OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

An Open Collaborative Approach for Rapid Evidence Generation

David K. Vawdrey, PhD
George Hripcsak MD, MS
Jon D. Duke MD, MS
Patrick Ryan PhD
Nigam H. Shah MBBS, PhD

AMIA Joint Summits on Translational Science
March 25, 2015

Introduction

David K. Vawdrey, PhD
NewYork-Presbyterian Hospital
Columbia University
New York, USA

What is OHDSI?

- Video Introduction of OHDSI

What is OHDSI?

- The Observational Health Data Sciences and Informatics (OHDSI) collaborative is an international network of researchers and observational health databases

- The goal of OHDSI is to bring out the value of health data through large-scale analytics

What is OHDSI?

- OHDSI builds on the **Observational Medical Outcomes Partnership (OMOP)**, and maintains the OMOP **Common Data Model (CDM)**
- OHDSI provides a suite of tools and algorithms for conducting observational research using large data sets
- All OHDSI solutions are open-source

OHDSI Mission

To **transform medical decision-making** by creating reliable scientific evidence about disease natural history, healthcare delivery, and the effects of medical interventions through large-scale analysis of observational health databases for **population-level** estimation and **patient-level** predictions.

OHDSI Vision

OHDSI collaborators access a network of **one billion patients** to generate evidence about all aspects of healthcare.

Patients and clinicians and all other decision-makers around the world use OHDSI tools and evidence every day.

OHDSI Objectives

1. To establish a research community for observational health data sciences that enables active engagement across multiple disciplines and stakeholder groups

OHDSI Objectives

2. To develop and evaluate analytical methods that use observational health data to study the effects of medical interventions and predict health outcomes for patients, and to generate the empirical evidence base necessary to establish best practices in observational analysis

OHDSI Objectives

3. To apply scientific best practices in the design and implementation of open-source systems for observational analysis to enable medical product risk identification, comparative effectiveness research, patient-level predictions, and healthcare improvement

OHDSI Objectives

4. To generate evidence about disease natural history, healthcare delivery, and the effects of medical interventions, supporting medical decision-making in a way that is credible, consistent, transparent, and personalized to patients and providers

OHDSI Objectives

5. To establish educational opportunities to train students, practitioners, and consumers about the foundational science of observational health data analysis

George Hripcsak MD, MS
Professor and Chair
Department of Biomedical Informatics
Columbia University

Jon D. Duke MD, MS
Senior Scientist
Director, Drug Safety Informatics Program
Regenstrief Institute

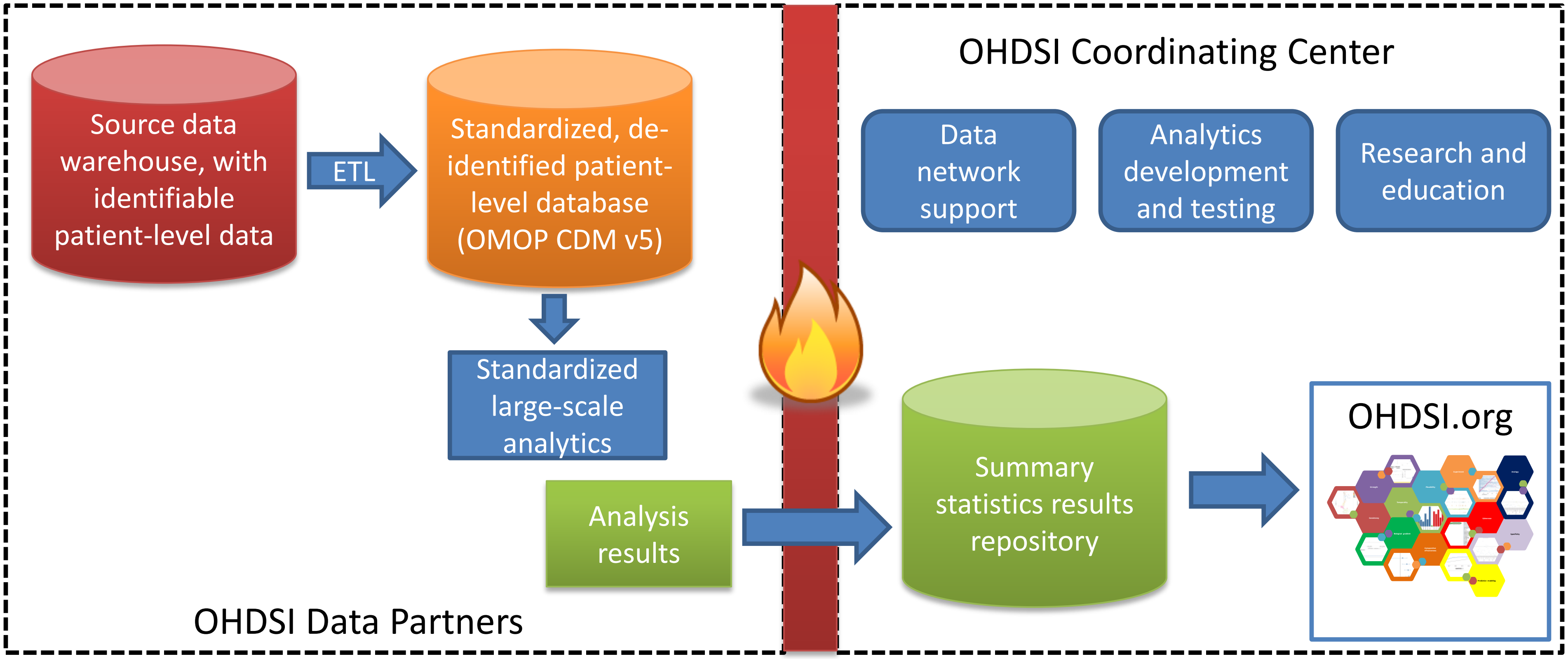
Patrick Ryan, PhD
Sr. Director and Head, Epidemiology Analytics
Janssen Research and Development

Nigam H. Shah MBBS, PhD
Assistant Professor
Dept. of Medicine (Biomedical Informatics)
Stanford University

OHDSI OMOP CDM at Columbia and Use of CDM in Clinical Data Research Networks (CDRNs)

