



# 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

---

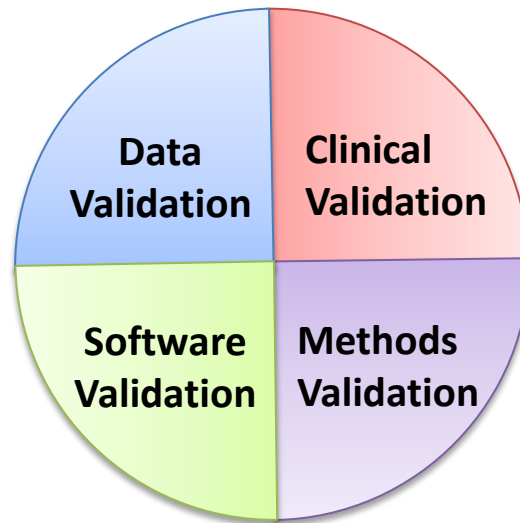


# What is Validation?

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



# Validation Components





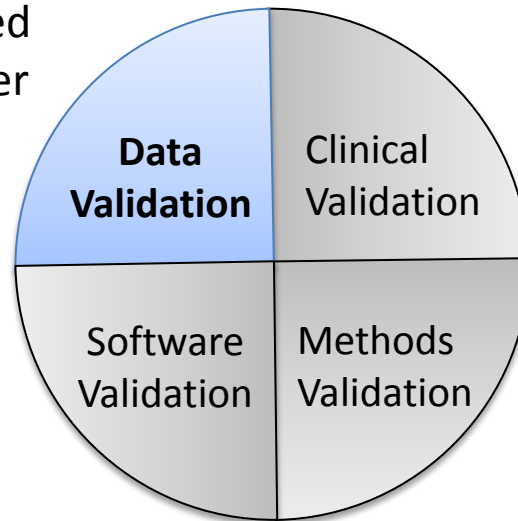
# Enabling Validation

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





# 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	41071	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?



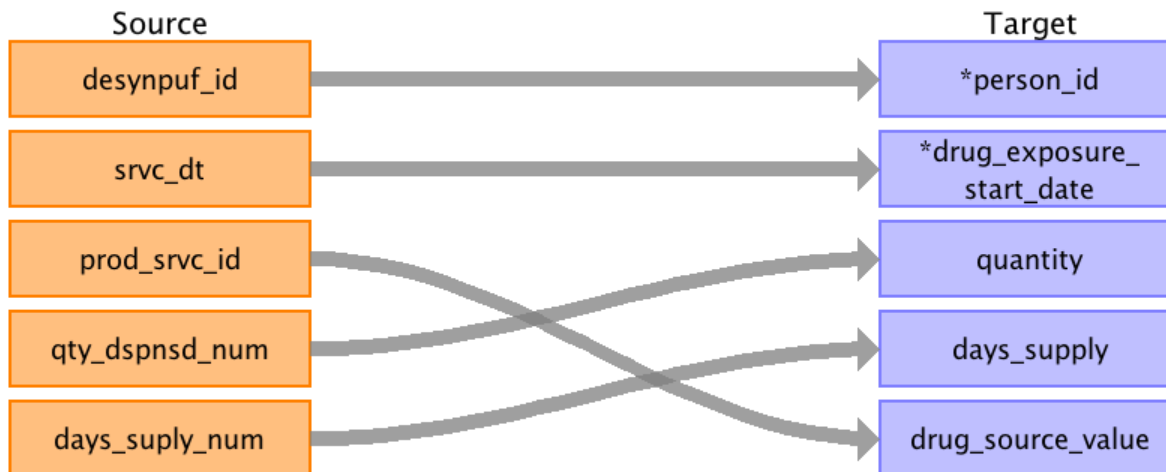
# Data Validation Tools

- 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

Overview Table

The screenshot displays the Usagi software interface. At the top, a table titled 'Overview Table' lists various source terms and their mappings to standard concepts. Below this, a 'Selected Mapping' window shows a detailed view of a specific mapping between a source code and a target concept. At the bottom, the 'Search Facility' is visible, including a query input field and several filter options.

Status	Source code	Source term	Frequency	Dutch term	Match score	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code
Unchecked	A	General and un...	26	Algemeen en ...	0.73	4244571	Generalized	Observation	Qualifier Value	SNOMED	9132005
Unchecked	B	Blood, bloodv.	4	Bloed en bloedr.	0.44	4133507	Organic medca...	Observation	Qualifier Value	SNOMED	278925002
Unchecked	D	Digestive	17	Tractus digesti...	1.00	436891	Adverse reacti...	Condition	Clinical Finding	SNOMED	218950001
Unchecked	F	Eye	218	Oog	1.00	373499	Disorder of eye...	Condition	Clinical Finding	SNOMED	371440005
Unchecked	H	Ear	87	Oor	1.00	4037611	Ear structure	Spec Anatomic	Body Structure	SNOMED	117590005
Unchecked	K	Cardiovascular	14	Tractus circula...	1.00	4014241	Structure of car...	Spec Anatomic	Body Structure	SNOMED	113257007
Unchecked	L	Musculoskeletal	38	Bewegingsapp.	0.93	4244662	Disorder of mu...	Condition	Clinical Finding	SNOMED	928009
Unchecked	N	Neurological	2044	Zinzesleutel	1.00	4192958	Neurology	Observation	Qualifier Value	SNOMED	394591006
Unchecked	P	Psychological	224	Psychische pro...	1.00	4249513	Psychologic	Observation	Qualifier Value	SNOMED	60224009
Unchecked	R	Respiratory	27	Tractus respir.	0.86	4024607	Respiratory fun...	Condition	Clinical Finding	SNOMED	196148009
Unchecked	S	Skin	59	Huid en subcutis	1.00	200174	Disorder of skin...	Condition	Clinical Finding	SNOMED	80659006

Source code	Source term	Frequency	Algemeen en un...
A	General and unspecified	26	Algemeen en un...

target concepts	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code
generalized	4244571	Generalized	Observation	Qualifier value	SNOMED	60132005

Search Facility

Query:

Filters

- Filter by automatically select concepts
- Filter by concept class: Admin Concept
- Filter by domain: Condition
- Filter by vocabulary: APC
- Filter invalid concepts

Results

Score	Synonym	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason
0.70	generalized	4244571	Generalized	Observation	Qualifier value	SNOMED	60132005	19700101	20991231	
0.87	Generalized and...	321882	Generalized and...	Condition	Clinical Finding	SNOMED	39823006	19700101	20991231	
0.59	Unspecified	4501594	Non-specific	Observation	Qualifier Value	SNOMED	10003008	19700101	20991231	
0.57	Generally contraindicated	444293	Contracted pelvis	Condition	Clinical Finding	SNOMED	871005	19700101	20991231	
0.55	General observat...	4021181	General observat...	Procedure	Procedure	SNOMED	225414002	19700101	20991231	
0.54	Production super...	4103933	Production super...	Observation	Social	SNOMED	25648006	19700101	20991231	

