OMOP CDM: The Validation Approach

Jon Duke, MD

Director, Center for Health Analytics and Informatics

Georgia Institute of Technology

Disclosure

- Have received prior or current funding from Celgene, Merck, Bayer, and Janssen.
- No COI related to today's presentation

“Loss of fidelity begins with the movement of data from the doctor’s brain to the medical record.”

-- Clem McDonald, MD

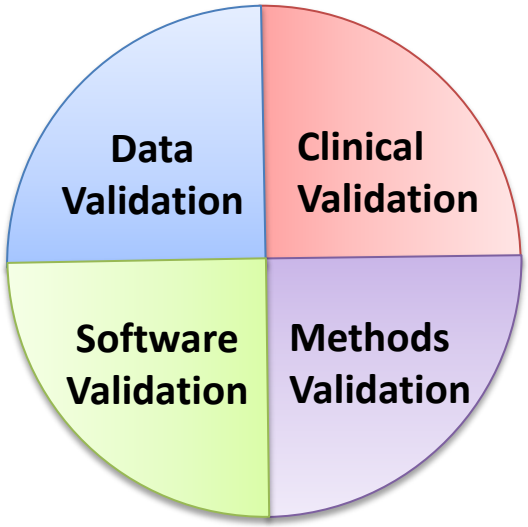
Director, Lister Hill Center for Biomedical Informatics

National Library of Medicine, USA

What is Validation?

- OHDSI views validation as a *holistic set of processes* necessary to achieve the highest quality reproducible evidence from diverse data sources

Validation Components

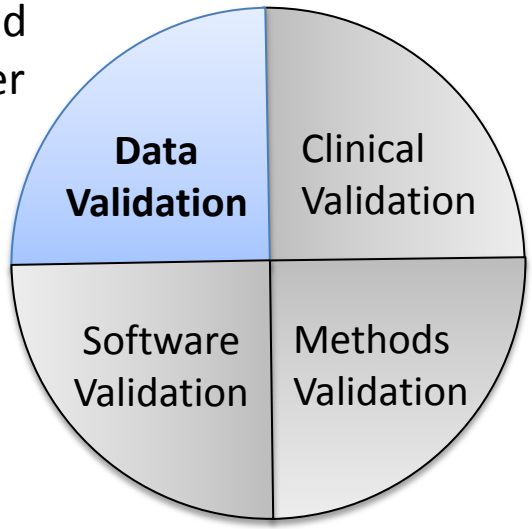


Enabling Validation

- OHDSI *enables* validation by the community through tools, workgroups, shared data discovery, and collaboration platforms
- A research network within OHDSI may *enforce* given validation processes as appropriate for that network's objectives

Data Validation

Are the data completely captured with plausible values in a manner that is conformant to agreed structure and conventions?



eGEMs

Generating Evidence & Methods to improve patient outcomes

A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data

Michael G. Kahn, MD, PhD;[i] Tiffany J. Callahan, MPH;[i] Juliana Barnard, MA;[i] Alan E. Bauck;[ii] Jeff Brown, PhD;[iii] Bruce N. Davidson, PhD;[iv] Hossein Estiri, PhD;[v] Carsten Goerg, PhD;[i] Erin Holve, PhD, MPH, MPP;[vi] Steven G. Johnson, MS;[vii] Siaw-Teng Liaw, MBBS, PhD, FRACGP, FACHI;[viii] Marianne Hamilton-Lopez, PhD, MPA;[ix] Daniella Meeker, PhD;[x] Toan C. Ong, PhD;[xi] Patrick Ryan, PhD;[xii] Ning Shang, PhD;[xiii] Nicole G. Weiskopf, PhD;[xiv] Chunhua Weng, PhD, FACMI;[xiii] Meredith N. Zozus, PhD;[xv] Lisa Schilling, MD[xi]

Data Validation

Truven CCAE: CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID	CONDITION_CONCEPT_ID
157033702	5/31/2000	41071	Inpatient claims - primary position	44825429	444406

Optum Extended SES: CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID	CONDITION_CONCEPT_ID
259000474406532	5/30/2000	41071	Inpatient claims - 1st position	44825429	444406

Premier : CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID	CONDITION_CONCEPT_ID
-171971409	1/1/2000	410.71	Hospital record - primary	44825429	444406

JMDC : CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID	CONDITION_CONCEPT_ID
4149337	4/11/2013	I214	Inpatient claims	45572081	444406

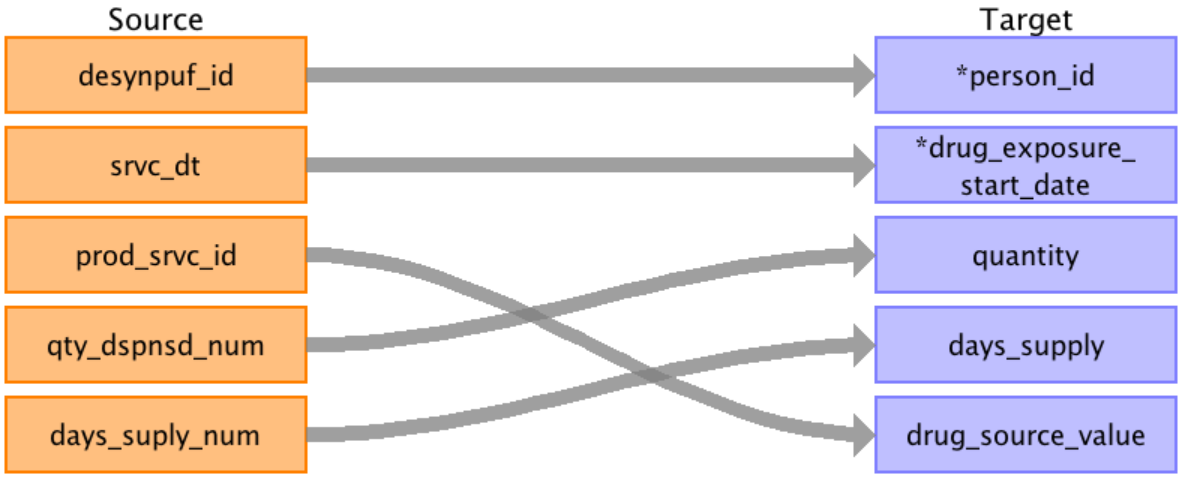
- How do we ensure preservation of source data into the CDM?
- How do we ensure ETL conventions are followed?
- How do we ensure vocabulary mappings are correct?
- How do we detect issues in the underlying data?

Data Validation Tools

- OHDSI has a set of tools to support data quality and CDM conformance both before and after ETL
- Pre-ETL
  - White Rabbit / Rabbit in a Hat
  - Usagi
- Post-ETL
  - ACHILLES

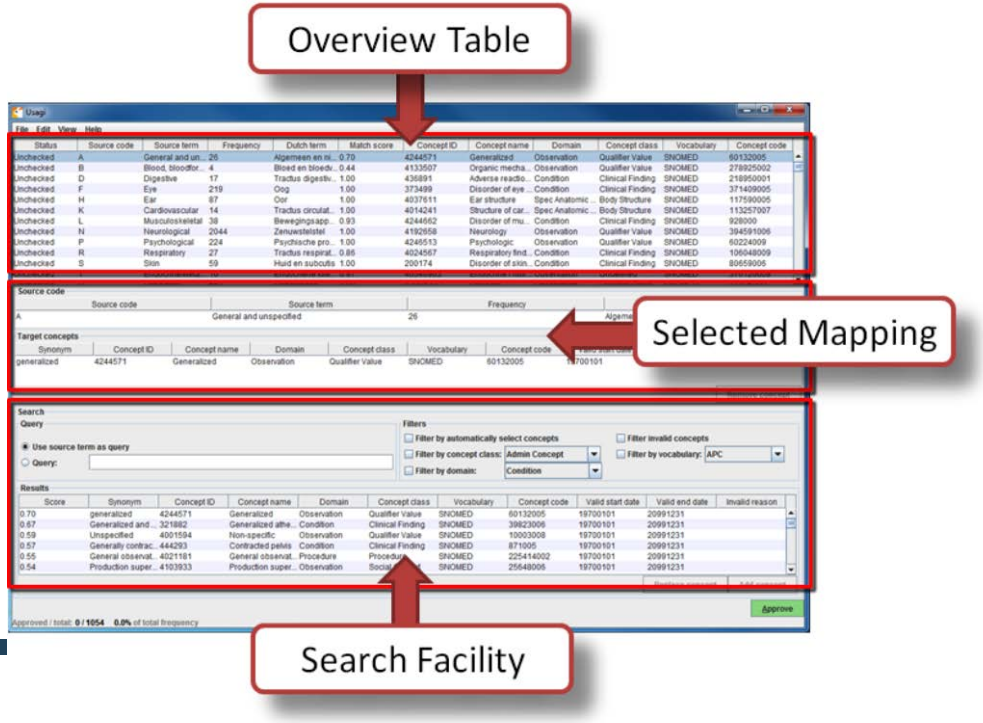
Pre-ETL Tools

- White Rabbit
  - Profiles source data and highlights patterns in source values (eg. variability, frequency)
- Rabbit in a Hat
  - Provides consistent mechanism for documenting ETL processes for each dataset to OMOP CDM



Pre-ETL Tools

- Usagi
  - Tool to support vocabulary mapping including mapping of local codes to standards
  - Exposes quality and density of fully mapped vs partially mapped vs unmapped content



Post-ETL Data Validation

- ACHILLES is OHDSI's standardized data characterization, quality, and CDM conformance package
- ACHILLES performs checks on every domain and every concept within the dataset
- OHDSI recommends running ACHILLES following ETL and subsequent refreshes

Example Quality Checks

- Demographics

3-Number of persons by year of birth; should not have year of birth in the future, (n=44,418)

- Date inconsistencies

711-Number of drug exposure records with end date < start date; count (n=15,922) should not be > 0

- Data drop off

902-Number of persons by drug era start month, by drug_concept_id; 17concepts have a 100% change in monthly count

- Data outside observation periods

410-Number of condition occurrence records outside valid observation period; count (n=29) should not be > 0

- Check for certain types of data

No body weight data in MEASUREMENT table (under concept_id 3,025,315 (LOINC code 29,463-7))