George Hripcsak, MD, MS
Biomedical Informatics
Columbia University
New York, USA

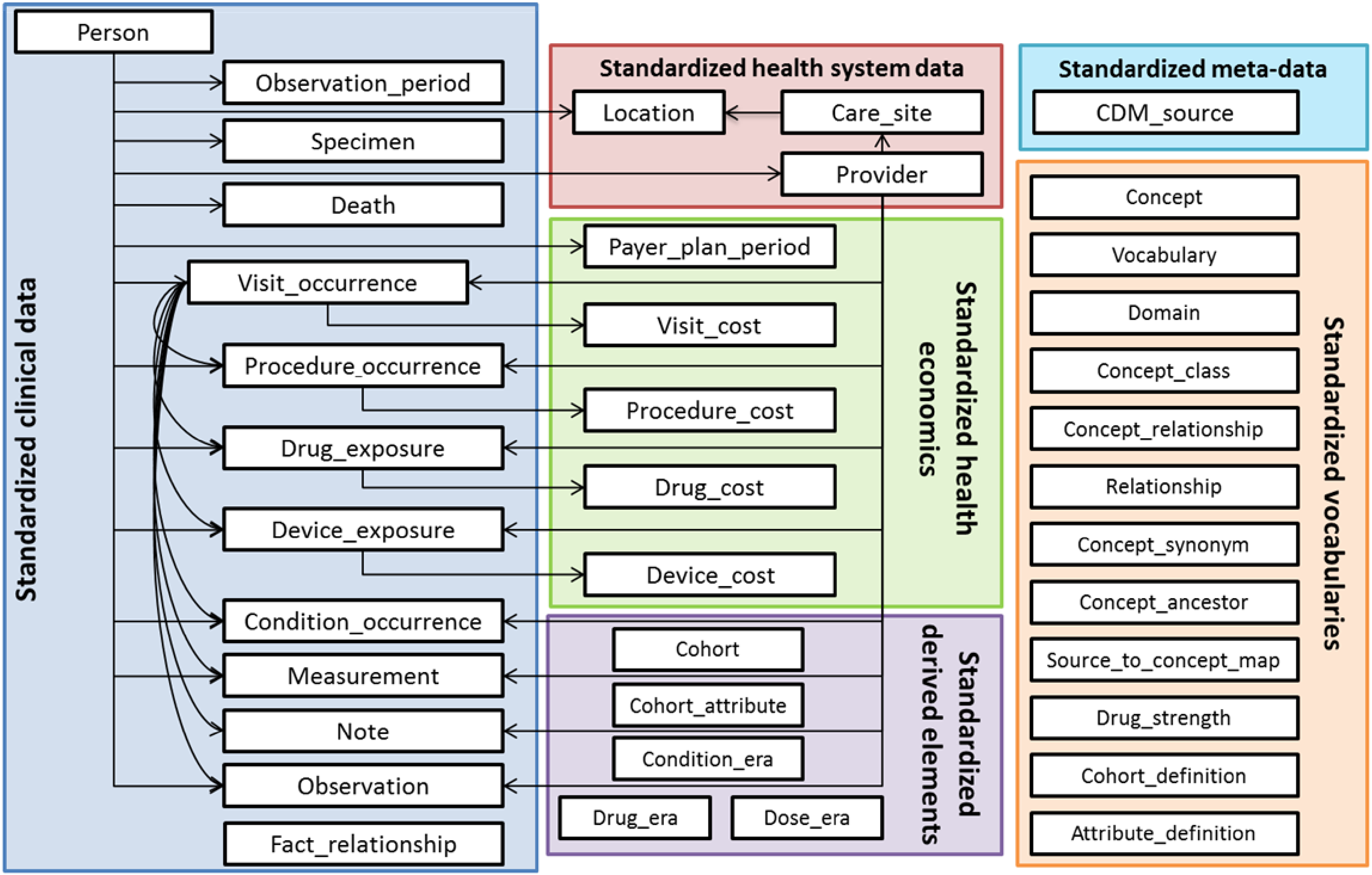
How OHDSI Works

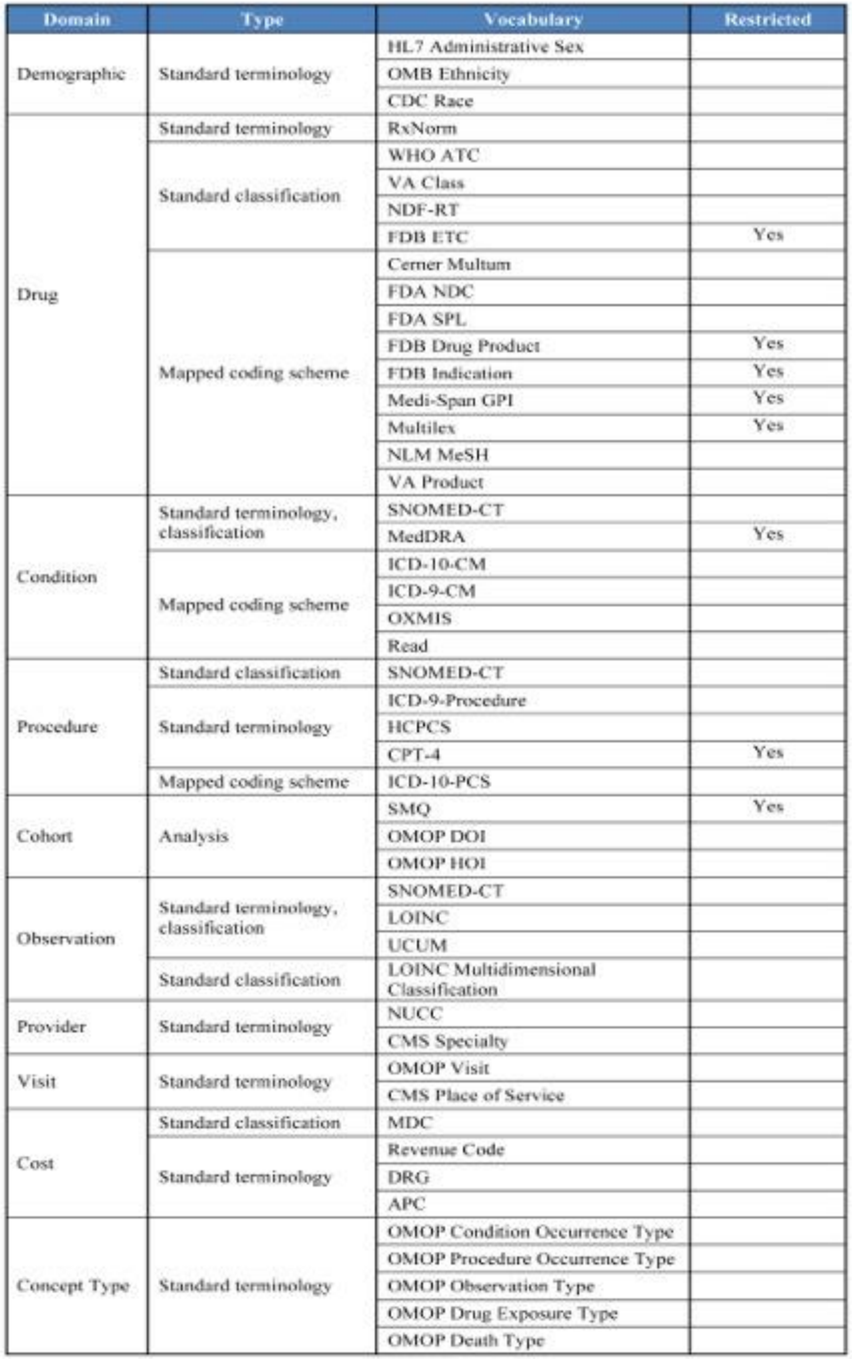

Source data warehouse, with identifiable patient-level data
ETL
Standardized, de-identified patient-level database (OMOP CDM v5)
Standardized large-scale analytics
Analysis results
OHDSI Data Partners
OHDSI Coordinating Center
Data network support
Analytics development and testing
Research and education
Summary statistics results repository
OHDSI.org

OHDSI Information Architecture

- Each site retains its own data
- Use a common information model
  - Concepts, terminologies, conceptual relations
  - “OMOP Common Data Model (v4, v5)”
  - Strictly defines terminology, mappings
  - Supports world-wide queries
- Advanced observational research methods
- Aggregate the results centrally