Selected Mapping

Search Facility



# 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))



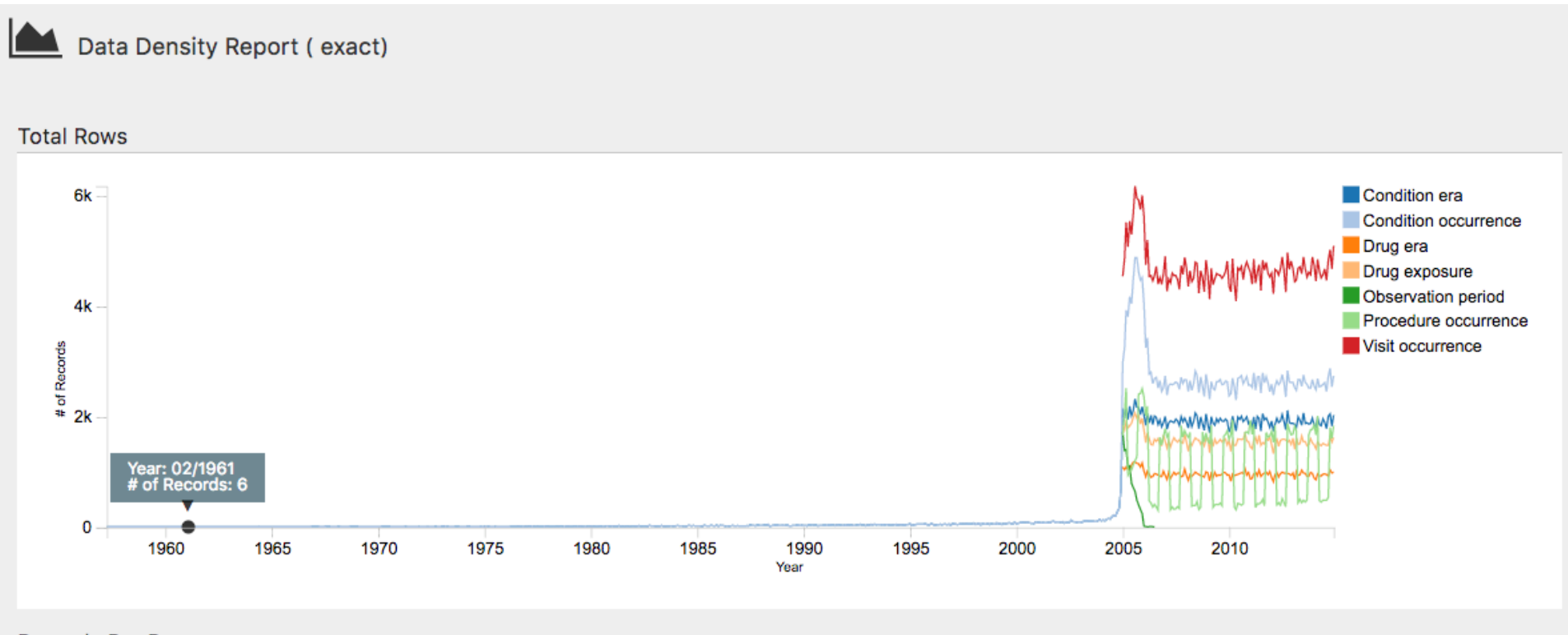
# Example Conformance Checks

```
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
```



# Post-ETL Validation

- Use ACHILLES visualizations to review anomalies







# Post-ETL Validation

- Use ACHILLES visualizations to review anomalies



Data Sources ▾ Reports ▾

## OPTUM

### Drug Era Report

#### Drug Prevalence

Treemap [Table](#)

BLOOD AND BLOOD FORMING ORGANS  
ANTITHROMBOTIC AGENTS  
VITAMIN K ANTAGONISTS

#### Warfarin

Prevalence: 0.91%  
Number of People: 368,077  
Length of Era: 193.07

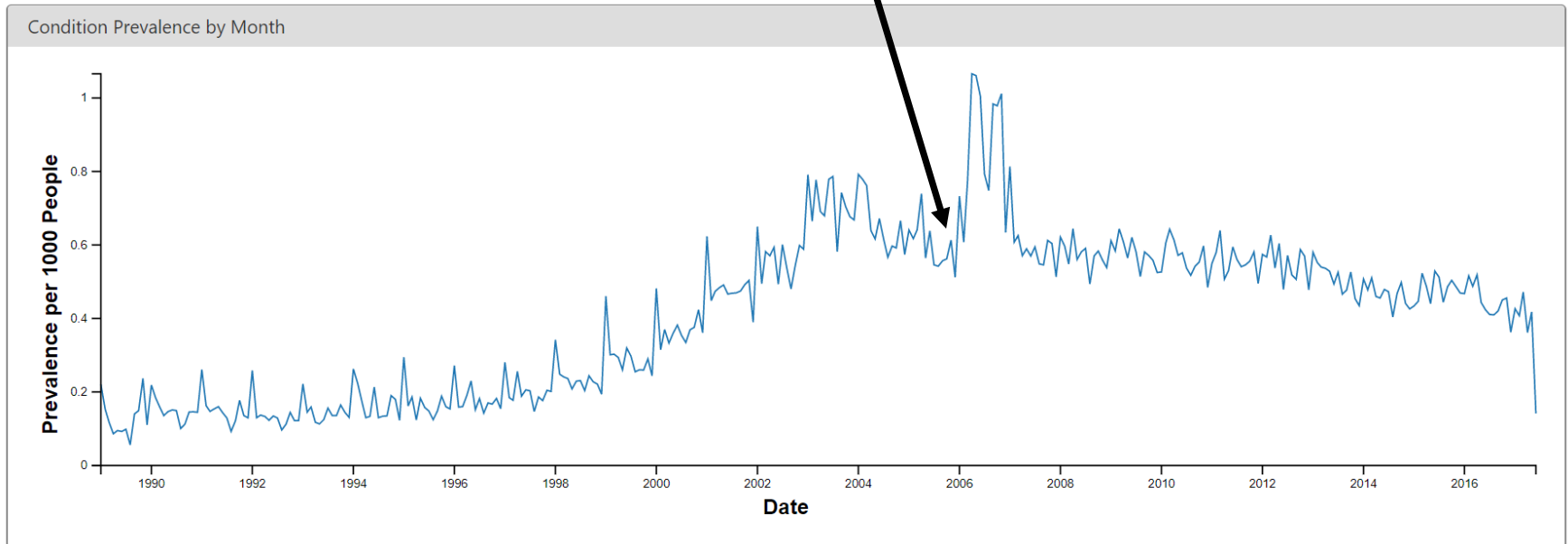
Box Size: Prevalence, Color: Records per Person (Blue to Orange = Low to High), Use Ctrl-Click to Zoom, Alt-Click to Reset Zoom



# Post-ETL Validation

- Use ACHILLES visualizations to review anomalies

Diabetes reimbursement change in UK





# 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
  - 2<sup>nd</sup>-level review by independent coding manager
- Quality and change evaluation scripts run prior to each release
- OHDSI Community review



# Rapidly Reviewable



SEARCH

SEARCH BY KEYWORD

atrial fibrillation

atrial fibrillation x

ICD9CM x

ICD10 x

Valid x

DOMAIN ▼

STANDARD CONCEPT ▼

CLASS ▼

VOCABULARY ▲

- HCPCS (3)
- ICD10 (5)
- ICD10CM (6)
- ICD9CM (2)
- Indication (12)
- LOINC (11)

DOWNLOAD RESULTS

ID ▼	CODE ▼	NAME ▼	CLASS ▼	CONCEPT ▼
44821957	427.31	Atrial fibrillation	5-dig billing code	Non-standard
44824248	427.3	Atrial fibrillation and flutter	4-dig nonbill code	Non-standard
45548021	I48.1	Persistent atrial fibrillation	ICD10 code	Non-standard
45581776	I48.0	Paroxysmal atrial fibrillation	ICD10 code	Non-standard
45591467	I48.2	Chronic atrial fibrillation	ICD10 code	Non-standard
45596206	I48	Atrial fibrillation and flutter	ICD10 Hierarchy	Non-standard
45755409	I48.9	Atrial fibrillation and atrial flutter, unspecified	ICD10 code	Non-standard



# 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 Ostropelets

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

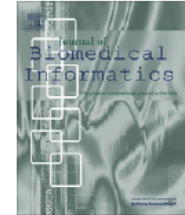


ELSEVIER

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: [www.elsevier.com/locate/yjbin](http://www.elsevier.com/locate/yjbin)



Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases <sup>☆</sup>

Christian Reich <sup>a,\*</sup>, Patrick B. Ryan <sup>a,b,1</sup>, Paul E. Stang <sup>a,b,1</sup>, Mitra Rocca <sup>c,2</sup>

<sup>a</sup> Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, 9650 Rockville Pike, Bethesda, MD 20814, USA

<sup>b</sup> Janssen Research & Development, LLC, 1125 Trenton-Harbourton Road, PO Box 200, MS K304, Titusville, NJ 08560, USA

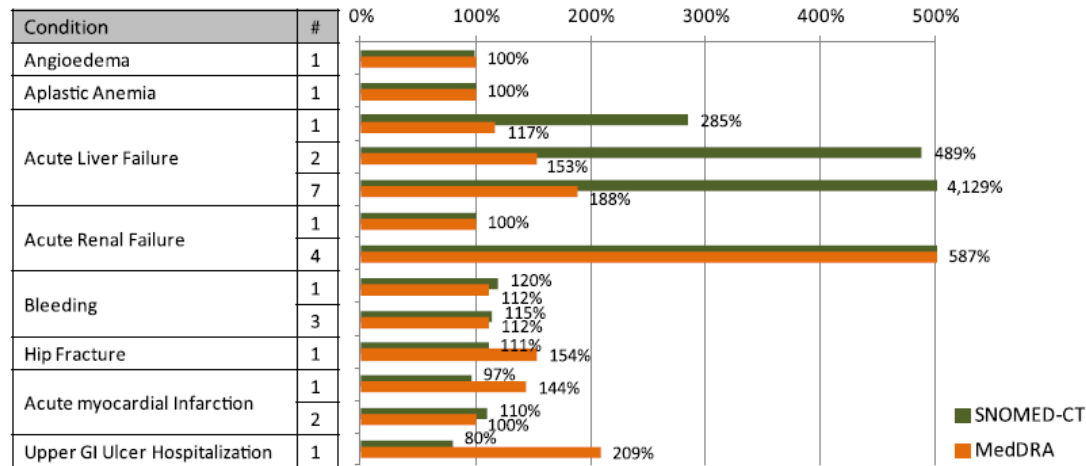
<sup>c</sup> Office of Translational Sciences, Center for Drug Evaluation and Research (CDER), US Food and Drug Administration, 10903 New Hampshire Ave., Bldg. 21, Rm. 4608, Silver Spring, MD 20933, USA