Full List at ohdsi.org

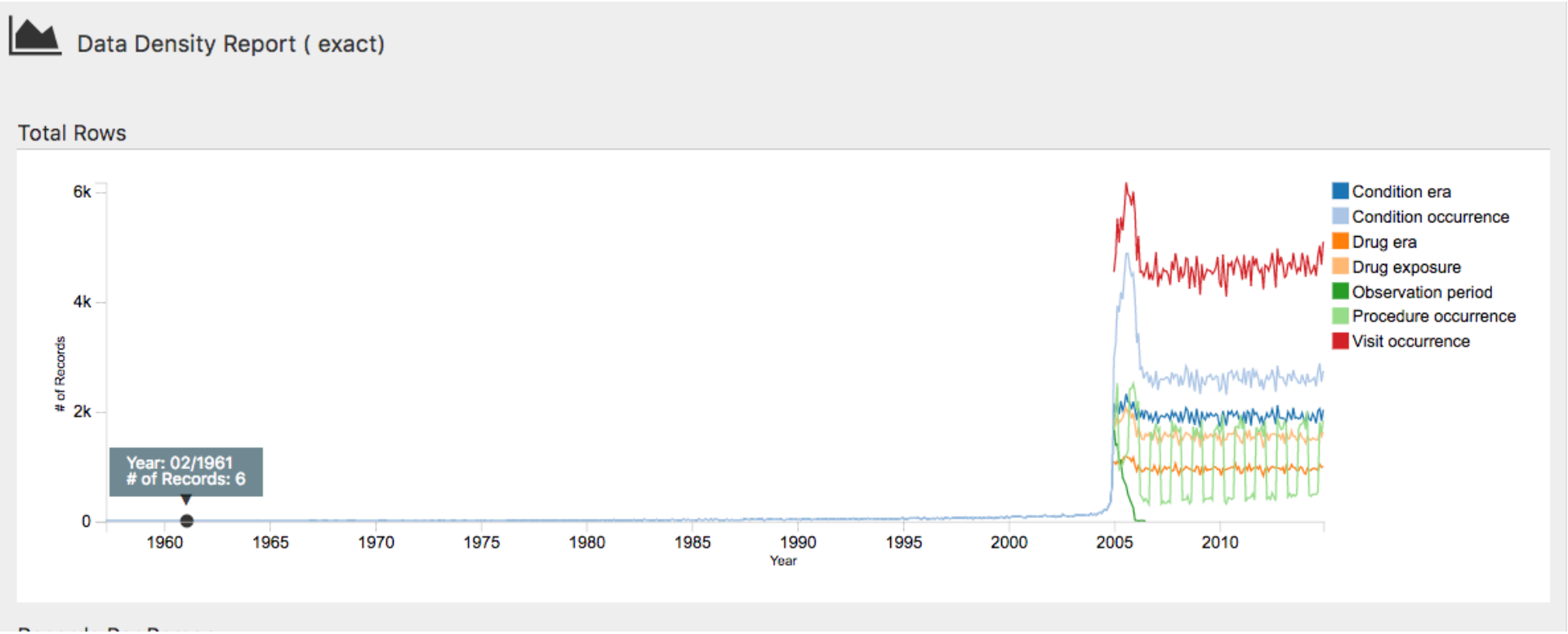
Example Conformance Checks

```
NOTIFICATION    Count of unmapped source values exceeds
threshold in: drug_exposure
NOTIFICATION    Count of unmapped source values exceeds
threshold in: measurement
WARNING 4-Number of persons by race; data with unmapped concepts
WARNING 301-Number of providers by specialty concept_id; data
with unmapped concepts
WARNING 400-Number of persons with at least one condition
occurrence, by condition_concept_id; data with unmapped concepts
WARNING 600-Number of persons with at least one procedure
occurrence, by procedure_concept_id; data with unmapped concepts
WARNING 700-Number of persons with at least one drug exposure,
by drug_concept_id; data with unmapped concepts
WARNING 800-Number of persons with at least one observation
occurrence, by observation_concept_id; data with unmapped
concepts
WARNING 1,000-Number of persons with at least one condition
era, by condition_concept_id; data with unmapped concepts
```

Full List at ohdsi.org

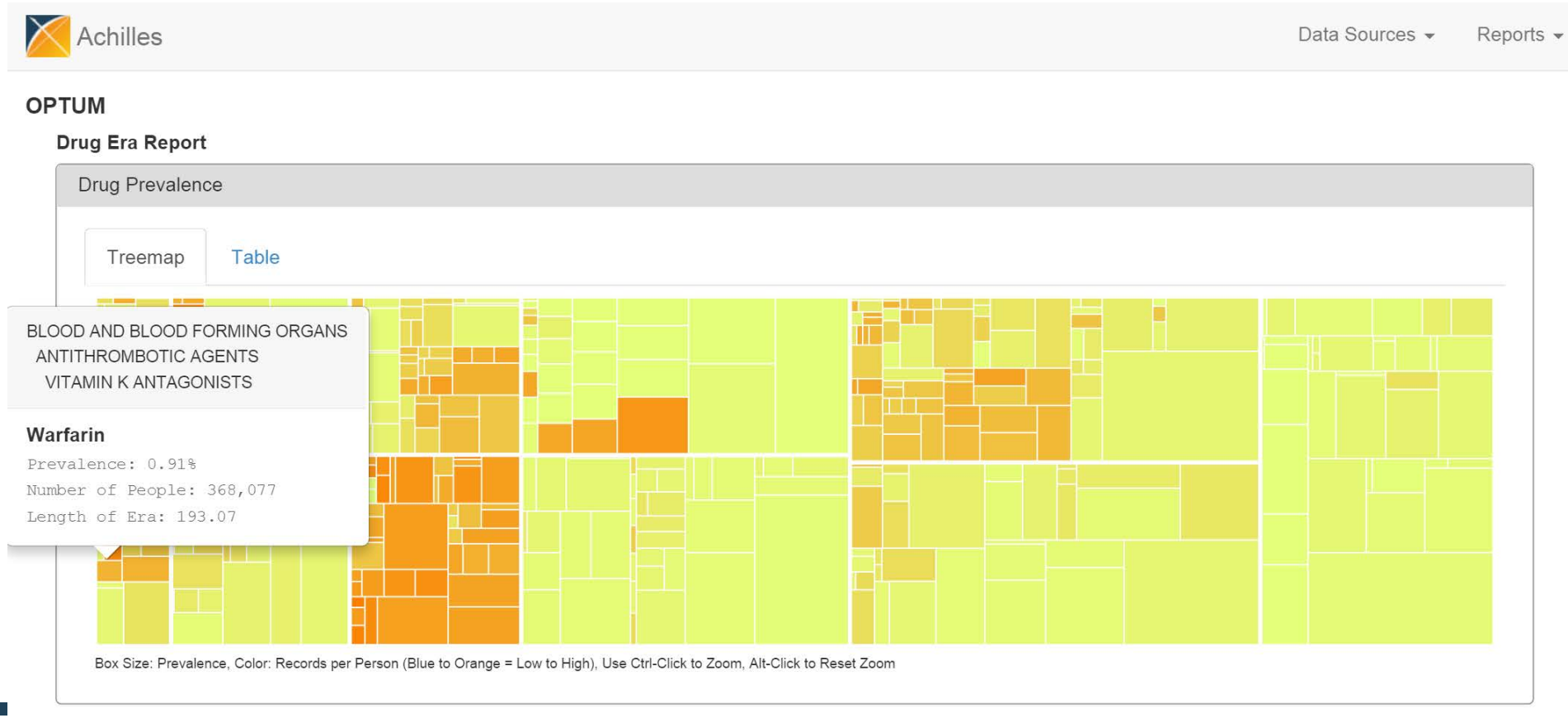
Post-ETL Validation

- Use ACHILLES visualizations to review anomalies



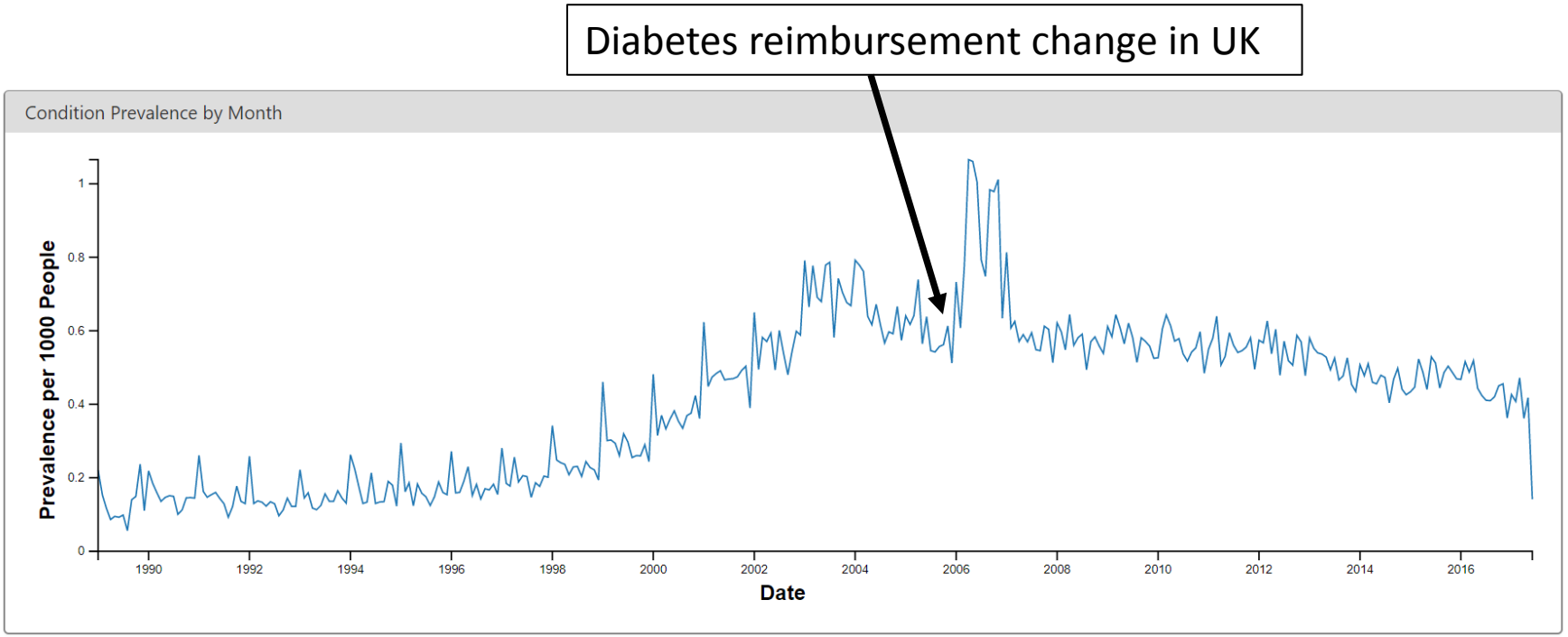
Post-ETL Validation

- Use ACHILLES visualizations to review anomalies



Post-ETL Validation

- Use ACHILLES visualizations to review anomalies



Checks Can Emerge from Community

```
--ruleid 42 DQ rule
--Percentage of outpatient visits (concept_id 9202) is
too low (for general population).
--This may indicate a dataset with mostly inpatient
data (that may be biased and missing some EHR events)
--Threshold was decided as 10th percentile in empiric
comparison of 12 real world datasets in the DQ-Study2
```

```
--ruleid 38 DQ rule; in a general dataset, it is
expected that more than providers with a wide range of
specialties
--(at least more than just one specialty) is present
--notification  may indicate that provider table is
missing data on specialty
--typical dataset has at least 28 specialties present
in provider table
```

```
--ruleid 44 DQ rule
--uses iris measure: patients with at least 1 Meas, 1 Dx and 1 Rx
```