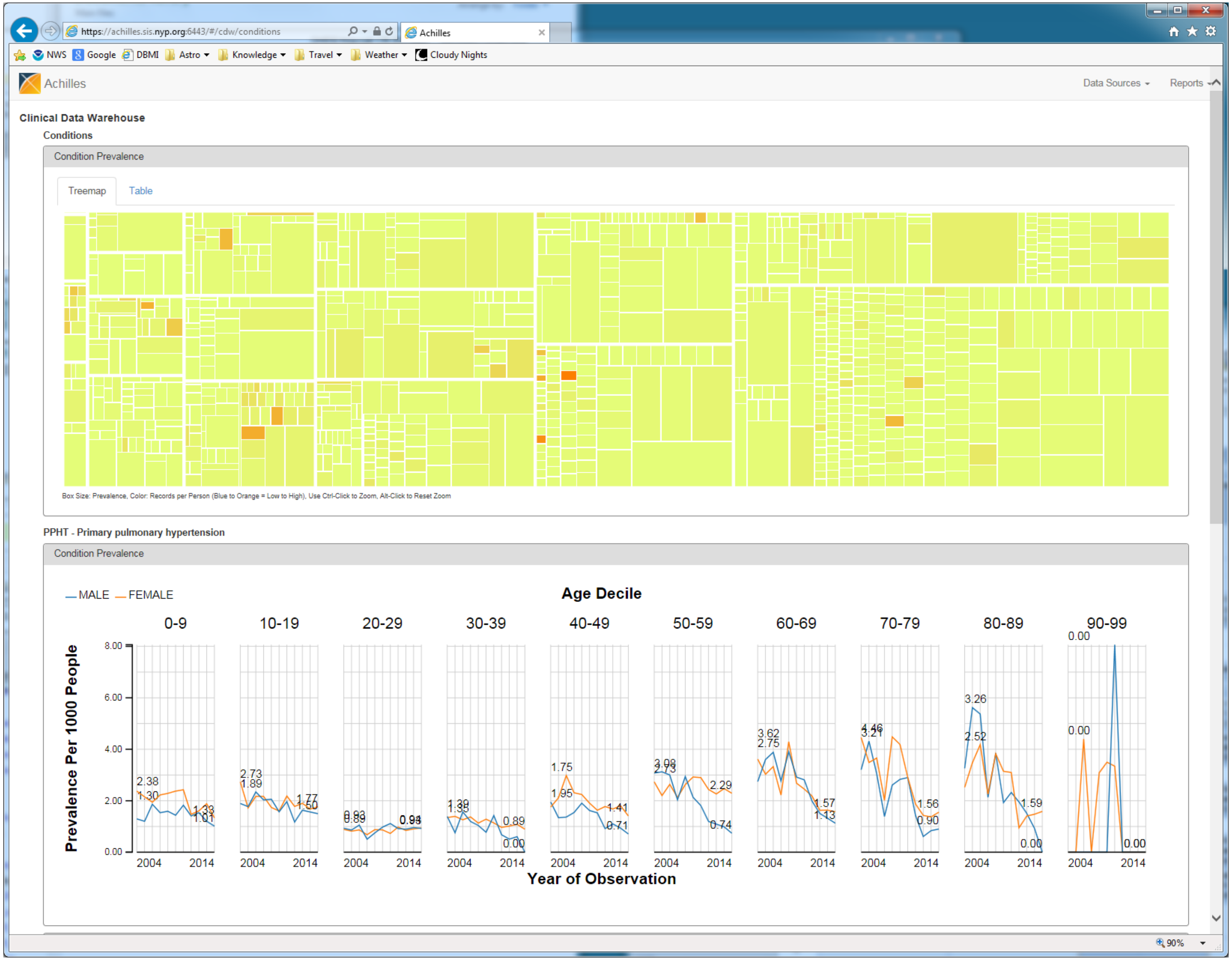
OMOP Common Data Model (CDM) v. 5.0





Domain	Type	Vocabulary	Restricted
Demographic	Standard terminology	HL7 Administrative Sex	
		OMB Ethnicity	
		CDC Race	
Drug	Standard terminology	RxNorm	
	Standard classification	WHO ATC	
		VA Class	
		NDF-RT	
		FDB ETC	Yes
	Mapped coding scheme	Cerner Multum	
		FDA NDC	
		FDA SPL	
		FDB Drug Product	Yes
		FDB Indication	Yes
		Medi-Span GPI	Yes
		Multilex	Yes
		NLM MeSH	
		VA Product	
Condition	Standard terminology, classification	SNOMED-CT	
		MedDRA	Yes
	Mapped coding scheme	ICD-10-CM	
		ICD-9-CM	
		OXMIS	
		Read	
Procedure	Standard classification	SNOMED-CT	
	Standard terminology	ICD-9-Procedure	
		HCPCS	
		CPT-4	Yes
	Mapped coding scheme	ICD-10-PCS	
Cohort	Analysis	SMQ	Yes
		OMOP DOI	
		OMOP HOI	
Observation	Standard terminology, classification	SNOMED-CT	
		LOINC	
		UCUM	
	Standard classification	LOINC Multidimensional Classification	
Provider	Standard terminology	NUCC	
		CMS Specialty	
Visit	Standard terminology	OMOP Visit	
		CMS Place of Service	
Cost	Standard classification	MDC	
	Standard terminology	Revenue Code	
		DRG	
		APC	
Concept Type	Standard terminology	OMOP Condition Occurrence Type	
		OMOP Procedure Occurrence Type	
		OMOP Observation Type	
		OMOP Drug Exposure Type	
		OMOP Death Type	

ACHILLES



Clinical Data Warehouse

Conditions

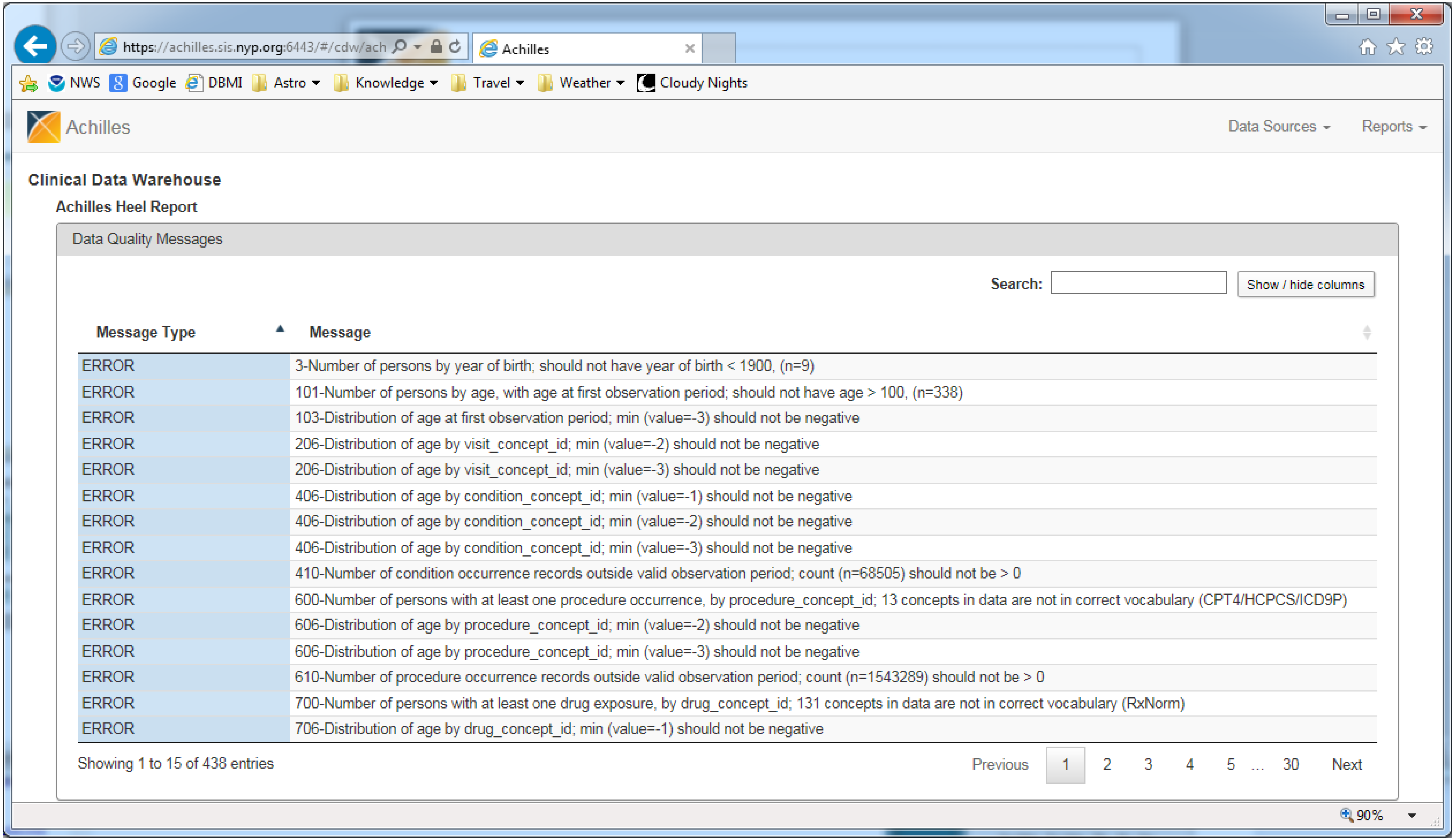
Condition Prevalence

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

PPHT - Primary pulmonary hypertension

Condition Prevalence

ACHILLES



Clinical Data Warehouse

Achilles Heel Report

Data Quality Messages

Message Type	Message
ERROR	3-Number of persons by year of birth; should not have year of birth < 1900, (n=9)
ERROR	101-Number of persons by age, with age at first observation period; should not have age > 100, (n=338)
ERROR	103-Distribution of age at first observation period; min (value=-3) should not be negative
ERROR	206-Distribution of age by visit_concept_id; min (value=-2) should not be negative
ERROR	206-Distribution of age by visit_concept_id; min (value=-3) should not be negative
ERROR	406-Distribution of age by condition_concept_id; min (value=-1) should not be negative
ERROR	406-Distribution of age by condition_concept_id; min (value=-2) should not be negative
ERROR	406-Distribution of age by condition_concept_id; min (value=-3) should not be negative
ERROR	410-Number of condition occurrence records outside valid observation period; count (n=68505) should not be > 0
ERROR	600-Number of persons with at least one procedure occurrence, by procedure_concept_id; 13 concepts in data are not in correct vocabulary (CPT4/HCPCS/ICD9P)
ERROR	606-Distribution of age by procedure_concept_id; min (value=-2) should not be negative
ERROR	606-Distribution of age by procedure_concept_id; min (value=-3) should not be negative
ERROR	610-Number of procedure occurrence records outside valid observation period; count (n=1543289) should not be > 0
ERROR	700-Number of persons with at least one drug exposure, by drug_concept_id; 131 concepts in data are not in correct vocabulary (RxNorm)
ERROR	706-Distribution of age by drug_concept_id; min (value=-1) should not be negative