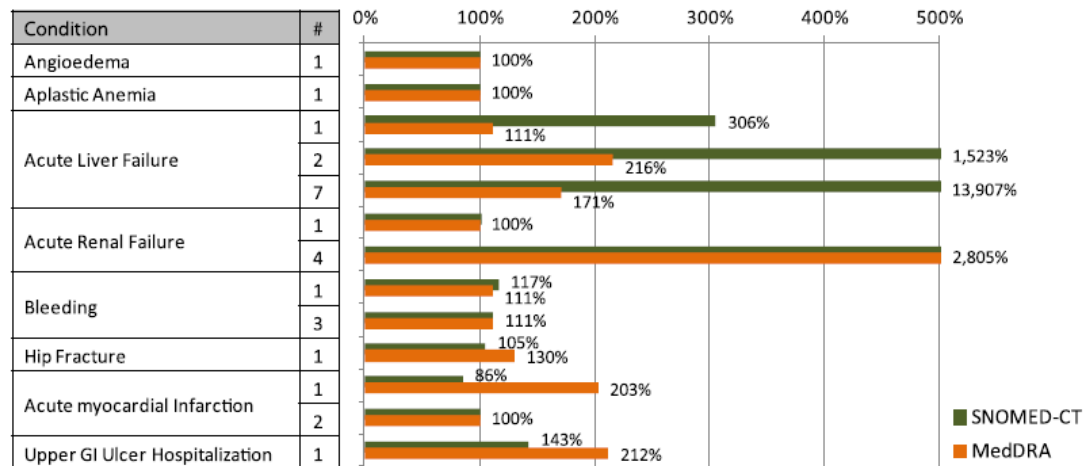


# Changing language may change your code list, which may change your cohort depending on the disease...

Cohort size of HOI in MSLR for different terminologies



Cohort size of HOI in GE for different terminologies





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

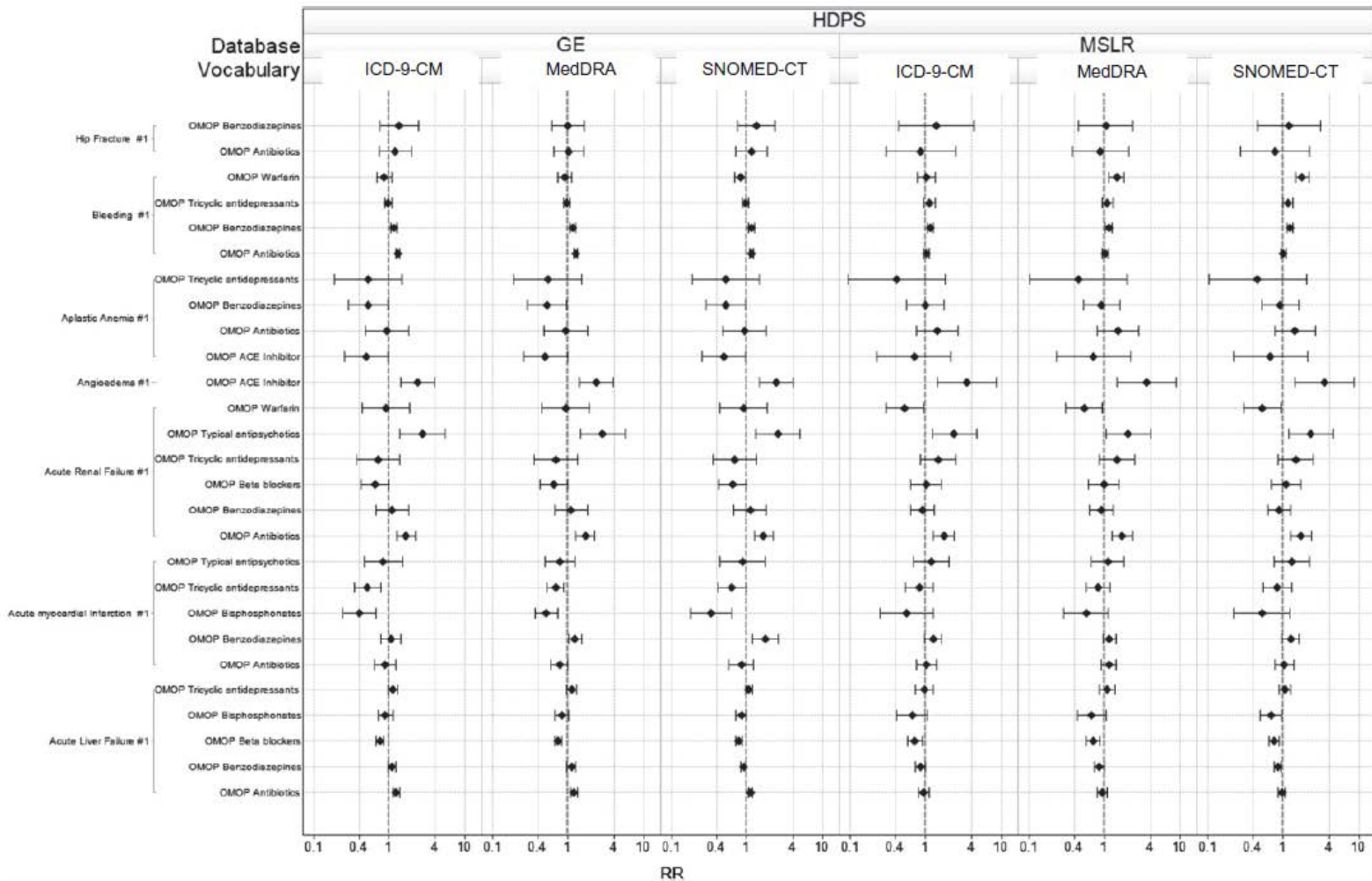


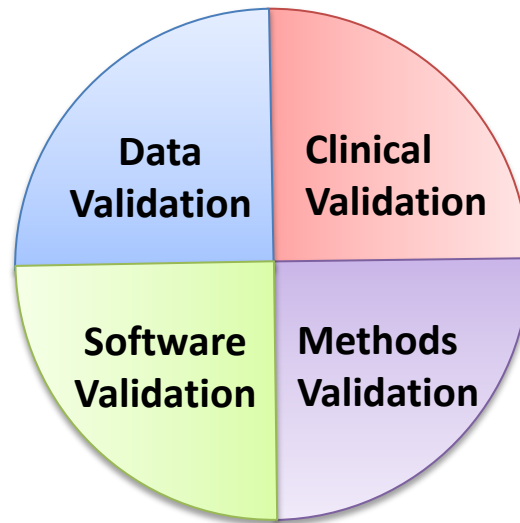
Fig. 3. Effect estimates and 95% confidence intervals for incident user design applied to MSLR and GE using ICD-9-CM, SNOMED-CT, and MedDRA as standard terminologies. Each dot represents the estimate of the effect of an individual HOI-drug combination (on the X-axis).



# More Data Quality Questions?

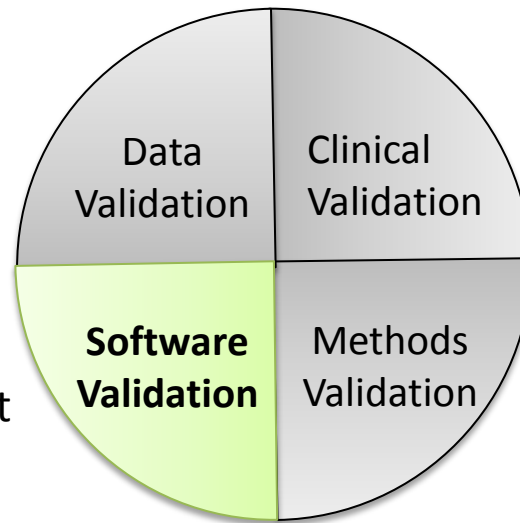
Check out our detailed FAQ at

[tinyurl.com/ohdsi-faq](https://tinyurl.com/ohdsi-faq)





# 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.

### 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.

packages

Methods Packages at  
[github.com/ohdsi](https://github.com/ohdsi)



# OHDSI METHODS LIBRARY: REGULATORY COMPLIANCE AND VALIDATION ISSUES

<b>What is the OHDSI Methods Library?</b>	<b>9</b>
Database connectivity	9
Large scale regularized regression	9
<b>Qualification and Validation of Systems for 21 CFR Part 11 Compliance</b>	<b>10</b>
Part 11: Electronic Records, Electronic Signatures	10
Validation	11
<b>Software Development Life Cycle (SDLC)</b>	<b>12</b>
Operational Overview	12
Source Code Management	12
Documentation	12
Testing and Validation	13
Availability of Current and Historical Archive Versions	13
Maintenance, Support and Retirement	13
Qualified Personnel	14
Physical and Logical Security	14
Disaster Recovery	14





# 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 is 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 update](#)



[ATLAS/Reporting tab does not render](#)



# Want to know more about an OHDSI software package?

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

The screenshot shows a Google search interface. The search bar contains the text "ohdsi sccs". Below the search bar, the "All" tab is selected. The search results show "About 199 results (0.29 seconds)". The first result is a GitHub repository: "GitHub - OHDSI/SelfControlledCaseSeries: An R package for ...". The second result is a PDF document: "[PDF] Evaluating the Comparative Self-Controlled Cases Series ... - OHDSI". The third result is a forum post: "The future of the (M)SCCS method - Researchers - OHDSI Forums".

Google

ohdsi sccs

All News Shopping Images Videos More Settings Tools

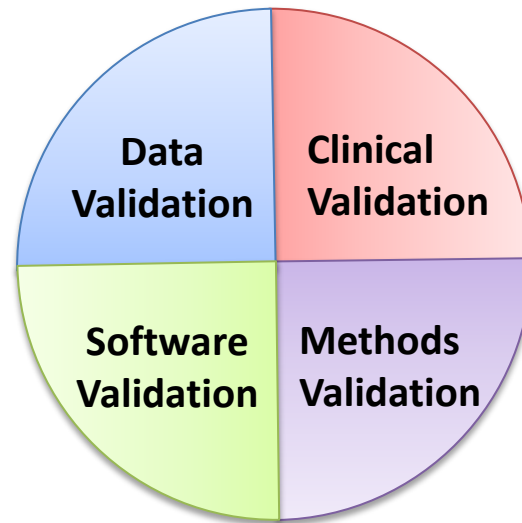
About 199 results (0.29 seconds)

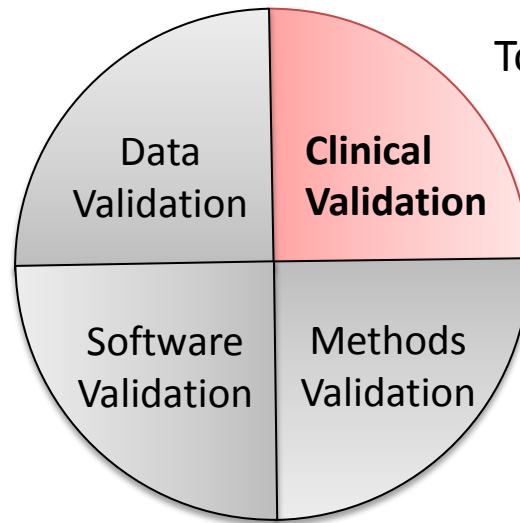
**GitHub - OHDSI/SelfControlledCaseSeries: An R package for ...**  
<https://github.com/OHDSI/SelfControlledCaseSeries> ▾  
Extracts the necessary data from a database in OMOP Common Data Model format. ...  
SelfControlledCaseSeries is an R package, with some functions implemented in C++. ... Libraries used in SelfControlledCaseSeries require Java.

**[PDF] Evaluating the Comparative Self-Controlled Cases Series ... - OHDSI**  
[www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=symposium\\_2016...](http://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=symposium_2016...) ▾  
The method runs a self-controlled case series (SCCS) analysis on a new user cohort of treatment cases and a new user cohort of comparator cases and then computes the ratio of rates as the effect estimate. The treatment and comparator groups are balanced by propensity score matching before the self-controlled case ...

**The future of the (M)SCCS method - Researchers - OHDSI Forums**  
[forums.ohdsi.org/t/the-future-of-the-m-sccs-method/322](http://forums.ohdsi.org/t/the-future-of-the-m-sccs-method/322) ▾  
13 Feb 2015 - 3 posts - 2 authors  
We've started working on the SCCS R package, but there are many questions still open. In the original MSCCS design we added all drugs to the model, we fitted the model once, and the beta for each drug was the relative...

Short video tutorials? - General 19 posts 13 Sep 2017  
Interesting article: Comparison of Humana claims data ... 3 posts 10 Jun 2015  
More results from forums.ohdsi.org





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)

[PHEKB] Type 2 Diabetes Save Close Copy Delete

Definition Concept Sets Generation Reporting Explore Export

**Cohort definition:** A cohort is defined as the set of persons satisfying one or more inclusion criteria for a duration of time. One person may have multiple cohort entry criteria and cohort exit criteria. Cohort entry criteria involve selecting one or more initial events, which determine the start date. Cohort exit criteria are applied to each cohort entry record to determine the end date when the person's episode no longer qualifies for the cohort.