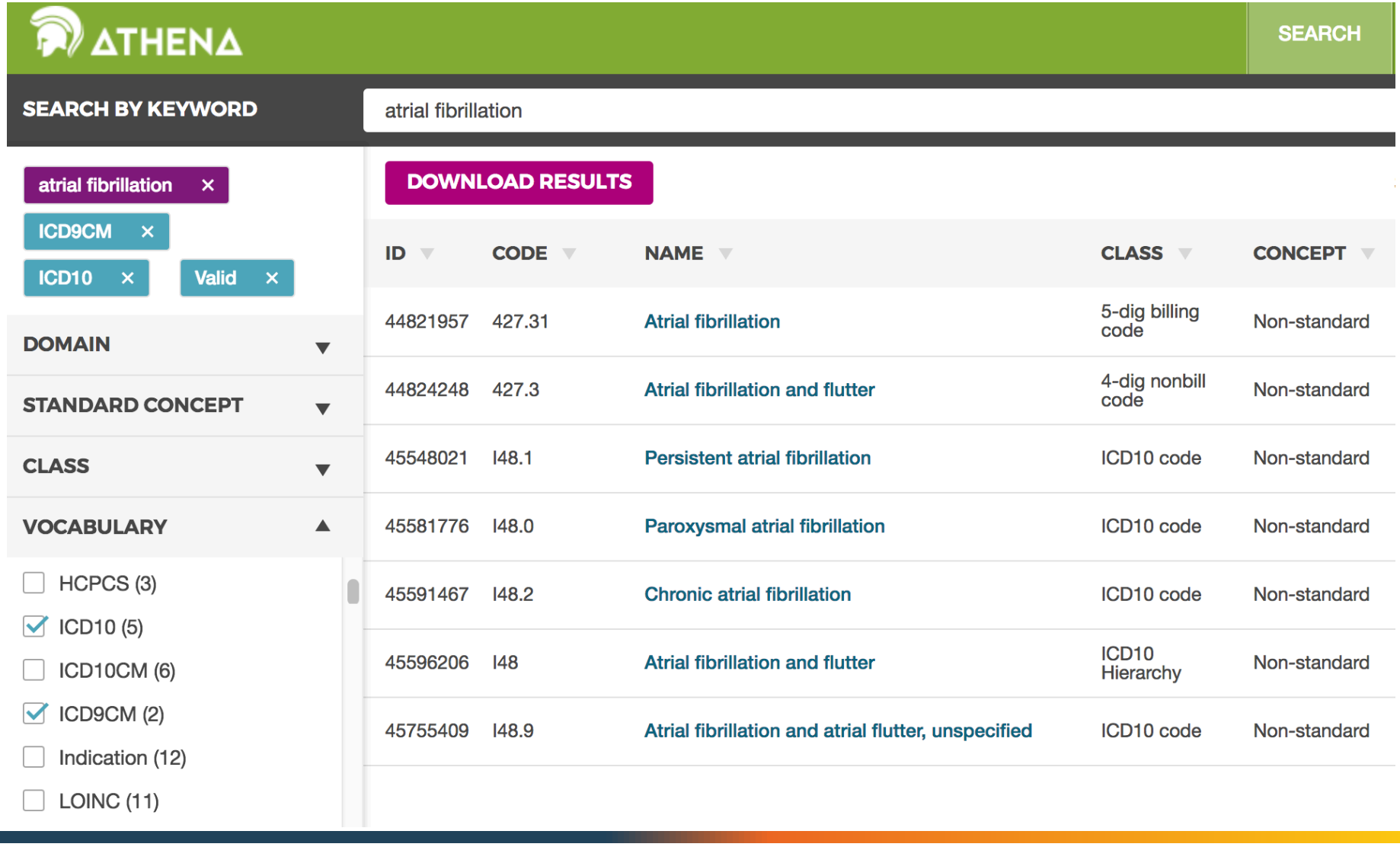
ETL Conventions

- In addition to the previously mentioned tools, ETL conventions are evolved by a public working group known as THEMIS
  - e.g. how to handle multiple birth dates, conflicting genders, events after death
- Any group of sites with data in the OMOP model (eg, a collaborative network) can require a common set of conventions

Vocabulary Validation

- The majority of OMOP vocabulary content is external and used without modification
  - Integration testing and all scripts publicly available
- Some mappings created manually
  - Eg. ICD-9-CM to SNOMED, ICD-10 to SNOMED, local drug forms
  - 2nd-level review by independent coding manager
- Quality and change evaluation scripts run prior to each release
- OHDSI Community review

Rapidly Reviewable



Community Vetted

Gemscript to RxNorm mapping error

Vocabulary Users

DTorok Don Torok — 12h

This is from v5.0 03-MAY-17, so maybe it is fixed. But gemscript code 55991020 (Levothyroxine sodium 100microgram tablets) maps to two different RxNorm concepts. 40169799 (Levothyroxine Sodium 1 MG Oral Tablet) and 40169766 (Levothyroxine Sodium 0.1 MG Oral Tablet). There are other example like this where the Gemscript code maps to more than one RxNorm code where the descriptions for the RxNorm codes are the same except for the dose amount.

Reply

created	last reply	1	17	2
12 hours	3 hours	reply	views	users

aostropolets Anna Ostropolets — 3h

@DTorok
Was fixed in August release.
Thanks for noticing!

Reply

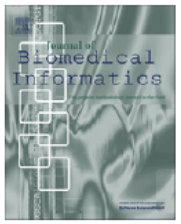
How does use of a standard terminology affect analysis results?

Journal of Biomedical Informatics 45 (2012) 689–696

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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



Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases ☆

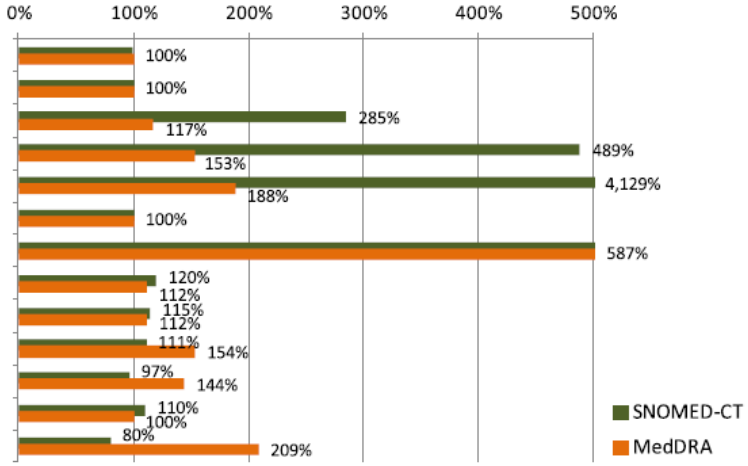
Christian Reich [a,*], Patrick B. Ryan [a,b,1], Paul E. Stang [a,b,1], Mitra Rocca [c,2]

[a] Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, 9650 Rockville Pike, Bethesda, MD 20814, USA
[b] Janssen Research & Development, LLC, 1125 Trenton-Harbourton Road, PO Box 200, MS K304, Titusville, NJ 08560, USA
[c] Office of Translational Sciences, Center for Drug Evaluation and Research (CDER), US Food and Drug Administration, 10903 New Hampshire Ave., Bldg. 21, Rm. 4608, Silver Spring, MD 20933, USA

Changing language may change your code list, which may change your cohort depending on the disease…

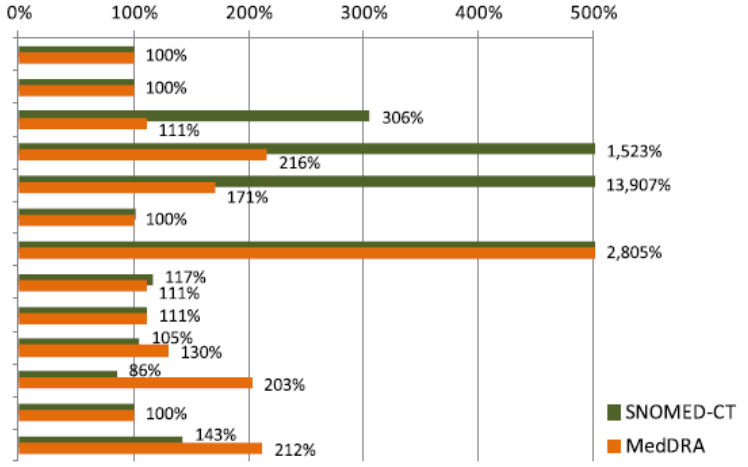
Cohort size of HOI in MSLR for different terminologies

Condition	#
Angioedema	1
Aplastic Anemia	1
Acute Liver Failure	1
	2
	7
Acute Renal Failure	1
	4
Bleeding	1
	3
Hip Fracture	1
Acute myocardial Infarction	1
	2
Upper GI Ulcer Hospitalization	1



Cohort size of HOI in GE for different terminologies

Condition	#
Angioedema	1
Aplastic Anemia	1
Acute Liver Failure	1
	2
	7
Acute Renal Failure	1
	4
Bleeding	1
	3
Hip Fracture	1
Acute myocardial Infarction	1
	2
Upper GI Ulcer Hospitalization	1



…but in practice, running an estimation analysis using source vs. standard vocabulary yields similar results

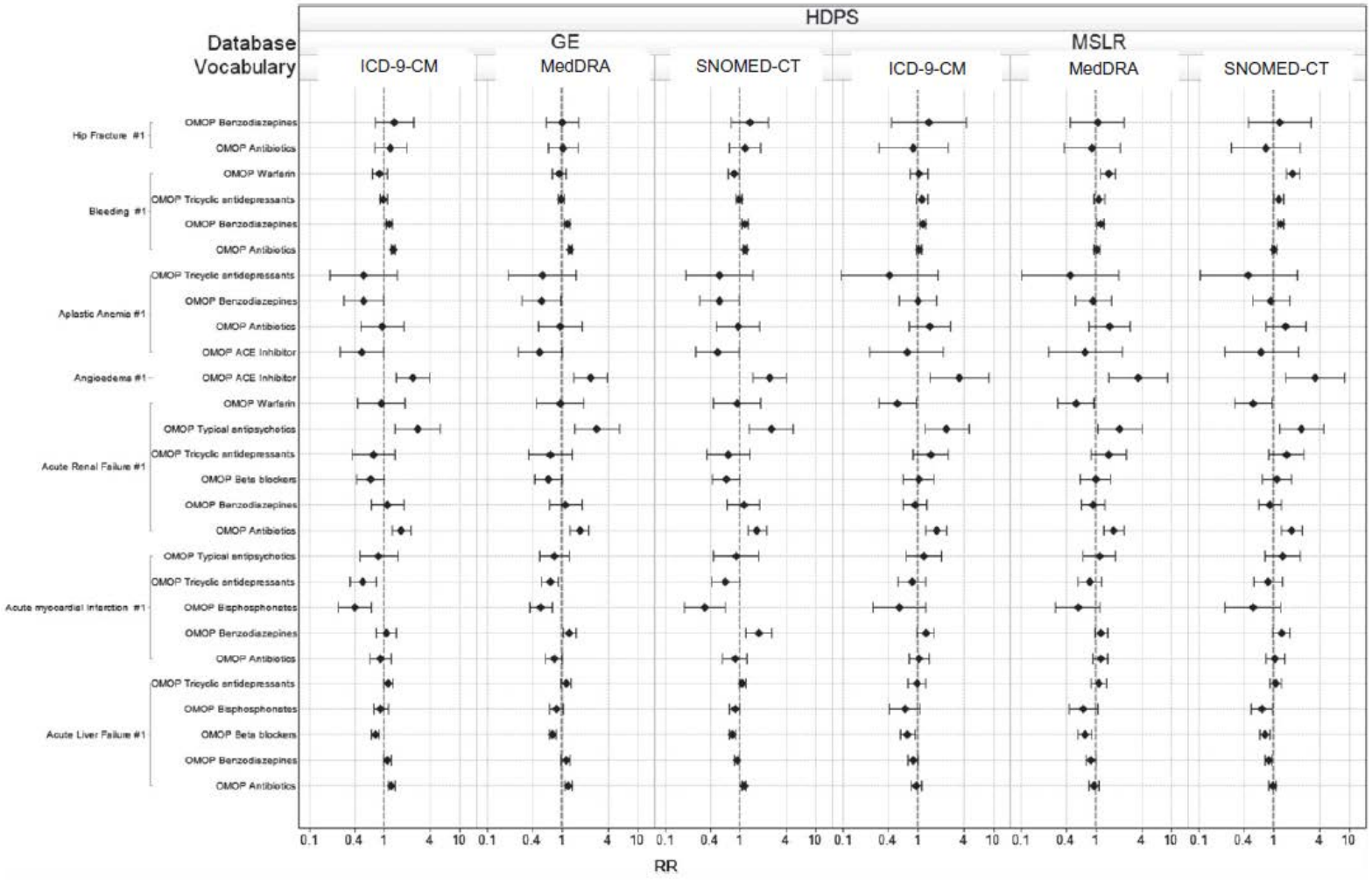
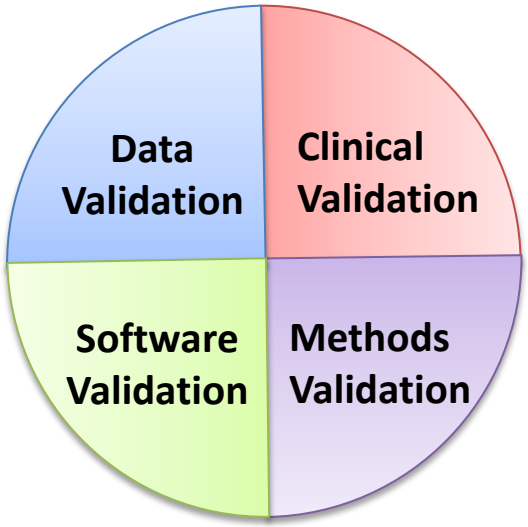