Showing 1 to 15 of 438 entries

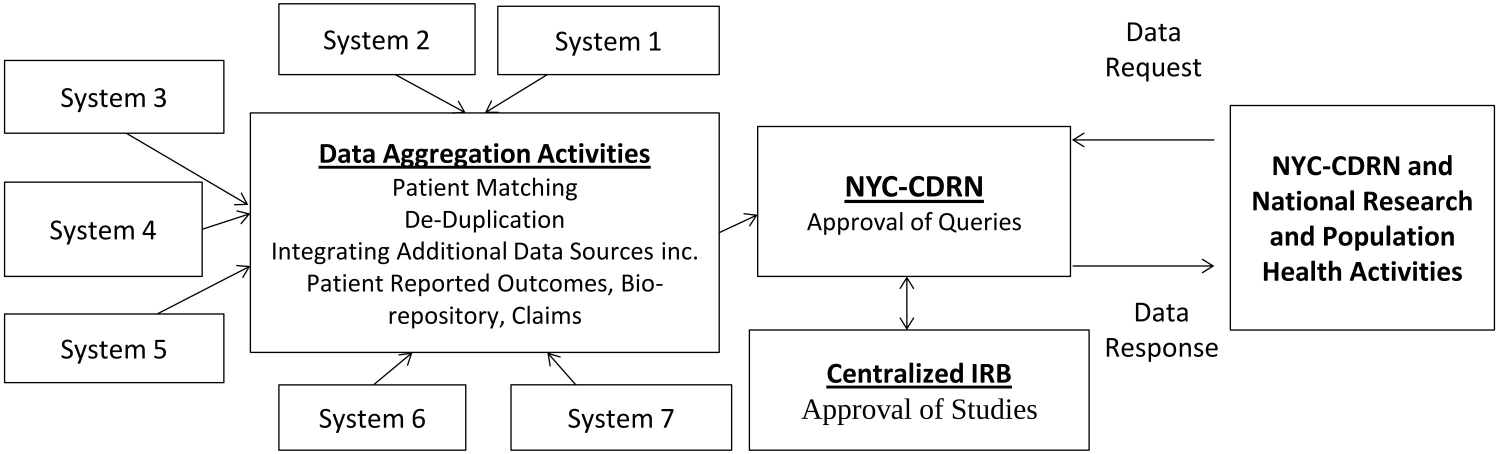
NYC-CDRN

New York City Clinical Data Research Network

Partner	Organization
Health System	• Clinical Directors Network • Columbia University College of Physicians and Surgeons • Montefiore Medical Center and Albert Einstein College of Med • Mount Sinai Health System and Icahn School of Medicine • NewYork-Presbyterian Hospital • NYU Langone Medical Center and NYU School of Medicine • Weill Cornell Medical College
Research Infrastructure	• Biomedical Research Alliance of New York • Cornell NYC Tech Campus • New York Genome Center • Rockefeller University
Health Information Exchange	• Bronx RHIO (Bronx Regional Informatics Center) • Healthix
Patient Organizations	• American Diabetes Association • Center for Medical Consumers • Consumer Reports • Cystic Fibrosis Foundation • New York Academy of Medicine • NYS Department of Health

NYC-CDRN

New York City Clinical Data Research Network



System 2

System 1

System 3

System 4

System 5

System 6

System 7

Data Aggregation Activities
Patient Matching
De-Duplication
Integrating Additional Data Sources inc. Patient Reported Outcomes, Bio-repository, Claims

NYC-CDRN
Approval of Queries

Centralized IRB
Approval of Studies

Data Request

Data Response

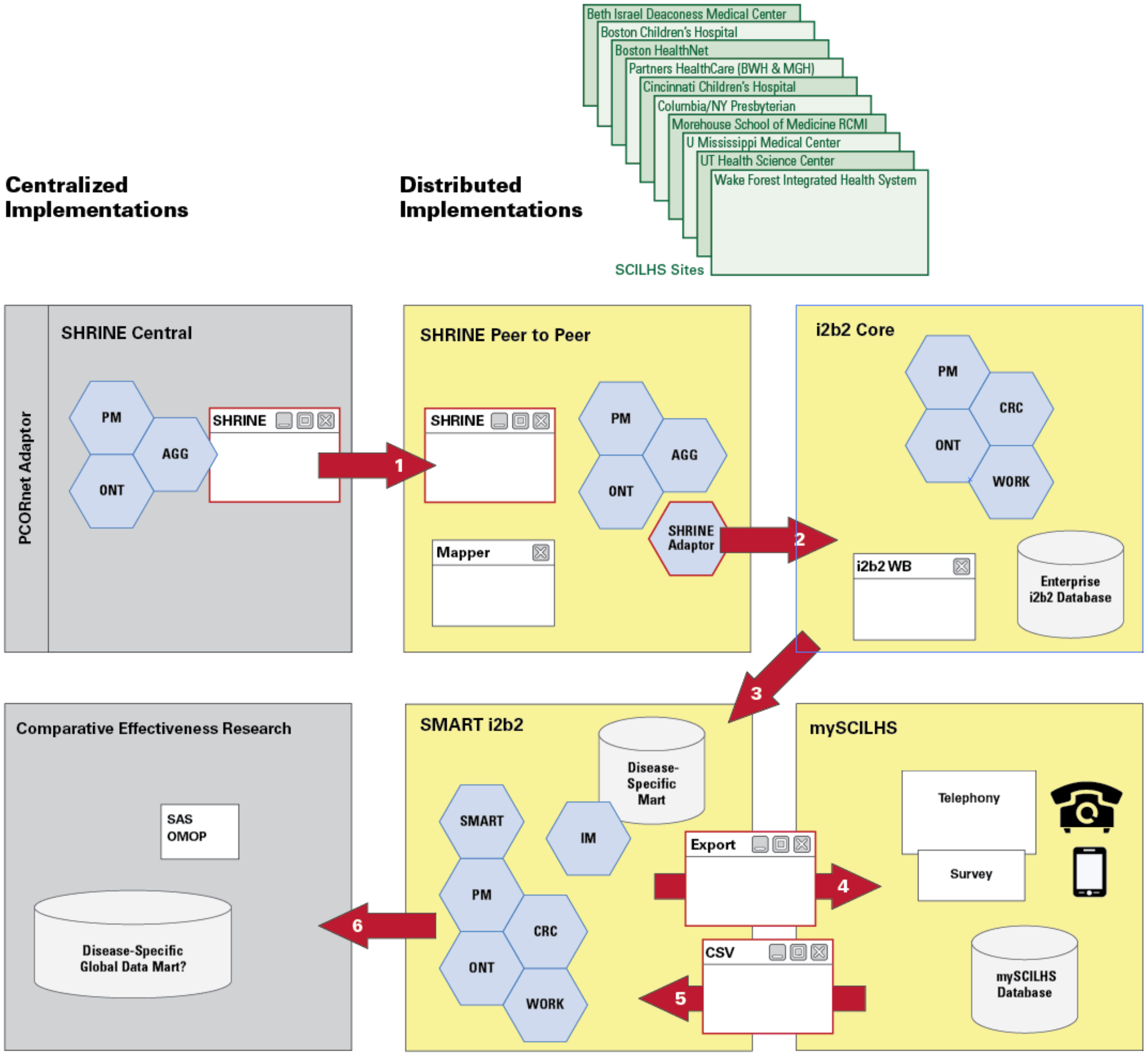
NYC-CDRN and National Research and Population Health Activities

Scalable Collaborative Infrastructure for a Learning Health Care System (SCILHS)

- Boston Children's Hospital
- Boston Health Net (Boston Med Center, etc.)
- Partners HealthCare System (Mass General, Brigham & Women's)
- Wake Forest Baptist Medical Center
- Beth Israel Deaconess Medical Center

- Cincinnati Children's Hospital
- University of Texas Health Science Center/Houston
- Columbia U Medical Center and NewYork-Presbyterian
- Morehouse/Grady/RCMI
- U Mississippi Medical Center/RCMI

SCILHS



Centralized Implementations

Distributed Implementations

Beth Israel Deaconess Medical Center
Boston Children's Hospital
Boston HealthNet
Partners HealthCare (BWH & MGH)
Cincinnati Children's Hospital
Columbia/NY Presbyterian
Morehouse School of Medicine RCMI
U Mississippi Medical Center
UT Health Science Center
Wake Forest Integrated Health System

SCILHS Sites

SHRINE Central

SHRINE Peer to Peer

i2b2 Core

Comparative Effectiveness Research

SMART i2b2

mySCILHS

Advance Clinical Trials (ACT)

- CTSA-driven, NCATS funded
- Promote innovation and efficiency in participant recruitment into multi-site studies
- 21 CTSA sites
- i2b2, SHRINE

OHDSI and i2b2 Opportunity

- Information model
  - Distinct from data schema
  - i2b2 flexible but slows cross-entity research
  - OHDSI highly defined
- Can use i2b2 or OHDSI schema, but OHDSI information model

OHDSI and i2b2

- PCORI Clinical Data Research Network (CDRN) in U.S.
  - 4 OHDSI/OMOP sites, 7 i2b2 sites (of 11)
  - Store in OHDSI or i2b2
  - Convert between them and convert to PCORnet