All Cohort Entry Criteria Cohort Exit Criteria

**Initial event cohort:** Events are recorded time-stamped observations for the persons, such as drug exposures, conditions, procedures, measurements, with the same value (such as procedures or measurements). The event index date is set to be equal to the event start date.

People having any of the following: Add Initial Event...

a condition occurrence of [PHEKB] T2DM Add

a drug exposure of [PHEKB] T2DM Medications Add

a measurement of [PHEKB] HBA1c Add

✗ with value as number Greater Than 6

✗ Unit is: percent Add Import

a measurement of [PHEKB] Lab: Random Glucose Add

✗ with an abnormal result

a measurement of [PHEKB] Lab: Fasting Glucose [Mass-Vol] Add

✗ with an abnormal result

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events to: all events per person.

**Initial event inclusion criteria:** From among the initial events, include:

People having all of the following criteria: Add New Criteria...

with exactly 0 using all occurrences of:

a condition occurrence of [PHEKB] T1DM Add

starting between All days Before and 0 days Before event index date

✗ and ending between All days Before and All days After event index date



# Readable View

## Initial Event Cohort

People having any of the following:

- a condition occurrence of [PHEKB] T2DM<sup>6</sup>
- a drug exposure of [PHEKB] T2DM Medications<sup>7</sup>
- a measurement of [PHEKB] HBA1c<sup>1</sup>
  - with value as number > 6
  - unit is any of: percent
- a measurement of [PHEKB] Lab: Random Glucose<sup>3</sup>
  - with an abnormal result
- a measurement of [PHEKB] Lab: Fasting Glucose [Mass-Volume]<sup>2</sup>
  - 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<sup>4</sup> starting between all days Before and 0 days Before event index date and ending between all days After and 0 days After event index date
- 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<sup>6</sup> starting between 0 days Before and 0 days After event index date and ending between all days After and 0 days After event index date
    - And people having any of the following criteria:
      - at least 1 occurrences of a drug exposure of [PHEKB] T2DM Medications<sup>7</sup> starting between all days Before and 1 days Before event index date and ending between all days After and 1 days After event index date
      - or at least 2 occurrences of a condition occurrence of [PHEKB] T2DM<sup>6</sup> starting between 0 days Before and all days After event index date and ending between all days After and 1 days After event index date
      - Or people having all of the following criteria:
        - exactly 0 occurrences of a drug exposure of [PHEKB] T1DM Medications starting between all days Before and 1 days Before event index date and ending between all days After and 1 days After event index date
        - and exactly 0 occurrences of a drug exposure of [PHEKB] T2DM Medications starting between all days Before and 1 days Before event index date and ending between all days After and 1 days After event index date
        - And people having any of the following criteria:
          - at least 1 occurrences of a measurement of [PHEKB] HBA1c<sup>1</sup> 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

te  
te



# Shareable Object

[PHEKB] Type 2 Diabetes Save Close Copy Delete

Definition Concept Sets Generation Reporting Explore Export

Text View Graphical View JSON SQL

```
{
  "ConceptSets": [
    {
      "id": 0,
      "name": "[PHEKB] T2DM",
      "expression": {
        "items": [
          {
            "concept": {
              "CONCEPT_CLASS_ID": "Clinical Finding",
              "CONCEPT_CODE": "422014003",
              "CONCEPT_ID": 443732,
              "CONCEPT_NAME": "Disorder due to type 2 diabetes mellitus",
              "DOMAIN_ID": "Condition",
              "INVALID_REASON": "V",
              "INVALID_REASON_CAPTION": "Valid",
              "STANDARD_CONCEPT": "S",
              "STANDARD_CONCEPT_CAPTION": "Standard",
              "VOCABULARY_ID": "SNOMED"
            },
            "includeDescendants": true
          },
          {
            "concept": {
              "CONCEPT_CLASS_ID": "Clinical Finding",
              "CONCEPT_CODE": "421750000",
              "CONCEPT_ID": 443734,
              "CONCEPT_NAME": "Ketoacidosis in type 2 diabetes mellitus",
              "DOMAIN_ID": "Condition",
              "INVALID_REASON": "V",
              "INVALID_REASON_CAPTION": "Valid",
              "STANDARD_CONCEPT": "S",
              "STANDARD_CONCEPT_CAPTION": "Standard",
              "VOCABULARY_ID": "SNOMED"
            },
            "isExcluded": true,
            "includeDescendants": true
          }
        ]
      }
    }
  ]
}
```



# 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

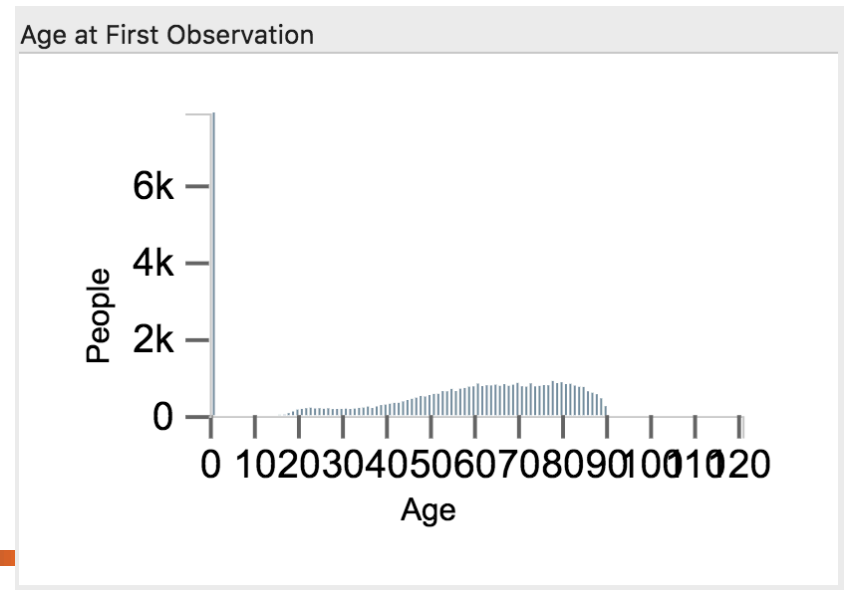
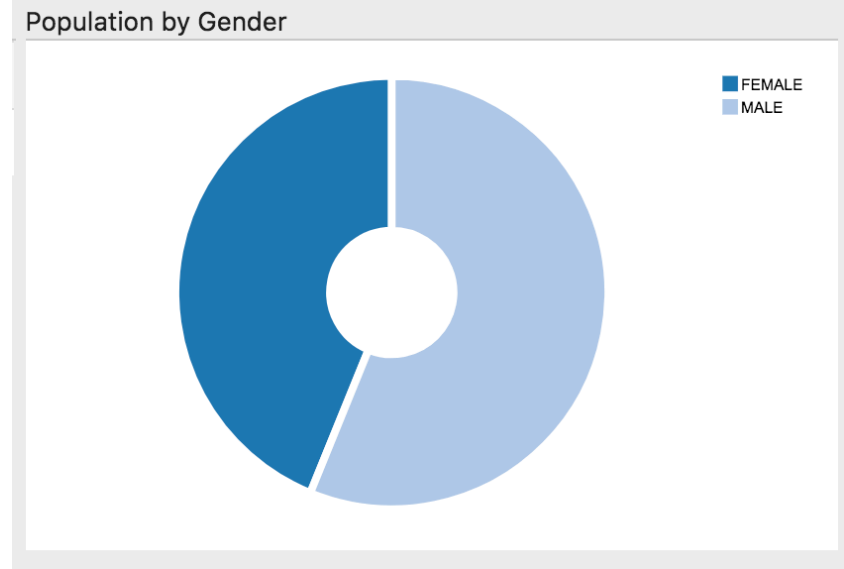
# Atlas Generates Comprehensive Cohort Statistics