Fig. 3. Effect estimates and 95% confidence intervals for incident user design applied to MSLR and GE using ICD-9-CM, SNOMED-CT, and MedDRA as standard terminologies. Each dot represents the estimate of the effect of an individual HOI-drug combination (on the X-axis).

More Data Quality Questions?

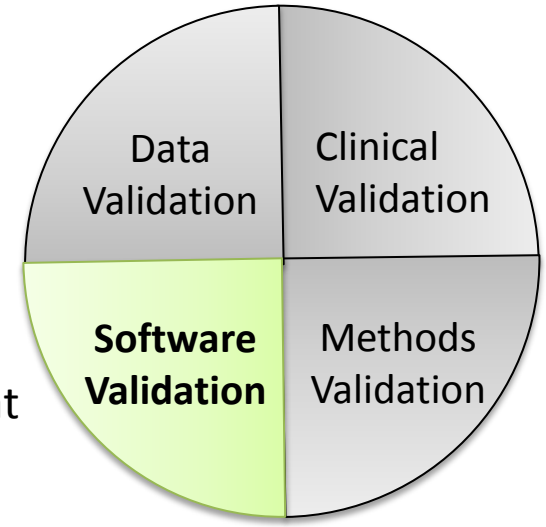
Check out our detailed FAQ at

tinyurl.com/ohdsi-faq


Data Validation
Clinical Validation
Software Validation
Methods Validation

Software Validation

Does the software do what it is expected to do?



Software Validation

- OHDSI creates two types of software
  - *ETL software* to transform source data to CDM
  - *Analytics software* that performs analyses on CDM
- Software should follow best-practice development guidelines
  - Unit Testing
  - Integration Testing
  - Functional Testing
  - Acceptance Testing
- The OHDSI community's computer scientists and IT professionals test their code and then have it evaluated by the community

Estimation methods

Cohort Method

New-user cohort studies using large-scale regression for propensity and outcome models

Self-Controlled Case Series

Self-Controlled Case Series analysis using few or many predictors, includes splines for age and seasonality.

Self-Controlled Cohort

A self-controlled cohort design, where time preceding exposure is used as control.

IC Temporal Pattern Disc.

A self-controlled design, but using temporal patterns around other exposures and outcomes to correct for time-varying confounding.

Case-control

Case-control studies, matching controls on age, gender, provider, and visit date. Allows nesting of the study in another cohort.

Case-crossover

Case-crossover design including the option to adjust for time-trends in exposures (so-called case-time-control).

Method characterization

Empirical Calibration

Use negative control exposure-outcome pairs to profile and calibrate a particular analysis design.

Method Evaluation

Use real data and established reference sets as well as simulations injected in real data to evaluate the performance of methods.

Methods Packages at github.com/ohdsi

packages

Database Connector

Connect directly to a wide range of database platforms, including SQL Server, Oracle, and PostgreSQL.

Sql Render

Generate SQL on the fly for the various SQL dialects.

Cyclops

Highly efficient implementation of regularized logistic, Poisson and Cox regression.

Feature Extraction

Automatically extract large sets of features for user-specified cohorts using data in the CDM.

OHDSI METHODS LIBRARY: REGULATORY COMPLIANCE AND VALIDATION ISSUES

Meeting Compliance Requirements

This document will address specific areas within the entirety of software systems employed in observational studies: It is intended to provide a reasonable consensus position relative to the use of the OHDSI Methods Library and to provide a common foundation for end users to meet their own internal standard operating procedures, documentation requirements and potential future regulatory obligations.

[Put into the reference list:

- 21 CFR Part 11 - Electronic Records; Electronic Signatures
- Guidance for Industry: Part 11, Electronic Records; Electronic Signatures - Scope and Application
- Guidance for Industry - Computerized Systems Used in Clinical Investigations (2007)
- General Principles of Software Validation;
- ICH E9 - Statistical Principles for Clinical Trials
- Guidance for Industry and FDA Staff - Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (2010)

Meeting Community Requirements

Cohort Exit Criteria Without Drug Concept Set #506

Open **ericaVoss** opened this issue 18 days ago · 0 comments

ericaVoss commented 18 days ago — Member

Expected behavior

If user fails to specify a drug concept set in Cohort Exit Criteria the UI should warn the user.

Actual behavior

User it allowed to run cohort and generation fails.
java.lang.RuntimeException: Drug Codeset ID can not be NULL.

Steps to reproduce behavior

1. Create Cohort
2. Add a cohort exit criteria "Cohort exit criteria based on the end of an era of persistent exposure to any drug within a defined concept set"
3. Set a persistence window (90) and surveillance window (30)
4. Do not set a Drug Concept Set
5. Generate
6. Run will immediately fail

Transparent Updates in Each Release

ATLAS Version 2.2.0 Release Notes
WebAPI Version 2.2.0 Release Notes

This latest release contains **20** feature enhancements and issue resolutions:


Error when creating criteria based on Observation Period

CIRCE UI Enhancements

PLP Specification Editor

UI functionality to choose collapse strategy exit criteria

Fixed cohort and IR report bindings related to D3 v4.

care site entropy

Atlas Charts Upgrade to D3 v4 (#417)

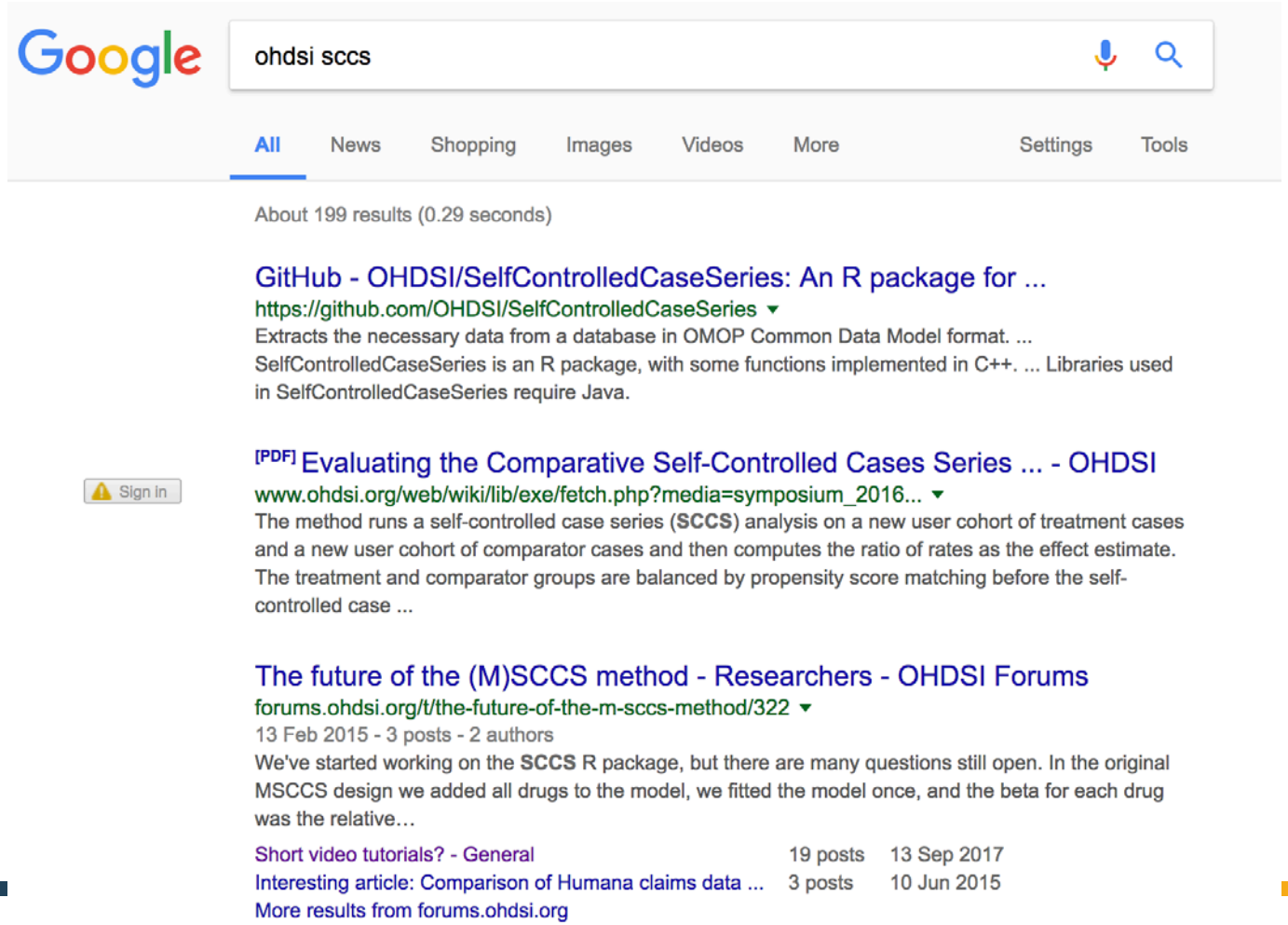
Improve client side caching reset

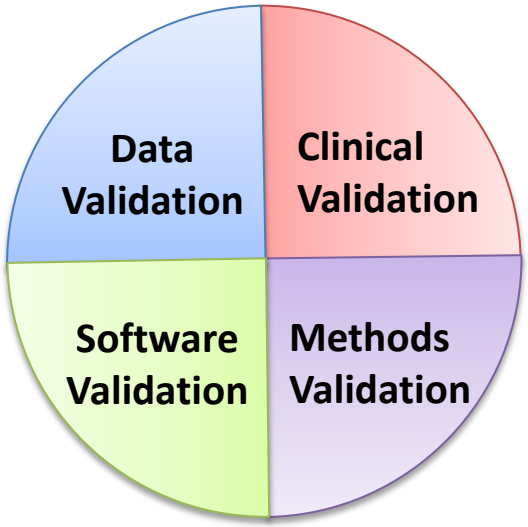
Databindings eventListener API change upate

