles heel is a report generated by the Achilles application that will run a series of data quality checks on the CDM using the Achilles data

Message Type	Message
ERROR	410-Number of condition occurrence records outside valid observation period; count (n=134) should not be >
ERROR	610-Number of procedure occurrence records outside valid observation period; count (n=11) should not be >
ERROR	710-Number of drug exposure records outside valid observation period; count (n=241) should not be > 0
ERROR	712-Number of drug exposure records with invalid provider_id; count (n=29,518) should not be > 0
ERROR	810-Number of observation records outside valid observation period; count (n=134) should not be > 0
ERROR	812-Number of observation records with invalid provider_id; count (n=8,518) should not be > 0
ERROR	909-Number of drug eras outside valid observation period; count (n=55) should not be > 0
ERROR	1,009-Number of condition eras outside valid observation period; count (n=134) should not be > 0
NOTIFICATION	[GeneralPopulationOnly] Not all deciles represented at first observation
NOTIFICATION	Unmapped data over percentage threshold in:Measurement
NOTIFICATION	Unmapped data over percentage threshold in:DrugExposure
NOTIFICATION	Unmapped data over percentage threshold in:Observation
NOTIFICATION	99+ percent of persons have exactly one observation period
NOTIFICATION	percentage of non-numerical measurement records exceeds general population threshold
NOTIFICATION	Unmapped data over percentage threshold in:Condition

Showing 1 to 15 of 25 entries Print Previous 1 2 Next

THANK YOU!

https://github.com/OHDSI/Achilles

https://github.com/OHDSI/DataQualityDashboard