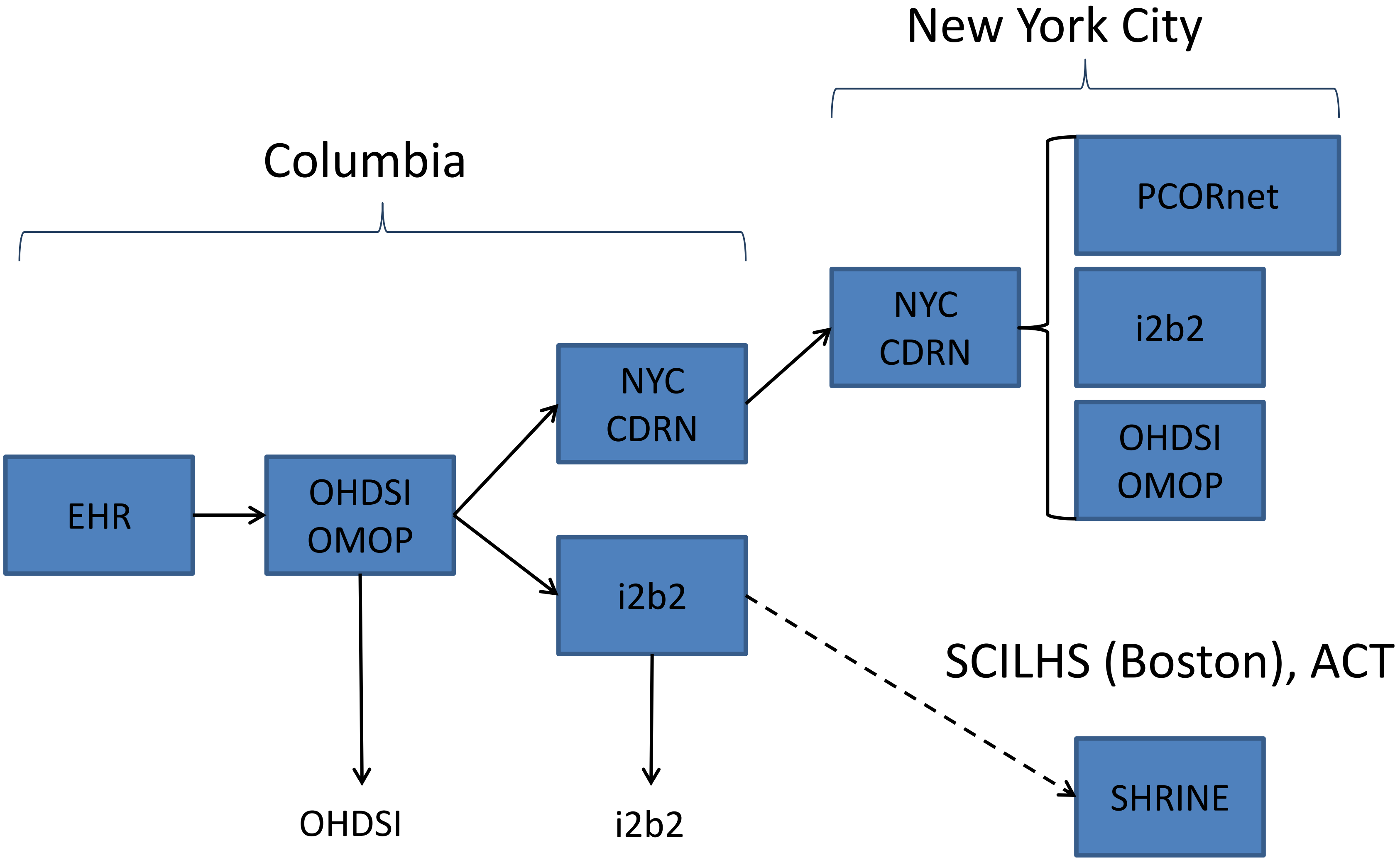
CDRN Alignment Tasks

- Construct CDRN Data Model (DM) and CDRN Vocabulary
  - Based on OMOP DM/Vocabulary
  - Address PCOR requirements
  - Address CDRN local needs
  - Align with OMOP V5 development
  - Align with other CDRN centers
  - Address versioning
- Develop Map-Sets
  - Develop vocabulary map-sets:
    - Sources-to-OMOP
    - i2b2-OMOP
    - PCOR-OMOP
  - Address versioning
  - Facilitate development of ETL processes
    - i2b2-OMOP
    - PCOR-OMOP

Deliverables

- ✓ Design Person table
- ✓ Design terminology back-end
- ✓ Select/create demographics controlled terminology
- ✓ Create mappings of site terminology to controlled terminology for submitting sites
- ✓ Provide QA recommendations
- ✓ Document decisions and artifacts

Columbia CDRN Approach



Population and Cohort Characterization Using the OMOP CDM

Jon D. Duke MD, MS
Regenstrief Institute

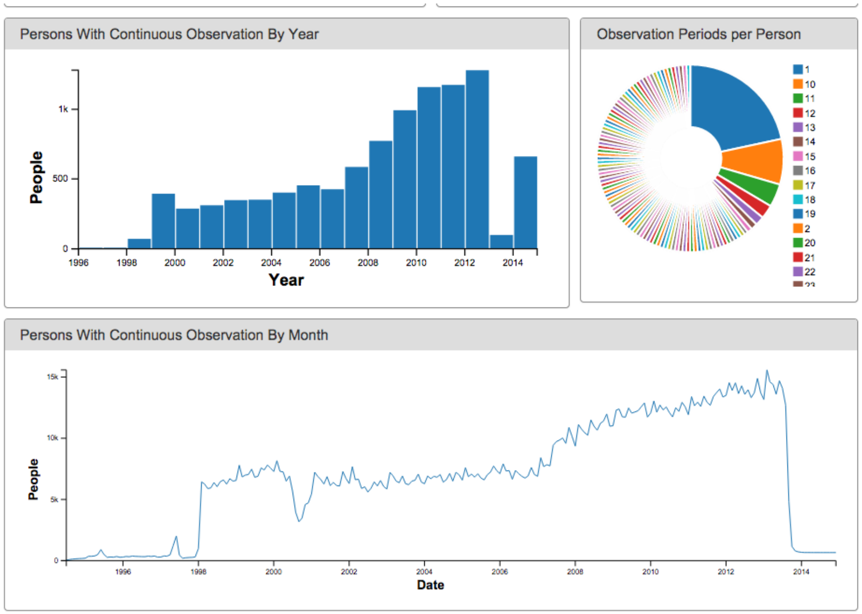
Characterization in OHDSI

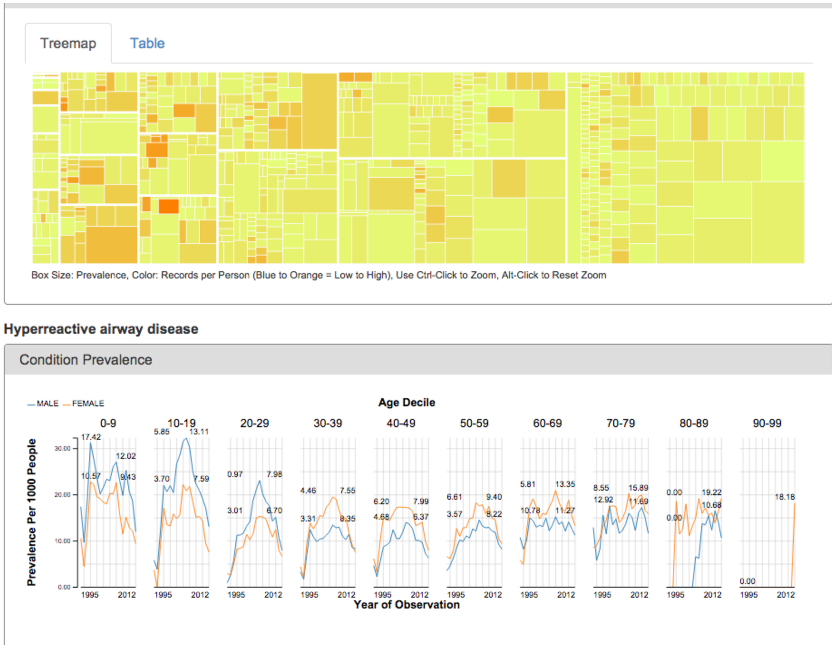
- In OHDSI, characterization = generating a comprehensive overview of a patient dataset
  - Clinical (e.g., conditions, medications, procedures)
  - Metadata (e.g., observation periods, data density)
- Supports
  - Feasibility studies
  - Hypothesis generation
  - Data quality assessment
  - Data sharing (aggregate-level)

OHDSI Tools for Characterization

- Population-Wide
  - ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems)
- Specific Cohorts
  - HERACLES (Health Enterprise Resource and Care Learning Exploration System)