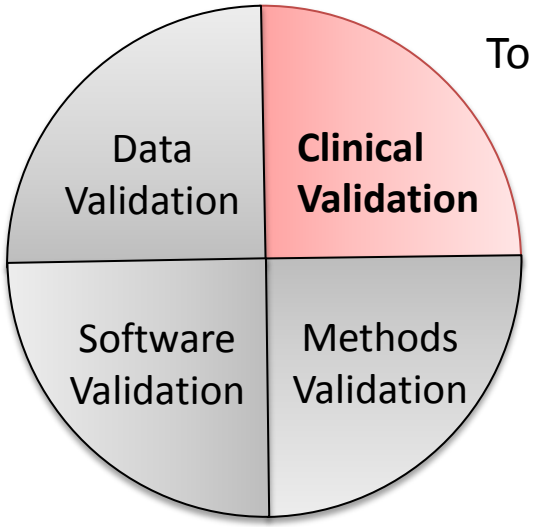
ATLAS/Reporting tab does not render

Want to know more about an OHDSI software package?

- Just Google it! It's easy to find!




Data Validation
Clinical Validation
Software Validation
Methods Validation



To what extent does the analysis conducted match the clinical intention?

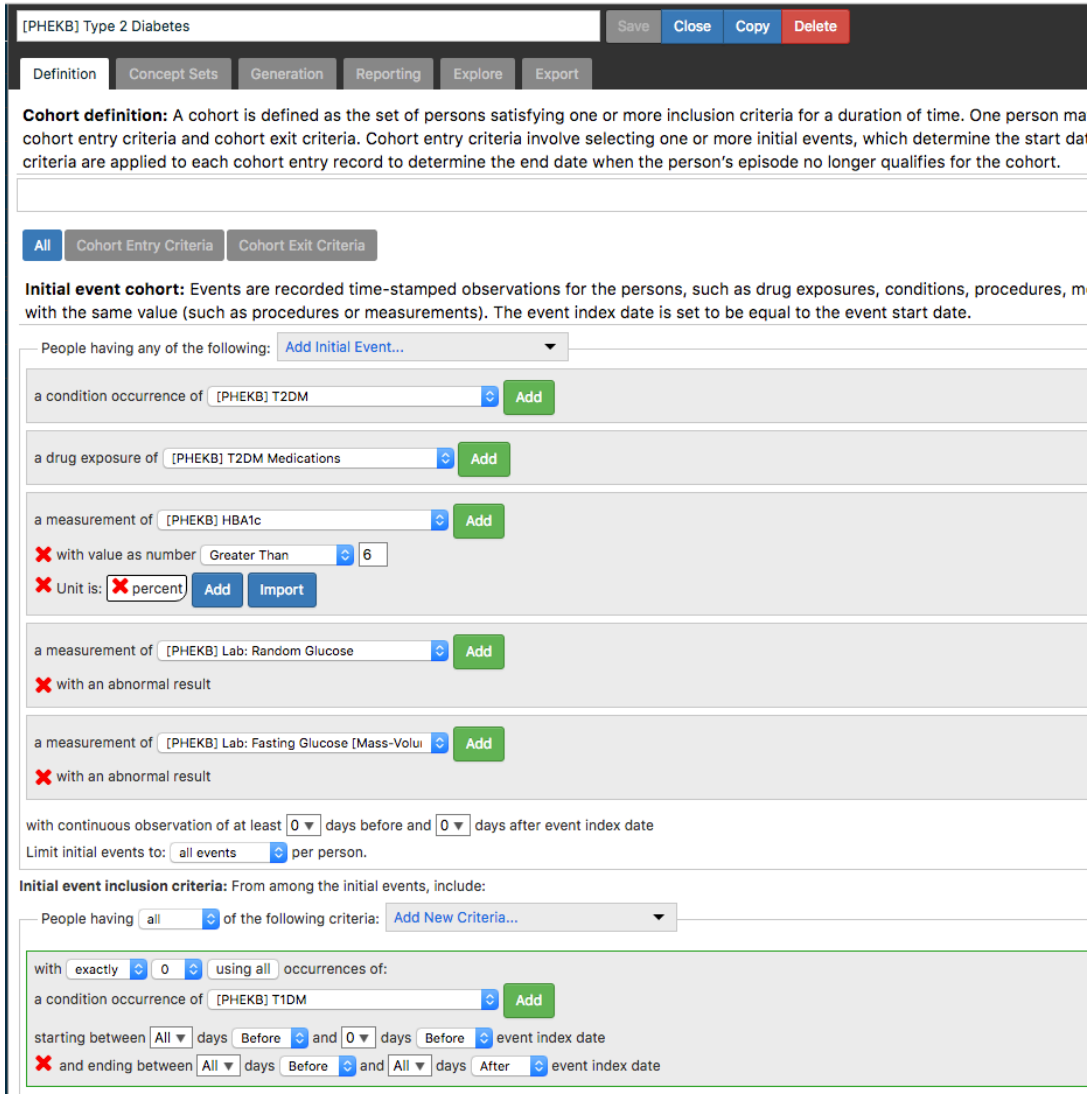
Clinical Validation

- Cohort definitions should be explicit and well-communicated for collaborator review prior to cohort generation
- Once cohorts have been generated, they must be validated in a standardized fashion

Cohort Definition Review

- OHDSI tools are designed to make clear and computable exactly what has been specified in a cohort definition

Create and Share via UI (even highly complex definitions)



Readable View

Initial Event Cohort

People having any of the following:

- a condition occurrence of [PHEKB] T2DM[6]
- a drug exposure of [PHEKB] T2DM Medications[7]
- a measurement of [PHEKB] HBA1c[1]
  - with value as number > 6
  - unit is any of: percent
- a measurement of [PHEKB] Lab: Random Glucose[3]
  - with an abnormal result
- a measurement of [PHEKB] Lab: Fasting Glucose [Mass-Volume][2]
  - with an abnormal result

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **all events per person.**

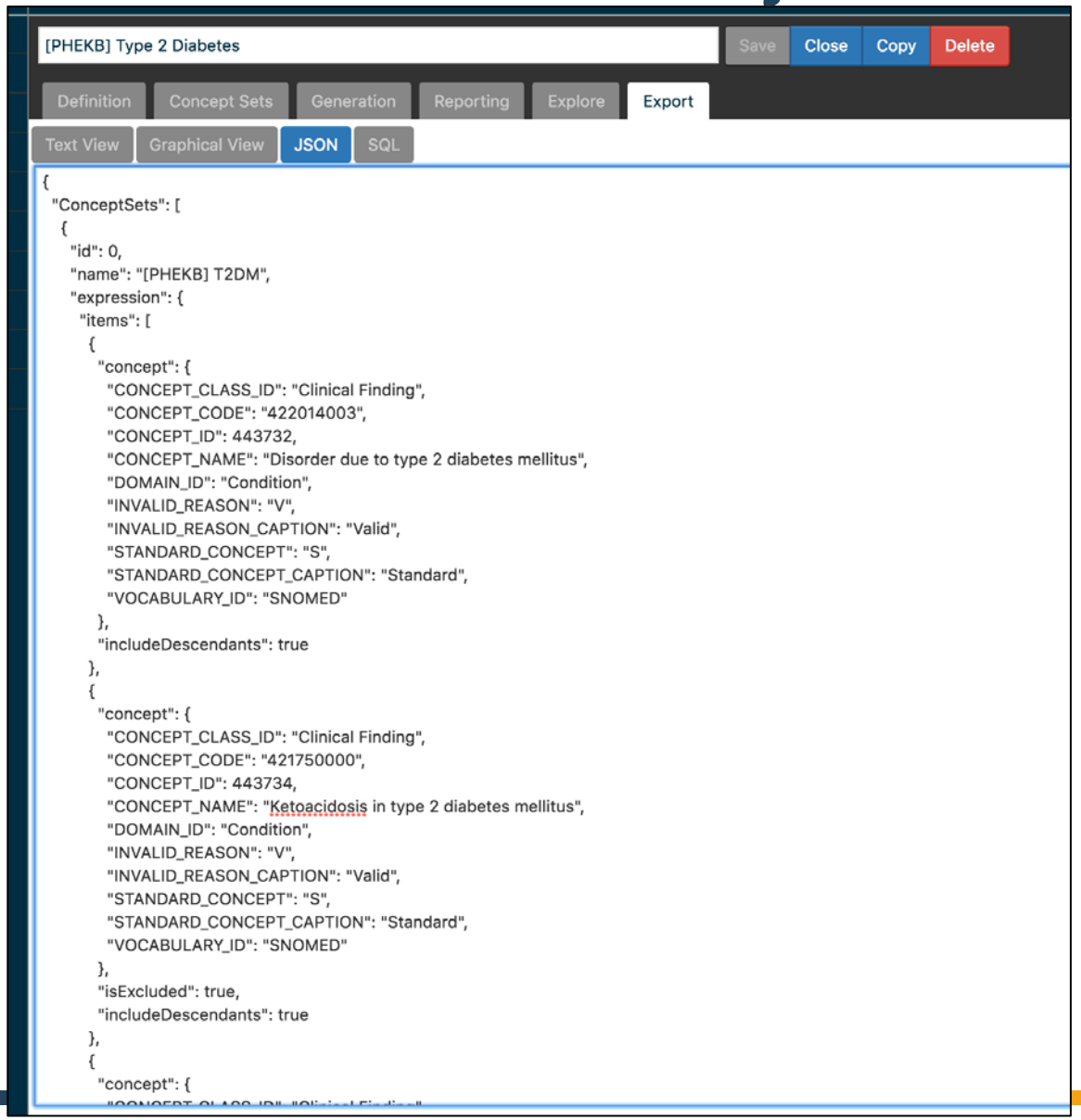
For people matching the Primary Events, include:
People having all of the following criteria:

- exactly 0 occurrences of a condition occurrence of [PHEKB] T1DM[4]
  starting between all days Before and 0 days Before event index date and ending between all days
- And people having any of the following criteria:
  - People having all of the following criteria:
    - at least 1 occurrences of a condition occurrence of [PHEKB] T2DM[6]
      starting between 0 days Before and 0 days After event index date and ending betwe
    - And people having any of the following criteria:
      - at least 1 occurrences of a drug exposure of [PHEKB] T2DM Medications[7]
        starting between all days Before and 1 days Before event index date and end
      - or at least 2 occurrences of a condition occurrence of [PHEKB] T2DM[6]
        starting between 0 days Before and all days After event index date and endi
      - Or people having all of the following criteria:
        - exactly 0 occurrences of a drug exposure of [PHEKB] T1DM Medicati
          starting between all days Before and 1 days Before event index date a
        - and exactly 0 occurrences of a drug exposure of [PHEKB] T2DM Med
          starting between all days Before and 1 days Before event index date a
        - And people having any of the following criteria:
          - at least 1 occurrences of a measurement of [PHEKB] HBA1c[1]
            - with an abnormal result