PTSD

Definition Concept Sets Generation Reporting Explore

Report	MIMIC
Cohort Specific	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Condition	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Condition Eras	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Conditions by Index	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Death	<input type="checkbox"/>
Drug Eras	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Drug Exposure	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Drugs by Index	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Heracles Heel	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Observation Periods	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Person	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Procedure	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Procedures by Index	<input checked="" type="checkbox"/> <input type="checkbox"/> <a href="#">view</a>
Data Completeness	<input type="checkbox"/>
Entropy	<input type="checkbox"/>

[Generate](#) [Export](#)





# Atlas Supports De-identified Patient Level Chart Review

♂ MALE | 1543 events | Age 67 at index

Hemoglobin (32)

Erythrocyte mean corpuscular volume [Entitic volume] by Automated count (32)

Leukocytes [#./volume] in Blood by Manual count (32)

Erythrocyte mean corpuscular hemoglobin concentration [Mass/volume] by Automated count (32)

Hematocrit [Volume Fraction] of Blood by Automated count (32)

Erythrocyte mean corpuscular hemoglobin [Entitic mass] by Automated count (32)

Erythrocyte distribution width [Ratio] by Automated count (32)

Platelets [#./volume] in Blood by Automated count (32)

Erythrocytes [#./volume] in Blood by Automated count (32)

Bicarbonate [Moles/volume] in Serum (31)

Creatinine serum/plasma (31)

Chloride serum/plasma (31)

Urea nitrogen serum/plasma (31)

Anion gap 4 in Serum or Plasma (30)

Sodium serum/plasma (30)

0 5 10 15 20 25 30 35 40 45

Column visibility Copy CSV Filter:

Show 15 entries

Showing 1 to 15 of 1,543 entries

Previous **1** 2 3 4 5 ... 103 Next

Domain	Concept Id	Concept Name	Start Day	End Day
procedure	2001353	Other bronchoscopy	0	0
conditionera	261881	Trauma and postoperative pulmonary insufficiency	0	22
condition	4108356	Cerebral infarction due to embolism of cerebral arteries	0	22
measurement	3025255	Bacteria [#./area] in Urine sediment by Microscopy high power field	0	0
condition	433163	Deficiency of macronutrients	0	22
drugera	40163661	pneumococcal capsular polysaccharide type 1 vaccine	0	47
drugera	40163674	pneumococcal capsular polysaccharide type 14 vaccine	0	47

# Standardized Manual Validation

Chart Review

Search

Cohorts

jduke

« BACK

#6649

ID: 6285

-102 yo MALE

Index: 10/22/76

Top | Bottom

Filters (9): all | none

- Condition (3)
- Conditionera (3)
- Death (0)
- Device (0)
- Drug (0)
- Drugera (0)
- Measurement (0)
- Observation (1)
- Procedure (2)
- Specimen (0)
- Visit (0)
- Documents (0)

inf

Day	Date	Data	
0	10/22/76	434.91	Cerebral artery occlusion, unspecified with cerebral infarction
0	10/22/76		Cerebral infarction due to thrombosis of cerebral arteries
0	10/22/76	96.6	Enteral infusion of concentrated nutritional substances
0	10/22/76	041.4	Escherichia coli [E. coli] infection in conditions classified elsewhere and of unspecified site
0	10/22/76		infection due to Escherichia coli
0	10/22/76	387	Inflammatory bowel disease w/o CC/MCC
0	10/22/76	99.10	Injection or infusion of thrombolytic agent
0	10/22/76	599.0	Urinary tract infection, site not specified
0	10/22/76		Urinary tract infectious disease

5 of 1169

« PREVIOUS

NEXT »

1. Did this patient have an ischemic stroke?

Yes

No

Not Sure

[Add comment»](#)

2. If yes, did this patient have an ischemic stroke on the index date (day 0)

Yes Day 0

Earlier than Day 0

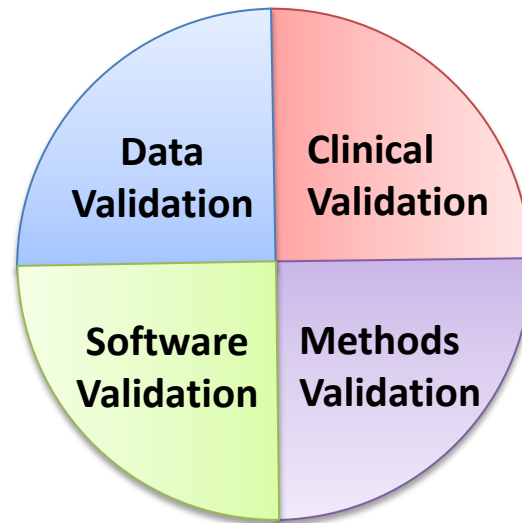
Later than Day 0



# Standardized Manual Validation

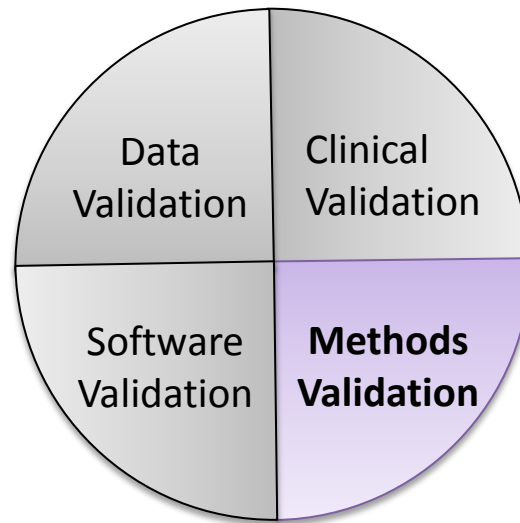
- Each cohort can be annotated using a custom set of questions
  - The 'question set' is a shareable object that can be sent in conjunction with the cohort definitions themselves, expediting validation
  - Can automate calculation of phenotype performance measures with goal of expediting generation of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







# 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

The screenshot shows the GitHub interface for the repository 'OHDSI / StudyProtocols'. The top navigation bar includes 'Unwatch' (29), 'Star' (7), and 'Fork' (11). Below the navigation bar, there are tabs for 'Code', 'Issues' (4), 'Pull requests' (0), 'Projects' (0), 'Pulse', 'Graphs', and 'Settings'. The current branch is 'master', and the repository path is 'StudyProtocols / KeppraAngioedema'. There are buttons for 'Create new file', 'Upload files', 'Find file', and 'History'. The commit history table shows the following entries:

Commit	Message	Time
schuemie	Added meta-analysis and forest plots to Keppra study	Latest commit e8f2fa7 on Feb 14
R	Adapting code for new version of CohortMethod	6 months ago
extras	Added meta-analysis and forest plots to Keppra study	2 months ago
inst	Added R environment snapshot for later replication.	3 months ago
man	Added population characteristics to output.	9 months ago
.Rbuildignore	Moved KeppraAngioedema from sandbox to StudyProtocols	11 months ago
.gitignore	Moved KeppraAngioedema from sandbox to StudyProtocols	11 months ago
DESCRIPTION	Added meta-analysis and forest plots to Keppra study	2 months ago
KeppraAngioedema.Rproj	Moved KeppraAngioedema from sandbox to StudyProtocols	11 months ago
NAMESPACE	Added writeReport to package functions	11 months ago
README.md	Update README.md	11 months ago