ACHILLES

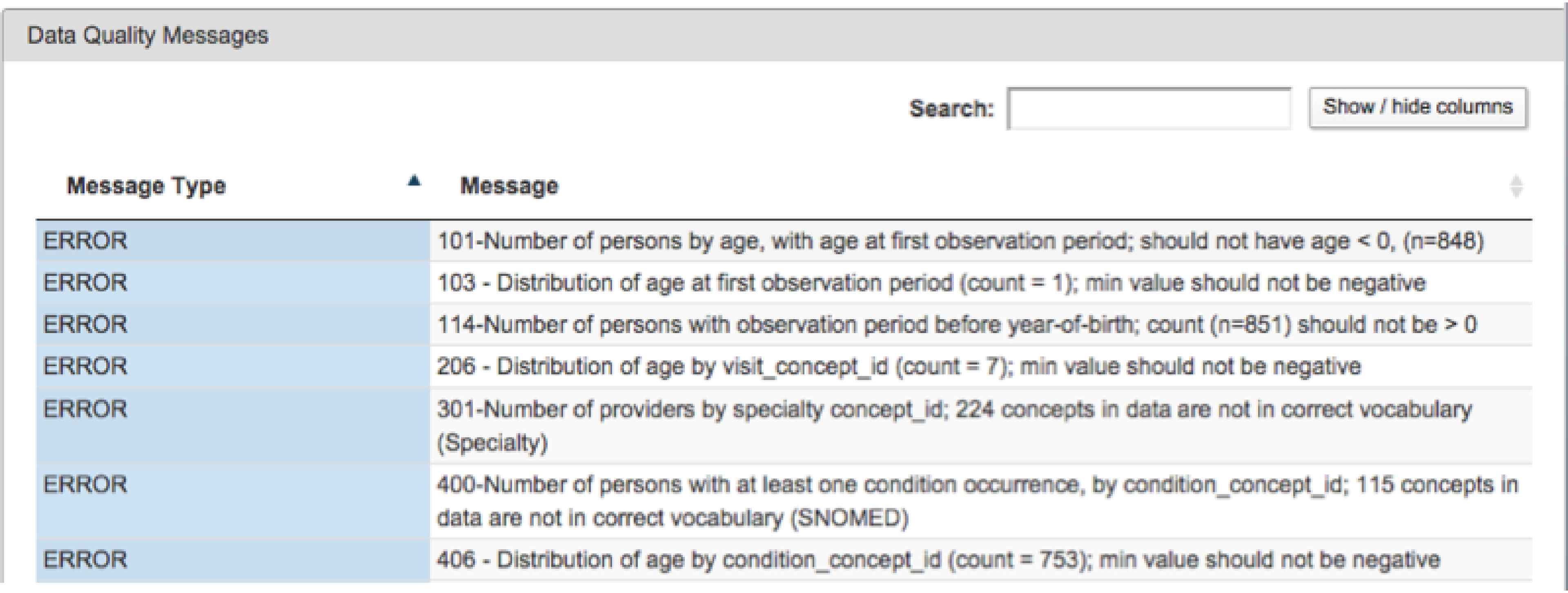




Box Size: Prevalence, Color: Records per Person (Blue to Orange = Low to High), Use Ctrl-Click to Zoom, Alt-Click to Reset Zoom

Hyperreactive airway disease

ACHILLES & Data Quality



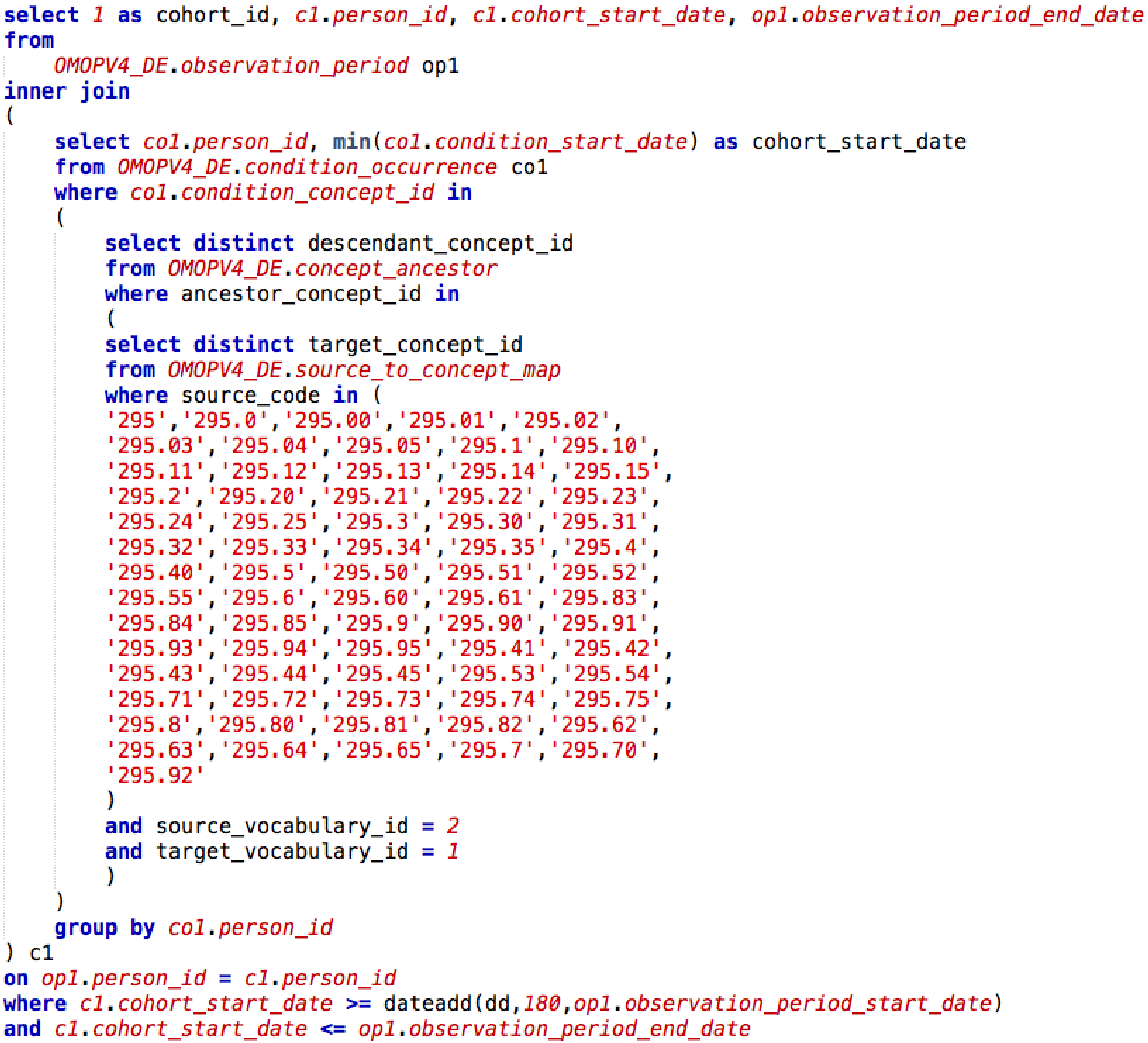
Message Type	Message
ERROR	101-Number of persons by age, with age at first observation period; should not have age < 0, (n=848)
ERROR	103 - Distribution of age at first observation period (count = 1); min value should not be negative
ERROR	114-Number of persons with observation period before year-of-birth; count (n=851) should not be > 0
ERROR	206 - Distribution of age by visit_concept_id (count = 7); min value should not be negative
ERROR	301-Number of providers by specialty concept_id; 224 concepts in data are not in correct vocabulary (Specialty)
ERROR	400-Number of persons with at least one condition occurrence, by condition_concept_id; 115 concepts in data are not in correct vocabulary (SNOMED)
ERROR	406 - Distribution of age by condition_concept_id (count = 753); min value should not be negative

ACHILLES

- Needs to be run only once per CDM
- Hybrid R / web-based application
- Can specify minimum cell size to enable sharing where possible