[PHEKB] T2DM Medications

Concept Id	Concept Name
1529331	Acarbose
1530014	Acetohexamide
1594973	Chlorpropamide
1583722	exenatide
1597756	glimepiride
1560171	Glipizide
1559684	Glyburide
1503297	Metformin
1510202	miglitol
1502826	nateglinide
1525215	pioglitazone
1516766	repaglinide
1547504	rosiglitazone
1580747	sitagliptin
1502809	Tolazamide
1515249	troglitazone

Shareable Object



Validate Generated Cohort

- Once the cohort has been generated, need to assess whether it indeed reflects the intended exposure or outcome of interest
- OHDSI tools provide support for both statistical and clinical review in a standardized fashion

Atlas Generates Comprehensive Cohort Statistics

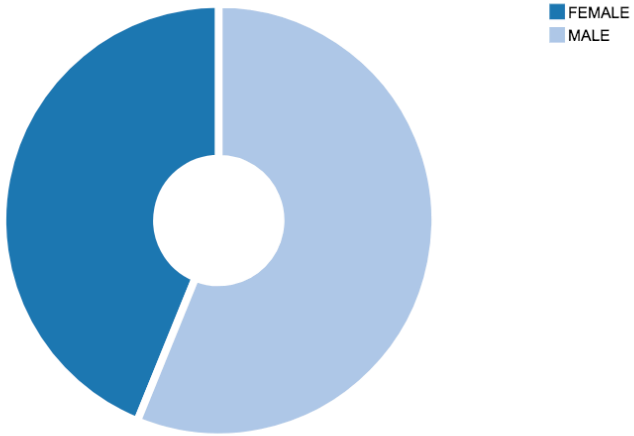
PTSD

Definition | Concept Sets | Generation | Reporting | Explore

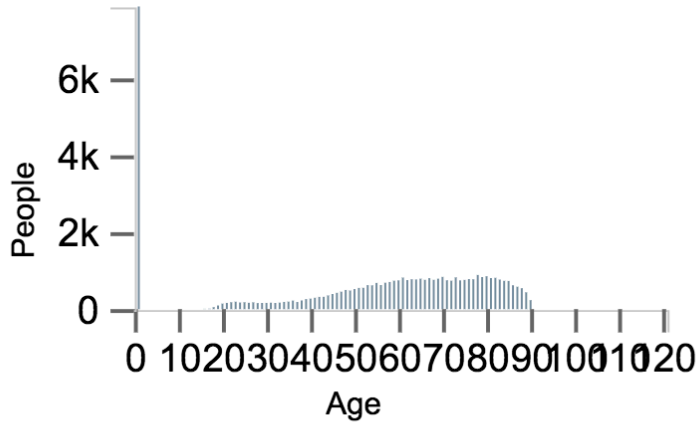
Report	MIMIC
Cohort Specific	view
Condition	view
Condition Eras	view
Conditions by Index	view
Death	
Drug Eras	view
Drug Exposure	view
Drugs by Index	view
Heracles Heel	view
Observation Periods	view
Person	view
Procedure	view
Procedures by Index	view
Data Completeness	
Entropy	