The README.md file content is as follows:

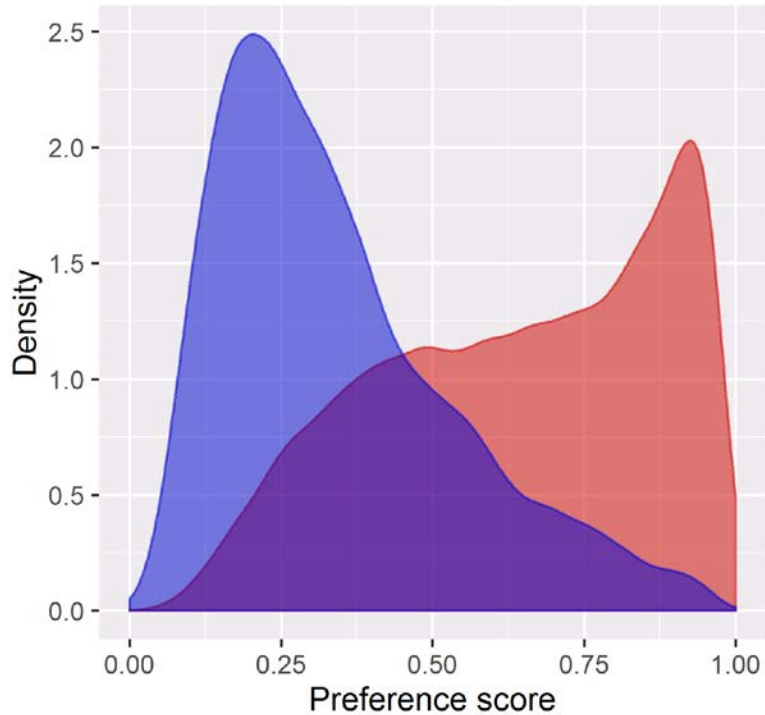
## OHDSI Keppra and the Risk of Angioedema study

This study aims to evaluate angioedema risk in seizure disorder patients exposed to Keppra (levetiracetam) compared with those exposed to phenytoin sodium. A potential link between levetiracetam and angioedema has been recently raised by the Food and Drug Administration in their review of spontaneous reporting data. In this study, we will analyze data from a distributed network using the OHDSI CohortMethod package.