Cohort Characterization

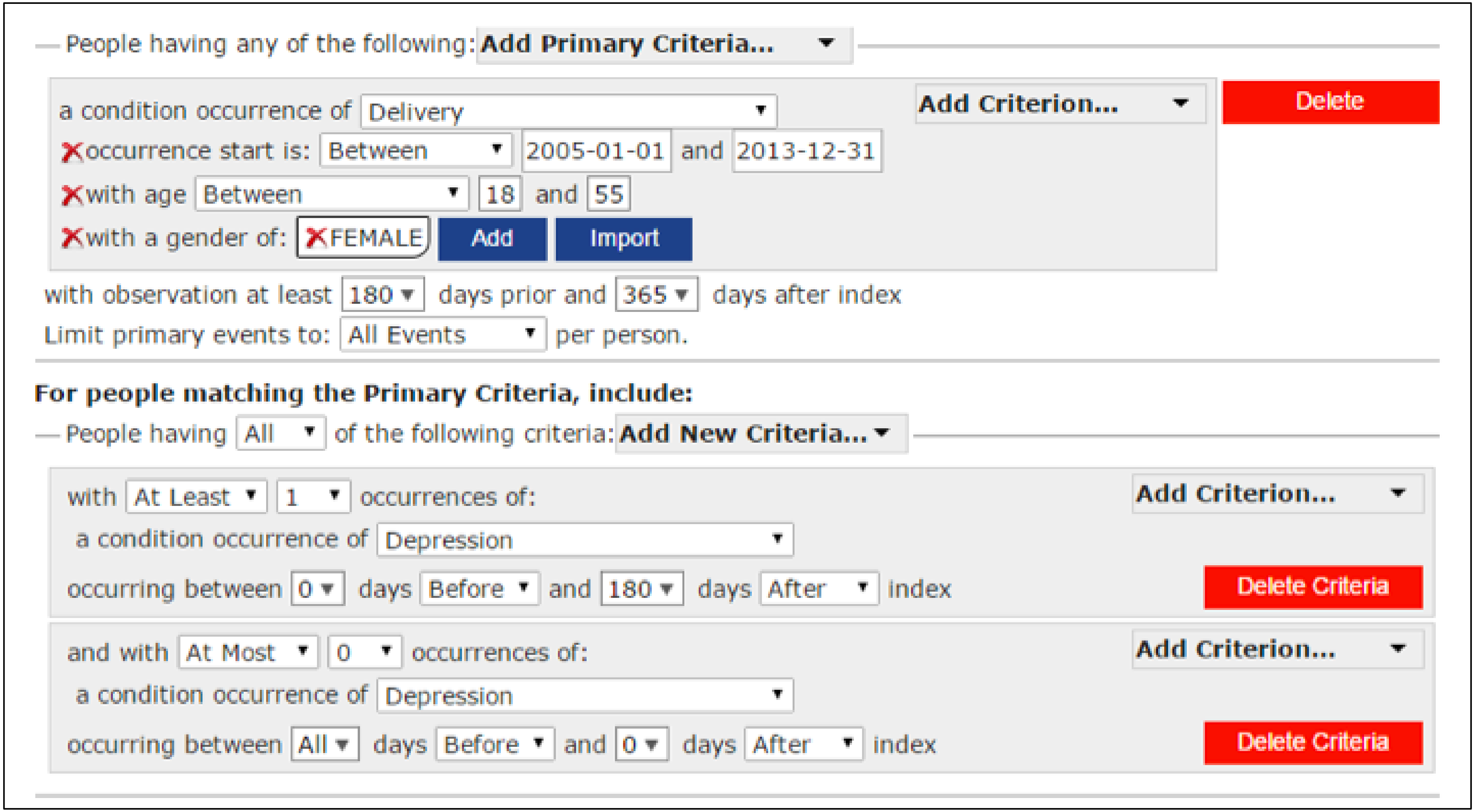
- CDM Cohorts can be created in a variety of ways
  - Manual queries



```
select 1 as cohort_id, c1.person_id, c1.cohort_start_date, op1.observation_period_end_date
from
    OMOPV4_DE.observation_period op1
inner join
(
    select co1.person_id, min(co1.condition_start_date) as cohort_start_date
    from OMOPV4_DE.condition_occurrence co1
    where co1.condition_concept_id in
    (
        select distinct descendant_concept_id
        from OMOPV4_DE.concept_ancestor
        where ancestor_concept_id in
        (
        select distinct target_concept_id
        from OMOPV4_DE.source_to_concept_map
        where source_code in (
        '295','295.0','295.00','295.01','295.02',
        '295.03','295.04','295.05','295.1','295.10',
        '295.11','295.12','295.13','295.14','295.15',
        '295.2','295.20','295.21','295.22','295.23',
        '295.24','295.25','295.3','295.30','295.31',
        '295.32','295.33','295.34','295.35','295.4',
        '295.40','295.5','295.50','295.51','295.52',
        '295.55','295.6','295.60','295.61','295.83',
        '295.84','295.85','295.9','295.90','295.91',
        '295.93','295.94','295.95','295.41','295.42',
        '295.43','295.44','295.45','295.53','295.54',
        '295.71','295.72','295.73','295.74','295.75',
        '295.8','295.80','295.81','295.82','295.62',
        '295.63','295.64','295.65','295.7','295.70',
        '295.92'
        )
        and source_vocabulary_id = 2
        and target_vocabulary_id = 1
        )
    )
    group by co1.person_id
) c1
on op1.person_id = c1.person_id
where c1.cohort_start_date >= dateadd(dd,180,op1.observation_period_start_date)
and c1.cohort_start_date <= op1.observation_period_end_date
```

Cohort Characterization

- CDM Cohorts can be created in a variety of ways
  - Manual queries
  - Cohort building tool (CIRCE)



People having any of the following:

a condition occurrence of Delivery

✗occurrence start is: Between 2005-01-01 and 2013-12-31

✗with age Between 18 and 55

✗with a gender of: ✗FEMALE

with observation at least 180 days prior and 365 days after index

Limit primary events to: All Events per person.

For people matching the Primary Criteria, include:

People having All of the following criteria:

with At Least 1 occurrences of:

a condition occurrence of Depression

occurring between 0 days Before and 180 days After index

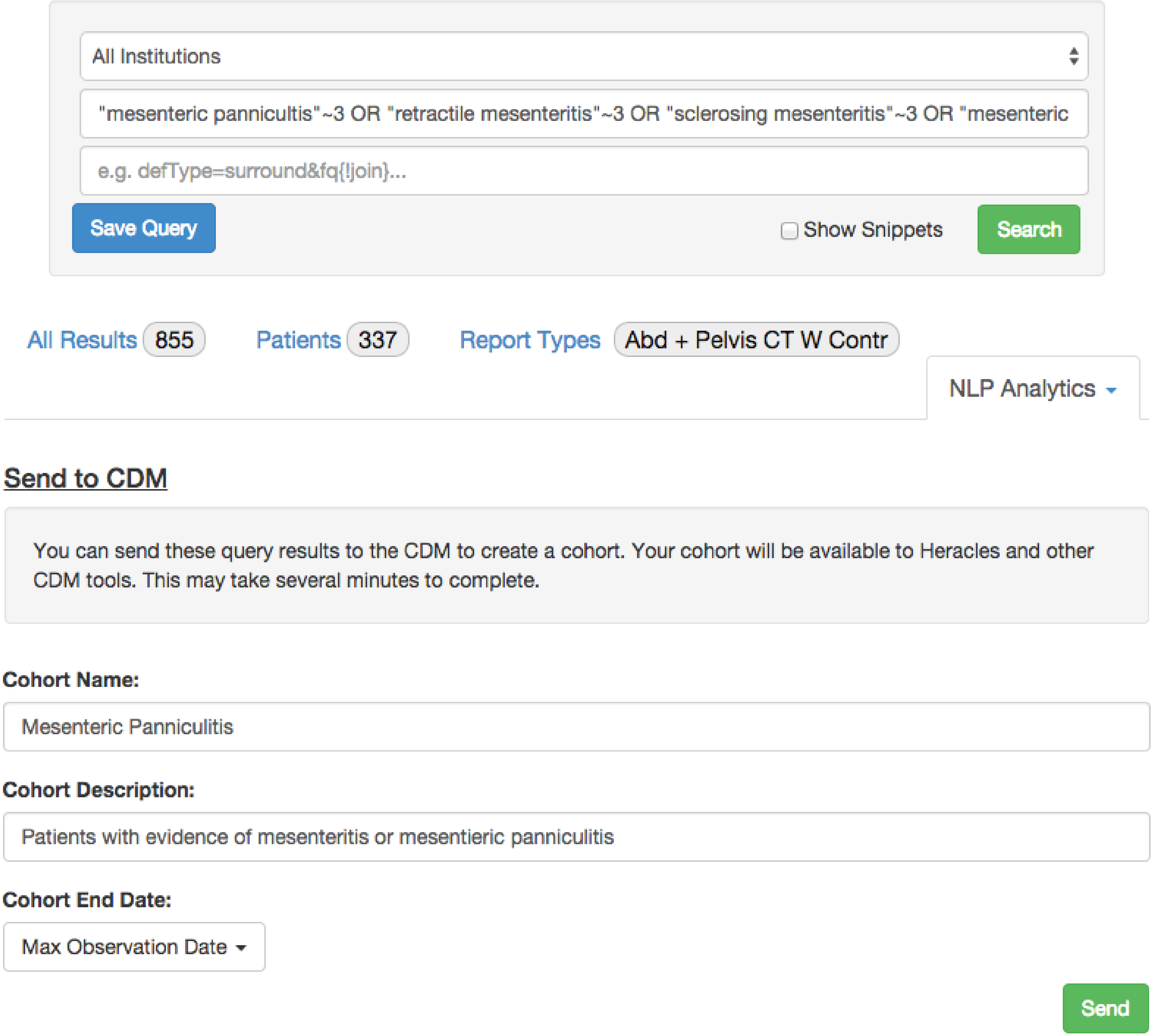
and with At Most 0 occurrences of:

a condition occurrence of Depression

occurring between All days Before and 0 days After index

Cohort Characterization

- CDM Cohorts can be created in a variety of ways
  - Manual queries
  - Cohort building tool (CIRCE)
  - Import of externally defined patient list



All Institutions

"mesenteric panniculitis"~3 OR "retractile mesenteritis"~3 OR "sclerosing mesenteritis"~3 OR "mesenteric

e.g. defType=surround&fq{!join}...