Generate | Export

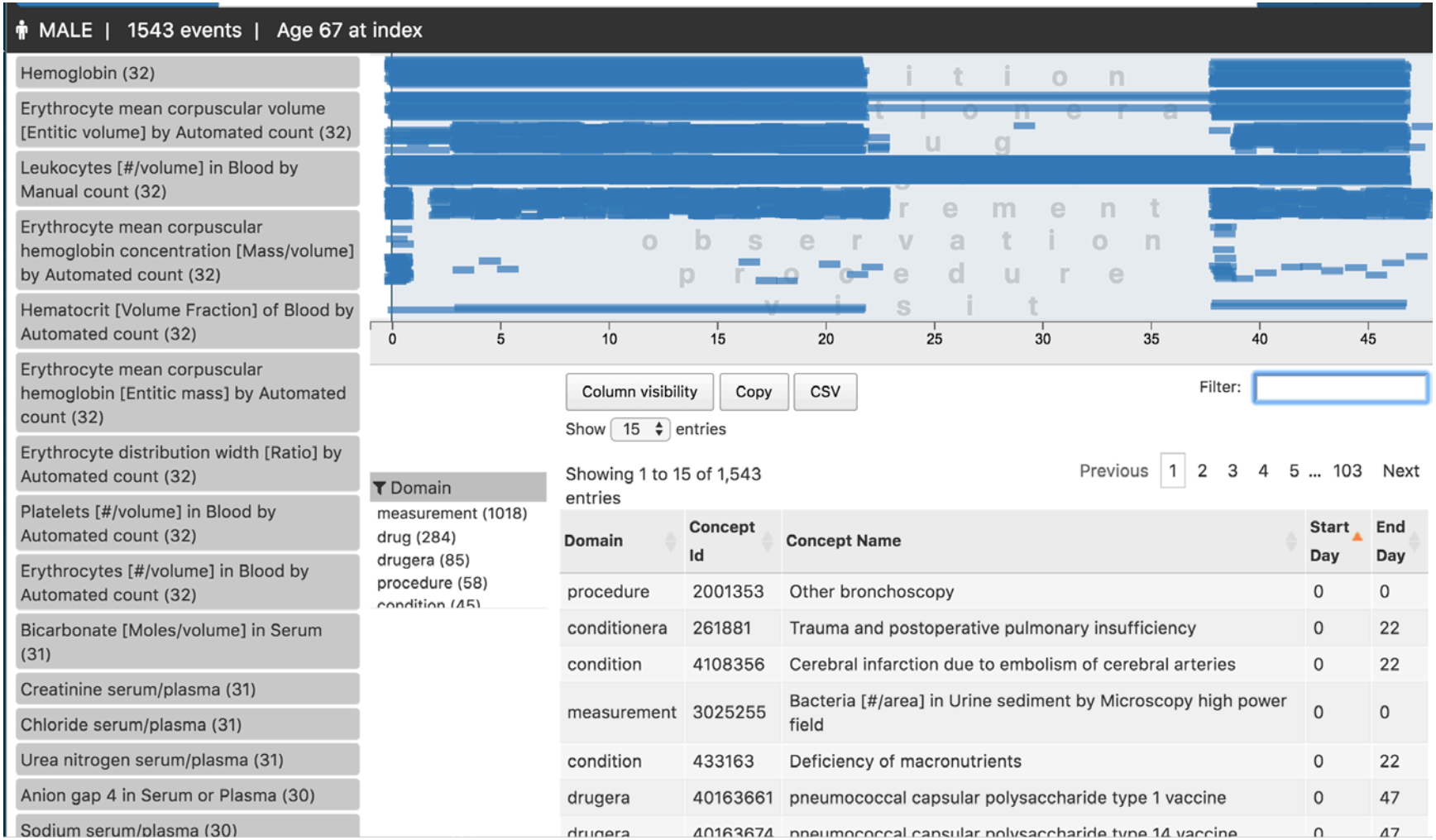
Population by Gender



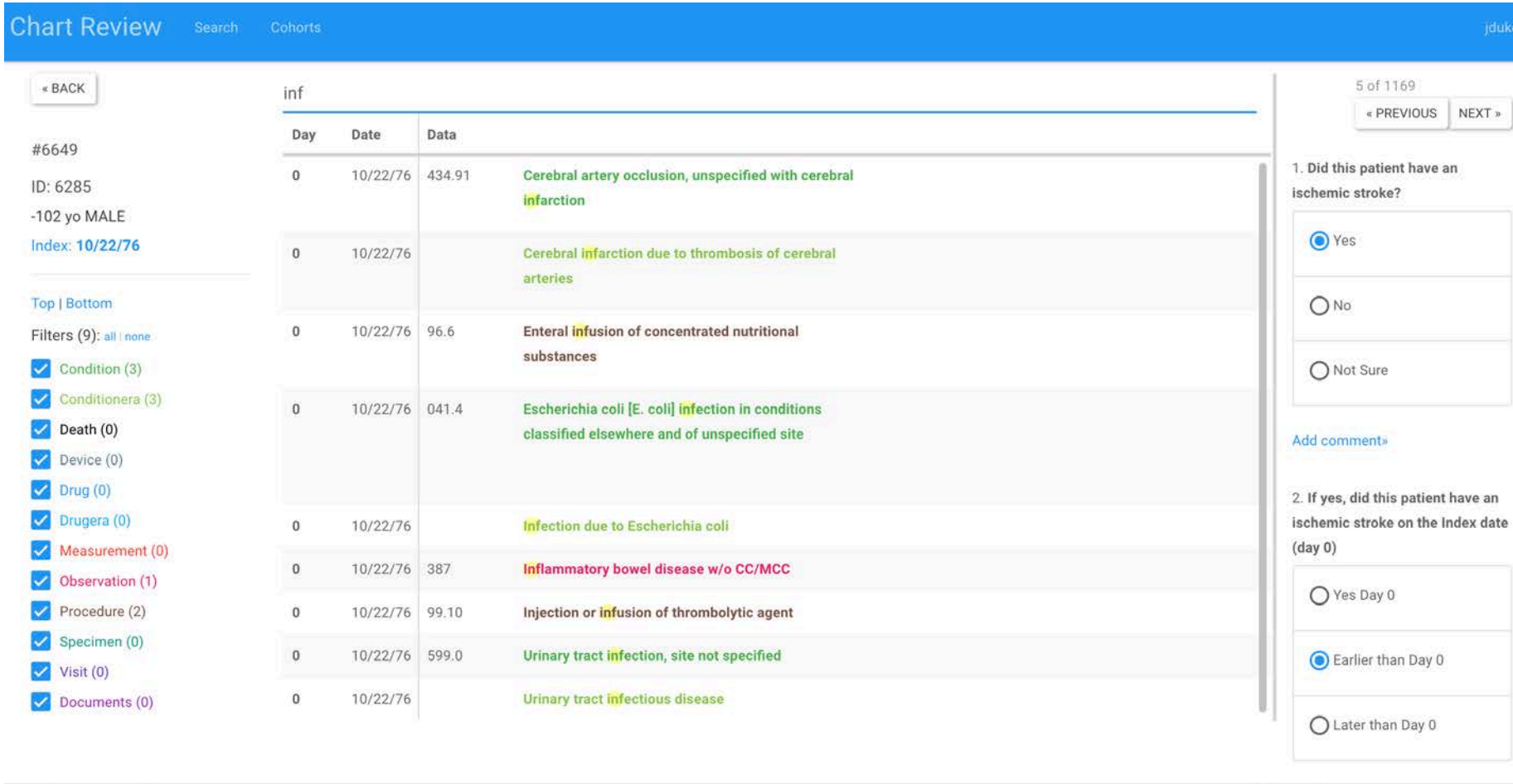
Age at First Observation



Atlas Supports De-identified Patient Level Chart Review

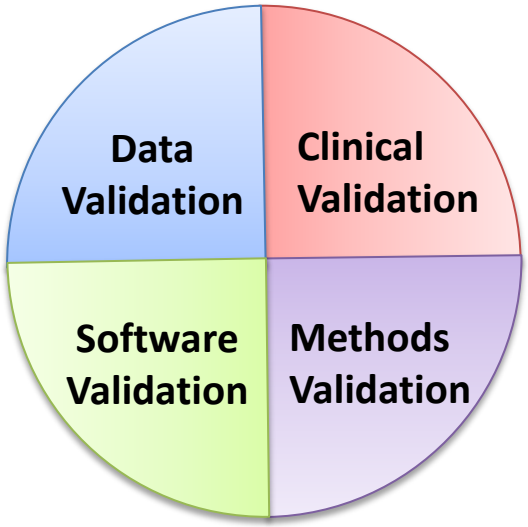


Standardized Manual Validation

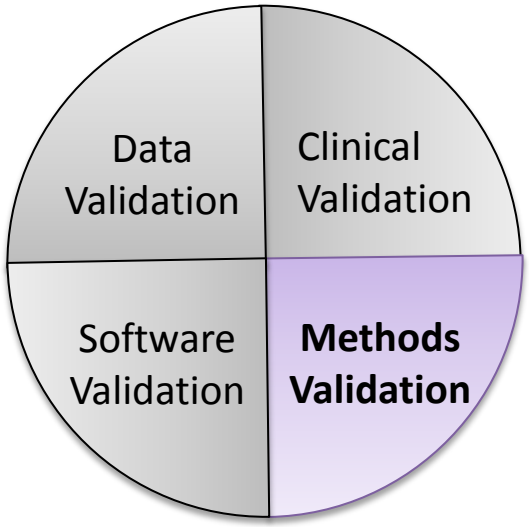


Standardized Manual Validation

- Each cohort can be annotated using a custom set of questions
- The ‘question set’ is a shareable object that can be sent in conjunction with the cohort definitions themselves, expediting validation
- Can automate calculation of phenotype performance measures with goal of expediting generation of cross-site performance metrics


Data Validation
Clinical Validation
Software Validation
Methods Validation

Validation Components



Do the estimates generated in an analysis measure what they purport to measure?

Methods Validation

- Methods Diagnostics
  - Is the planned statistical method valid given the data?
- Empirical Calibration
  - Are there systematic errors that require calibration of the results?

Keppra and Angioedema

Potential Signals of Serious Risks/New Safety Information Identified by the FDA Adverse Event Reporting System (FAERS) between October - December 2015

Keppra (levetiracetam) tablet, oral solution, injection	Angioedema	FDA is evaluating the need for regulatory action.

Request identified by OHDSI community member March 2016

Levetiracetam and Risk of Angioedema in patients with Seizure Disorder

Objective: To assess the risk between exposure to Keppra (levetiracetam) and angioedema.

Rationale: The Food and Drug Administration (FDA) has recently announced that they are evaluating the need for regulatory action regarding a potential association between exposure to the anti-seizure drug Keppra and angioedema. OHDSI seeks to support evidence generation for questions of importance to FDA and other stakeholders seeking to protect and promote the public's health.

Project Lead(s): Jon Duke, Patrick Ryan, Marc Suchard, George Hripcsak, [?Adler], Christian Reich, Yuriy Khoma, Marie-Sophie Schwalm, Yonghui Hu, [Stanford- Juan?], Martijn Schuemie.

Coordinating Institution(s): Regenstrief Institute / Georgia Tech

Participating Institution(s): Regenstrief Institute, Georgia Tech, Janssen Research and Development, Columbia University, University of California Los Angeles, University of Texas Houston, Stanford University, QuintilesIMS.

Full Protocol: Keppra and Angioedema Risk Protocol

Initial Proposal Date: 5/3/2016

Launch Date: 5/18/2016

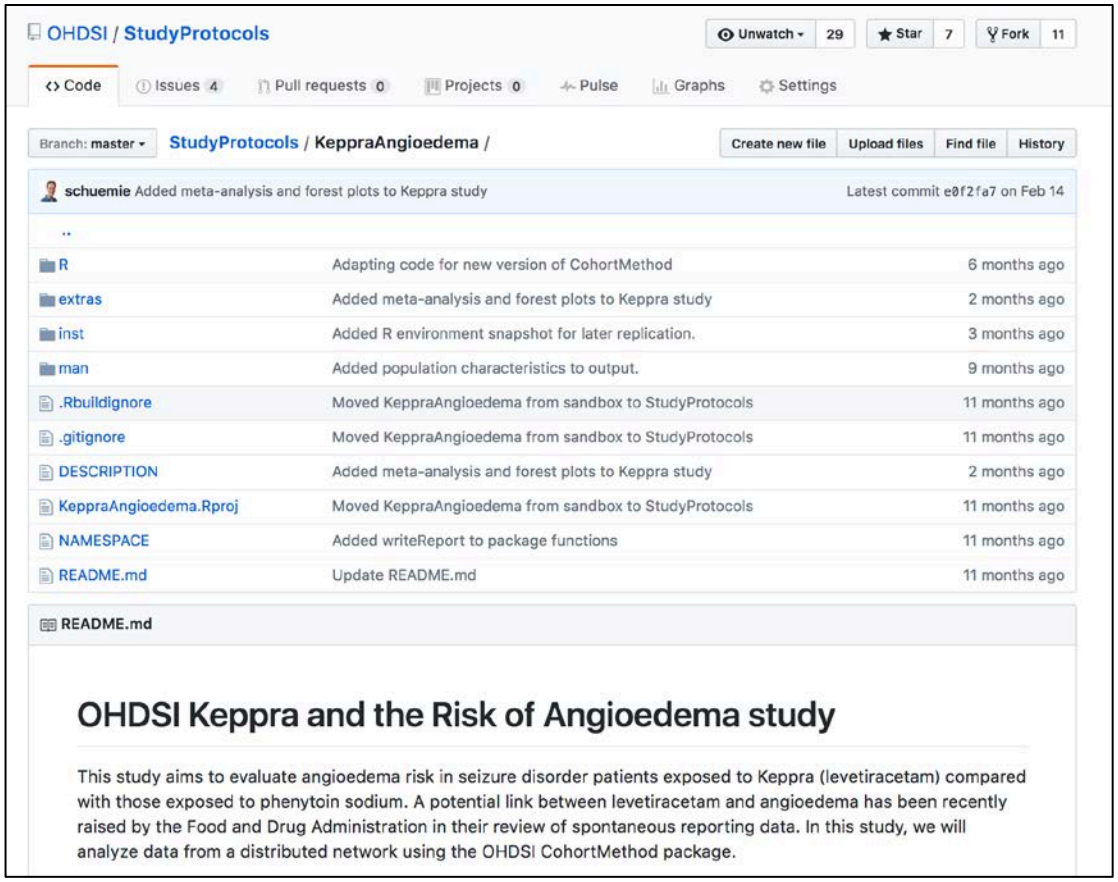
Receive Results for Analysis Date: 7/15/2016

Study Closure Date: 12/1/2016 (**Study closed**) Extended for 2 additional sites

Paper submitted 4/2017, Published 8/2017

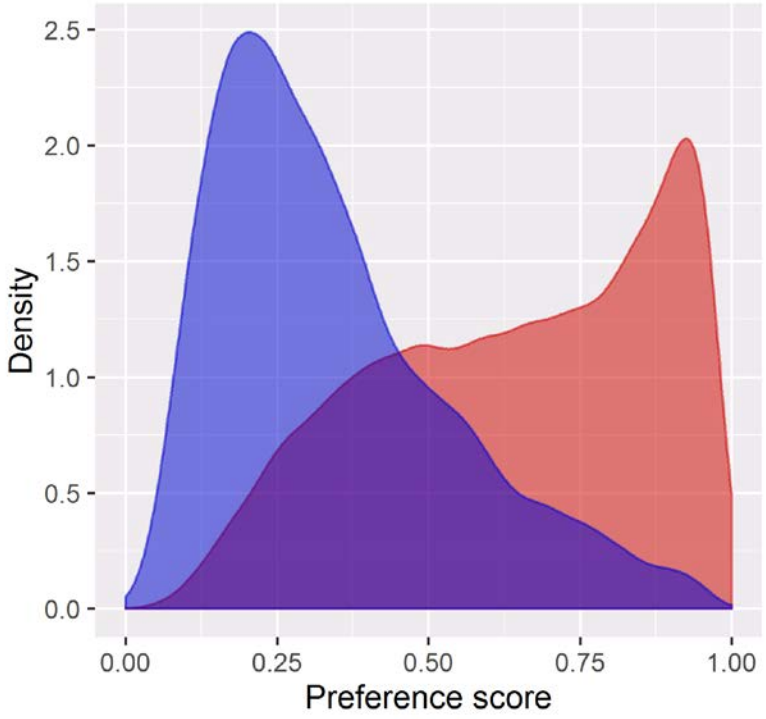
Leverage OHDSI CohortMethod

- Leveraged OHDSI CohortMethod R package
- Code tested at 2 sites prior to study start
- All code posted on GitHub