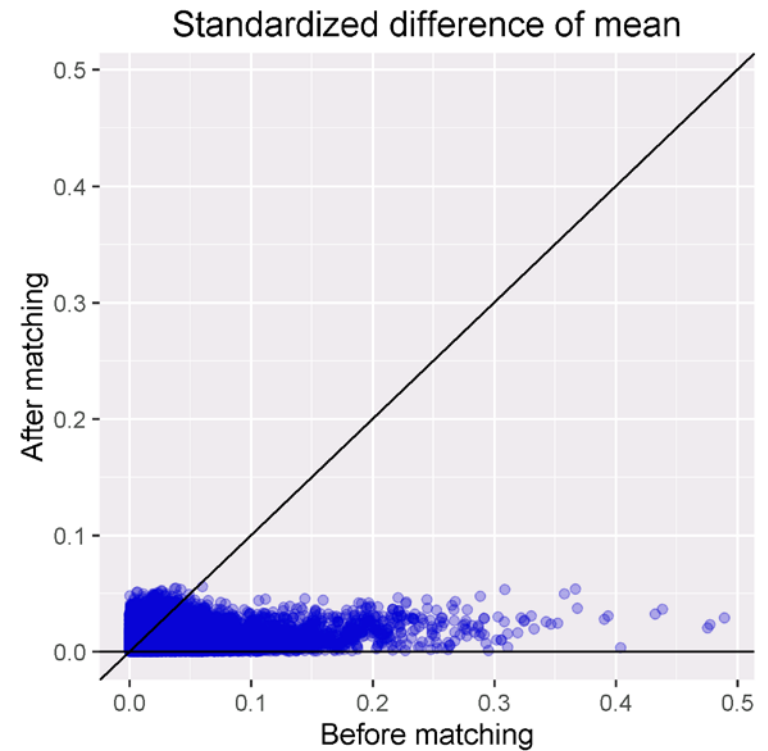


# Package Provides Diagnostic Checks

## Propensity Score Distribution



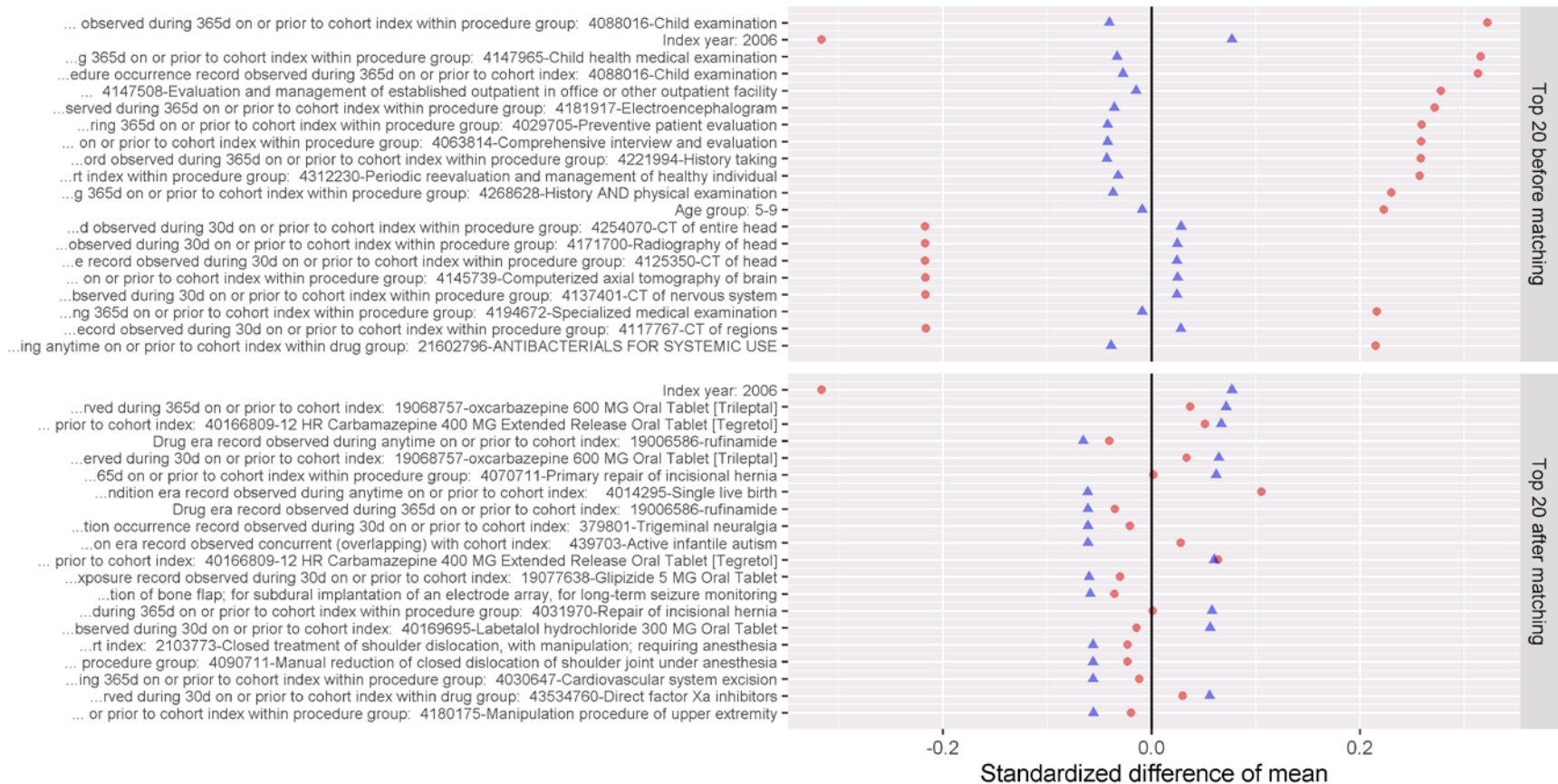
## Covariate Balance





# Covariate Balance

• before matching  
▲ after matching

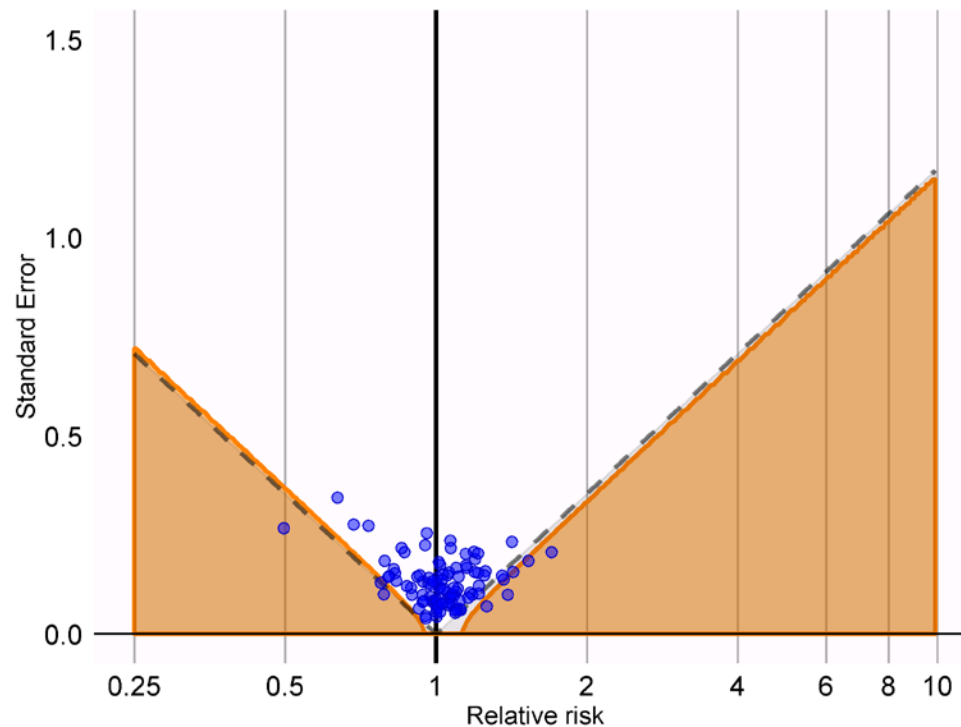






# Empirical Calibration

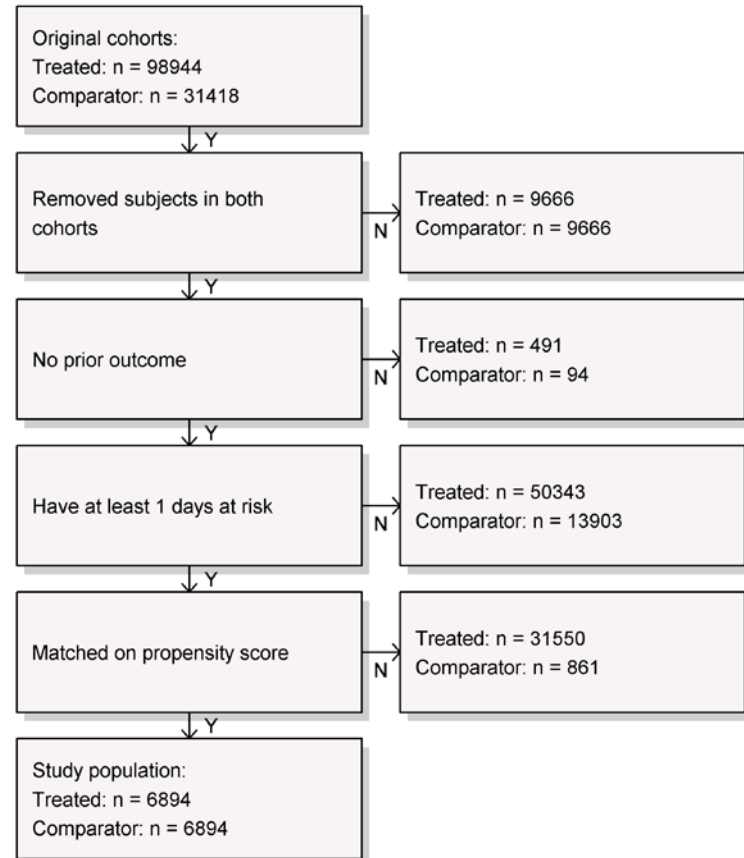
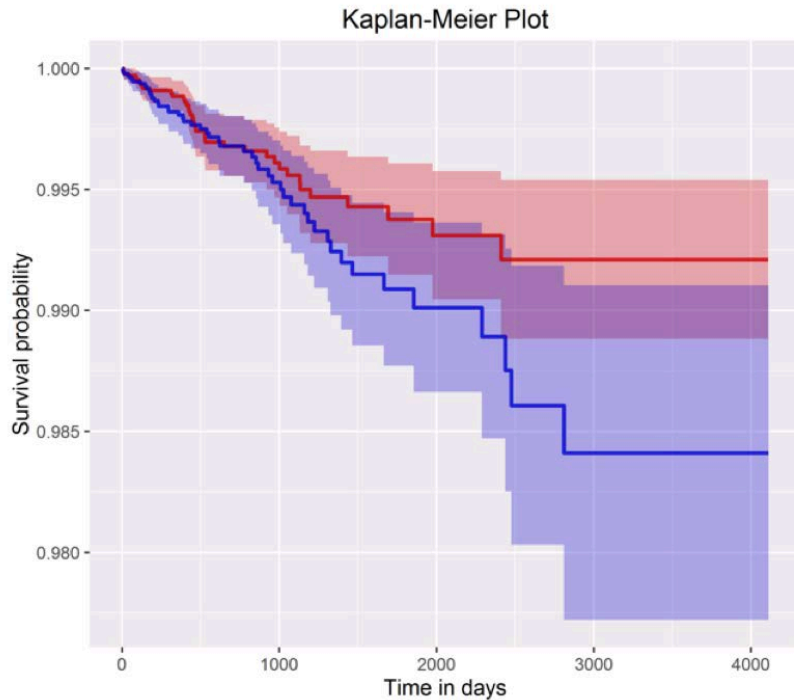
- Uses negative controls (conditions known not to be associated with the exposure) to calibrate p-values



This study is  
well calibrated!

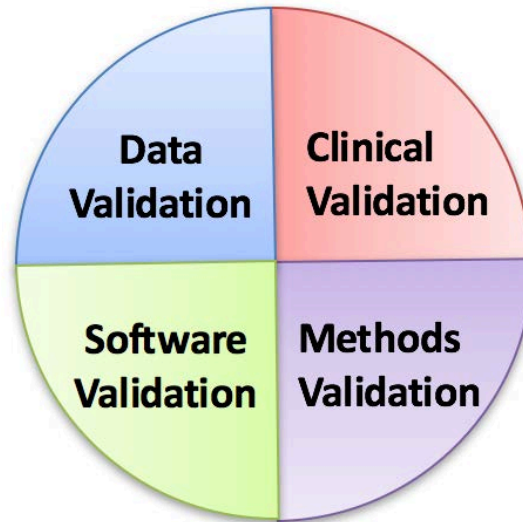


# Generates Results and Supporting Documents





“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](mailto:jon.duke@gatech.edu)