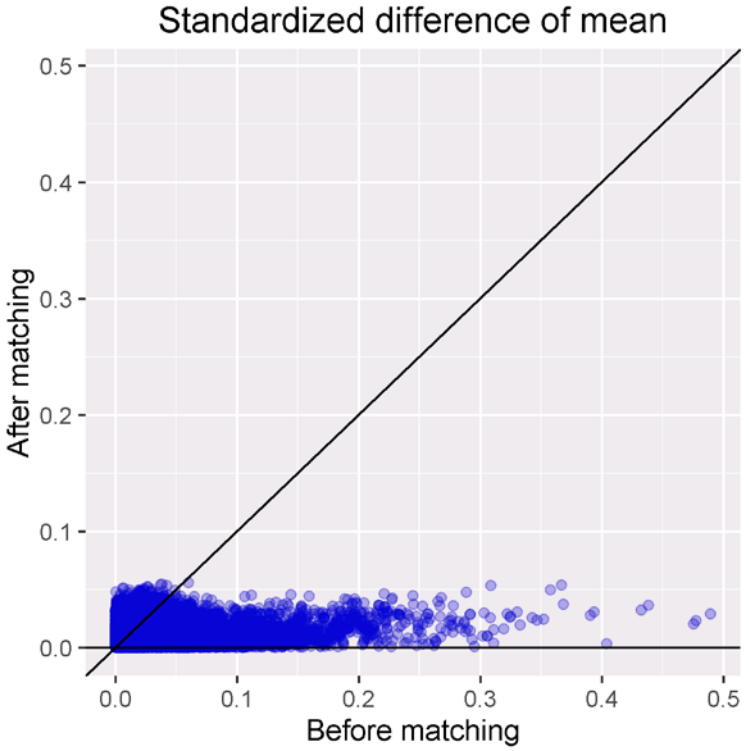


Package Provides Diagnostic Checks

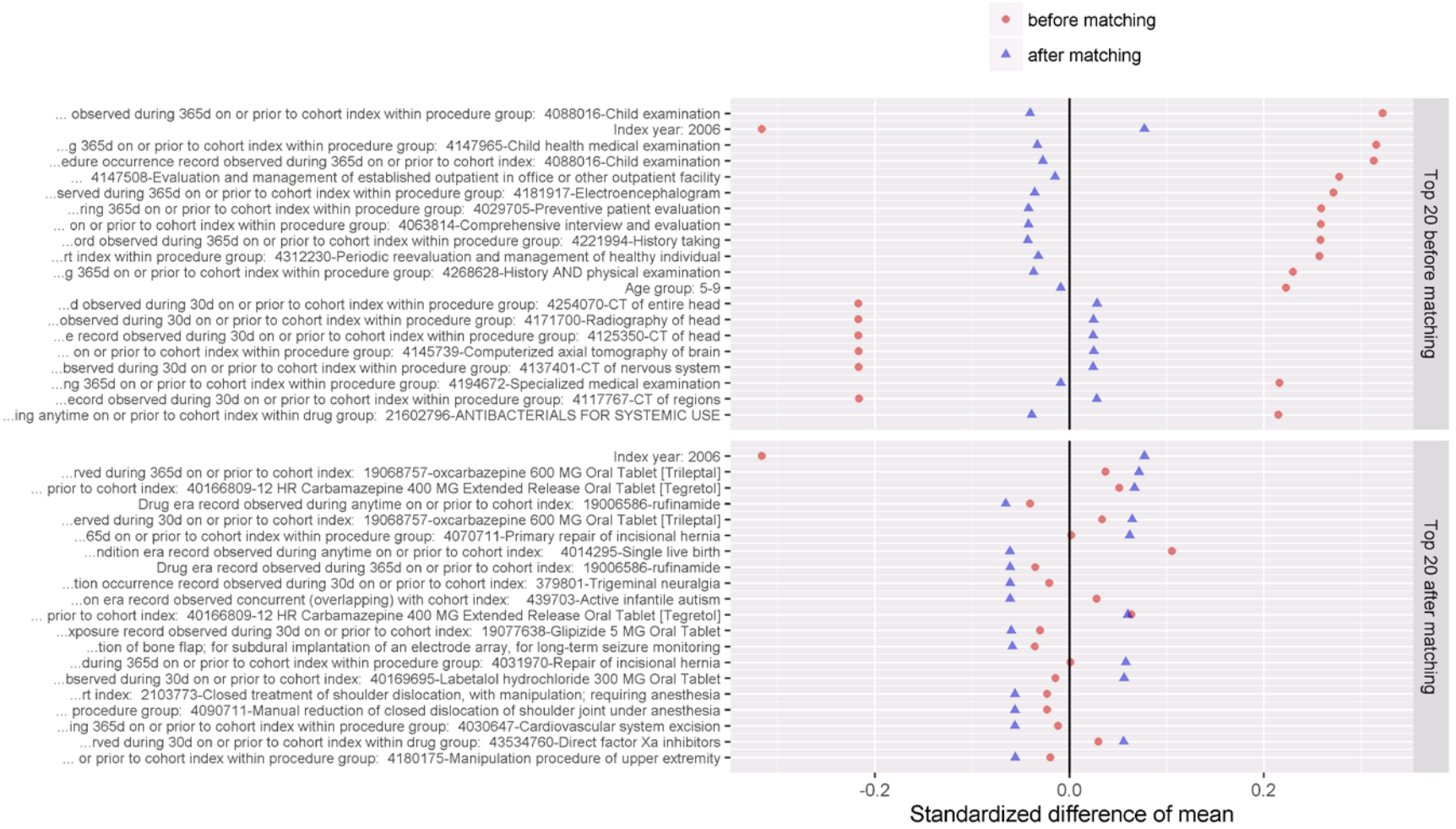
Propensity Score Distribution



Covariate Balance

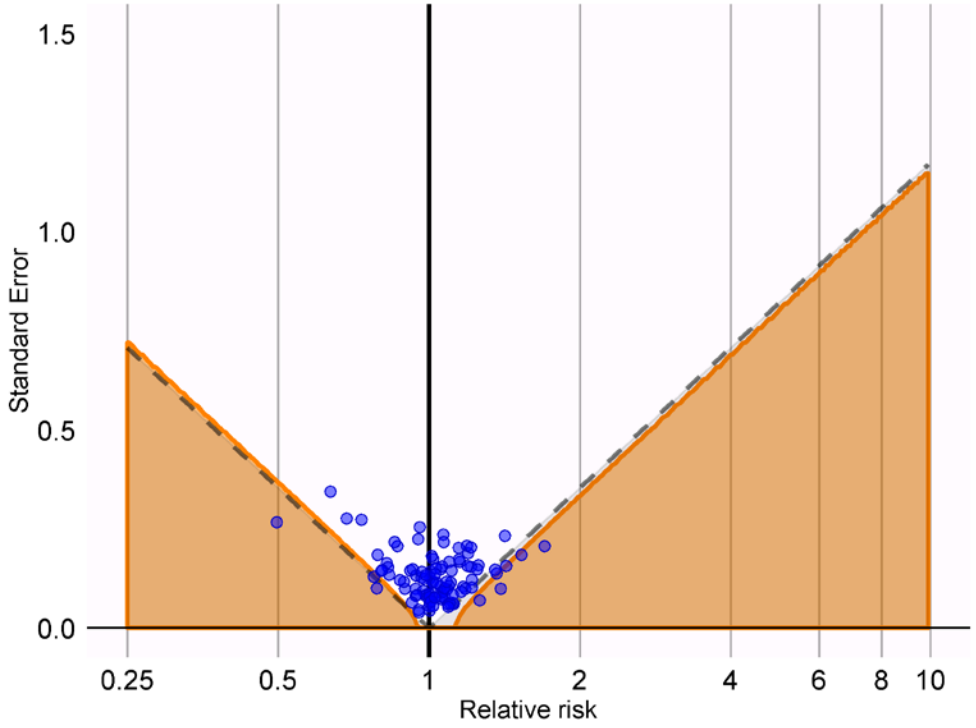


Covariate Balance



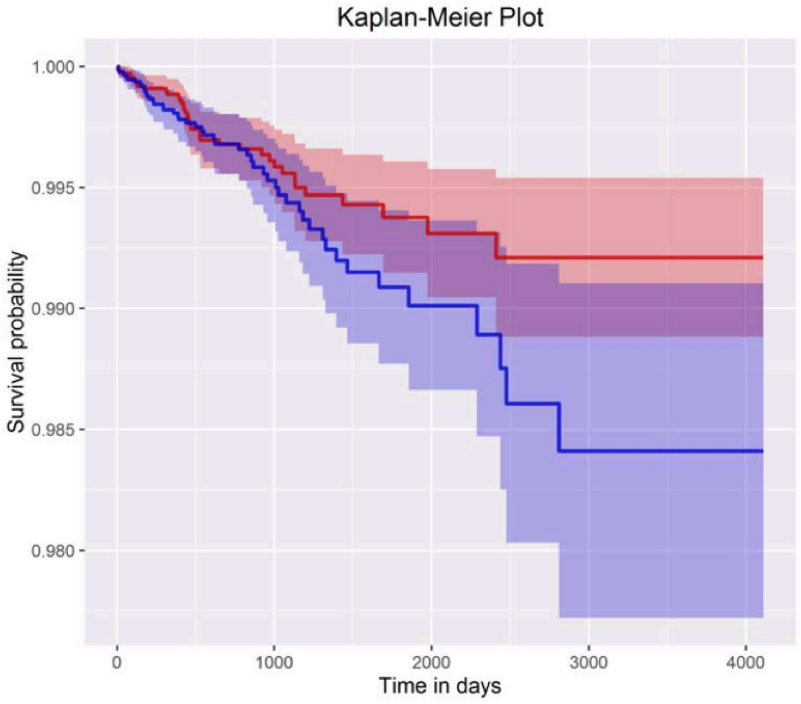
Empirical Calibration

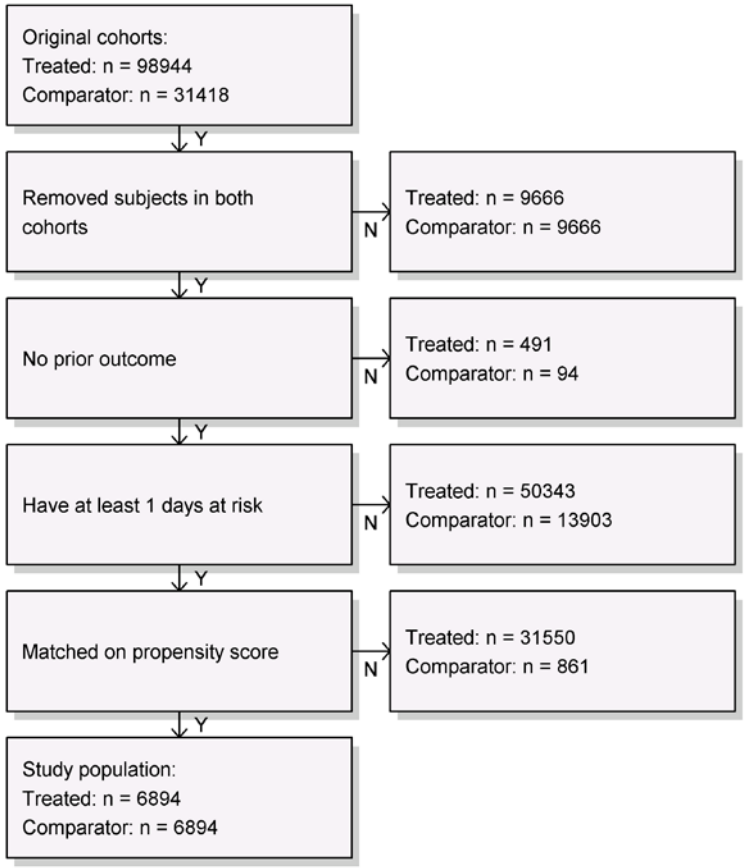
- Uses negative controls (conditions known not to be associated with the exposure) to calibrate p-values



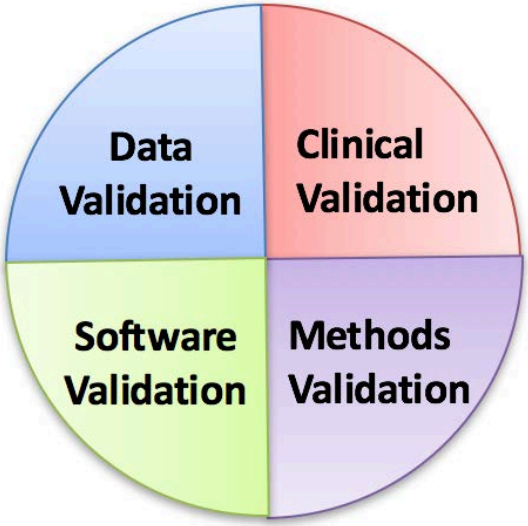
This study is well calibrated!

Generates Results and Supporting Documents





“Many eyes make all bugs shallow.”



Validation Requires Many Eyes

- Every aspect of validation is improved dramatically by transparency, openness to challenge, and engagement of a community that is passionate about making it better
- As OMOP adoption continues to grow rapidly from industry to academia to government (VA, DoD, CDC, FDA), our eye-count goes up and quality and innovation will continue to grow even further

Thanks!

jon.duke@gatech.edu