Report Types: Abd + Pelvis CT W Contr

Send to CDM

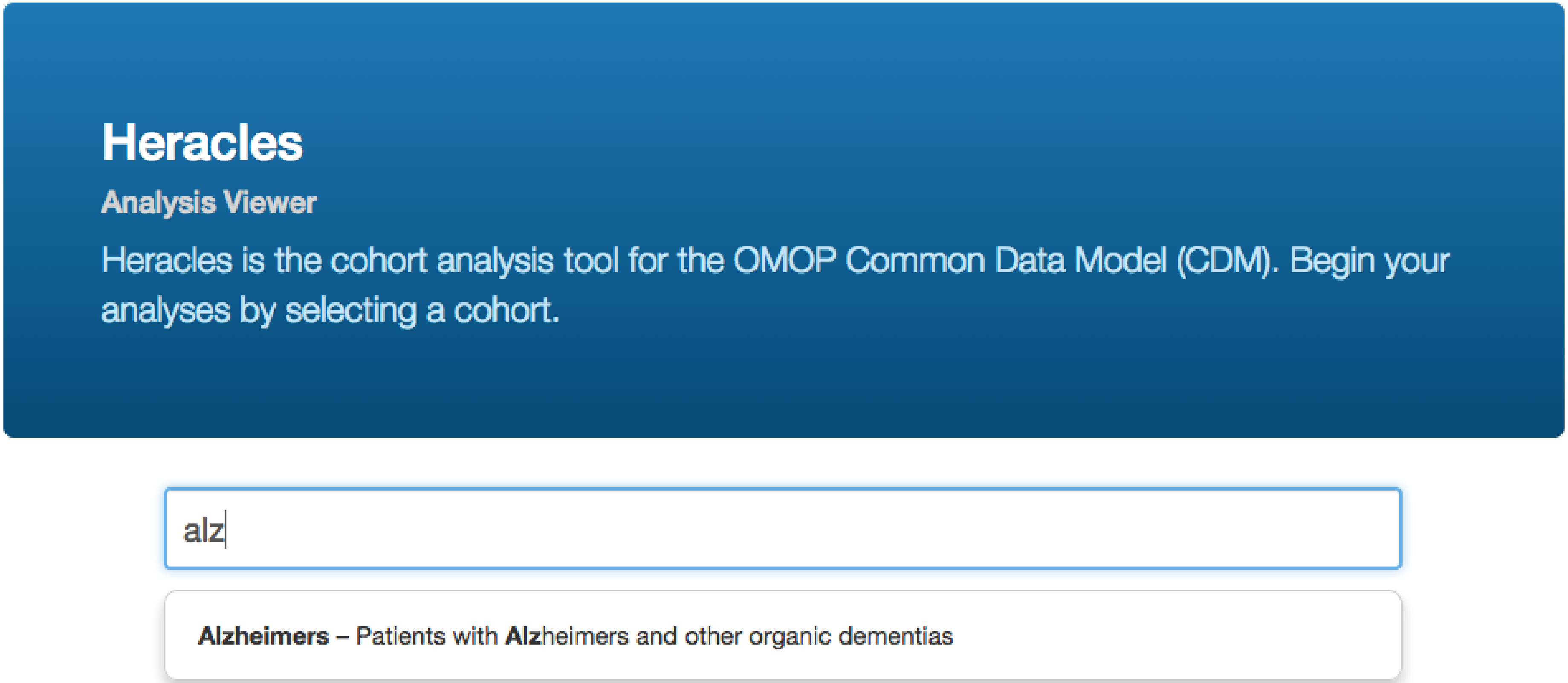
You can send these query results to the CDM to create a cohort. Your cohort will be available to Heracles and other CDM tools. This may take several minutes to complete.

Cohort Name: Mesenteric Panniculitis

Cohort Description: Patients with evidence of mesenteritis or mesentieric panniculitis

Cohort End Date: Max Observation Date

HERACLES



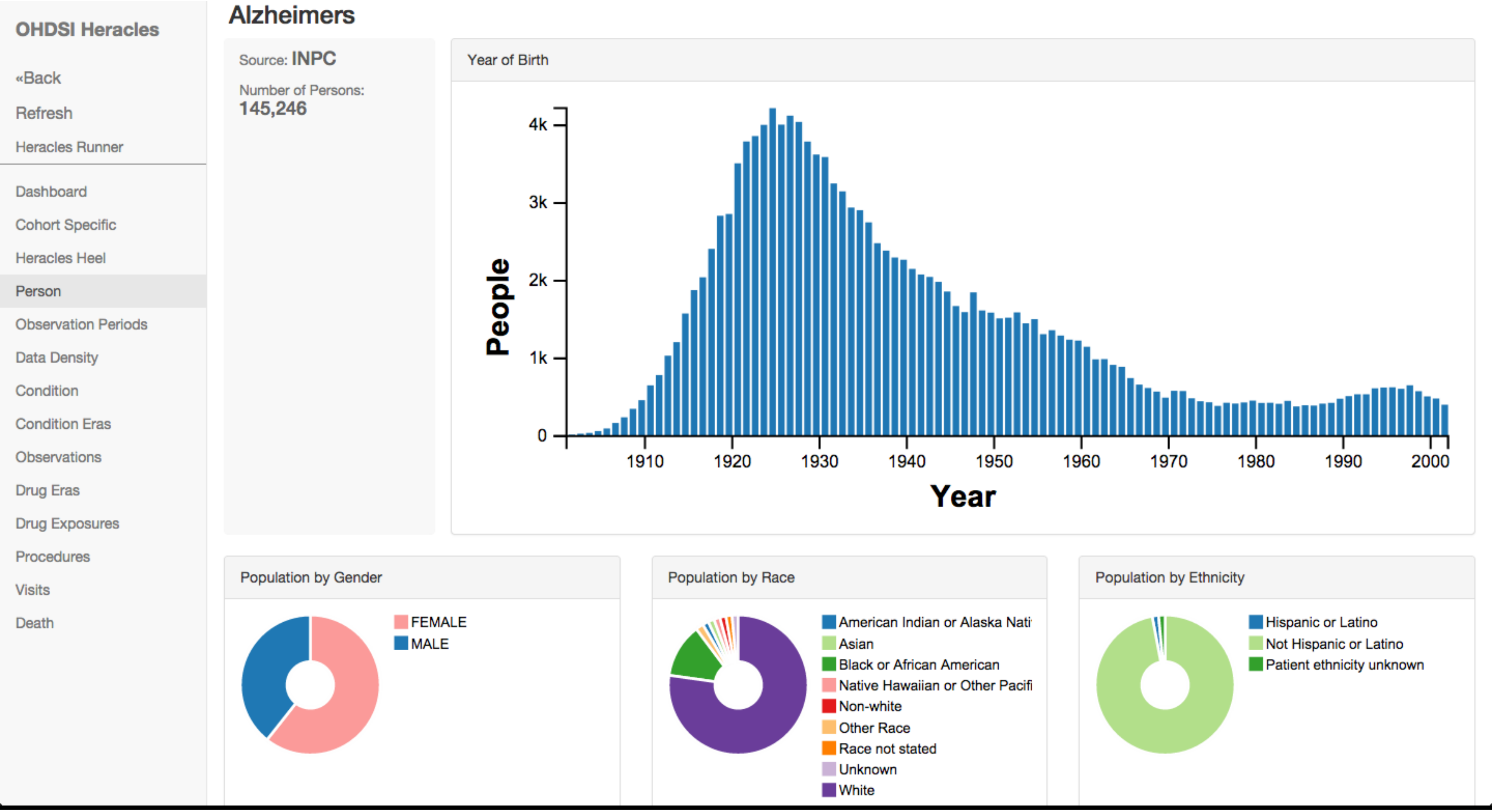
Heracles

Analysis Viewer

Heracles is the cohort analysis tool for the OMOP Common Data Model (CDM). Begin your analyses by selecting a cohort.

alz

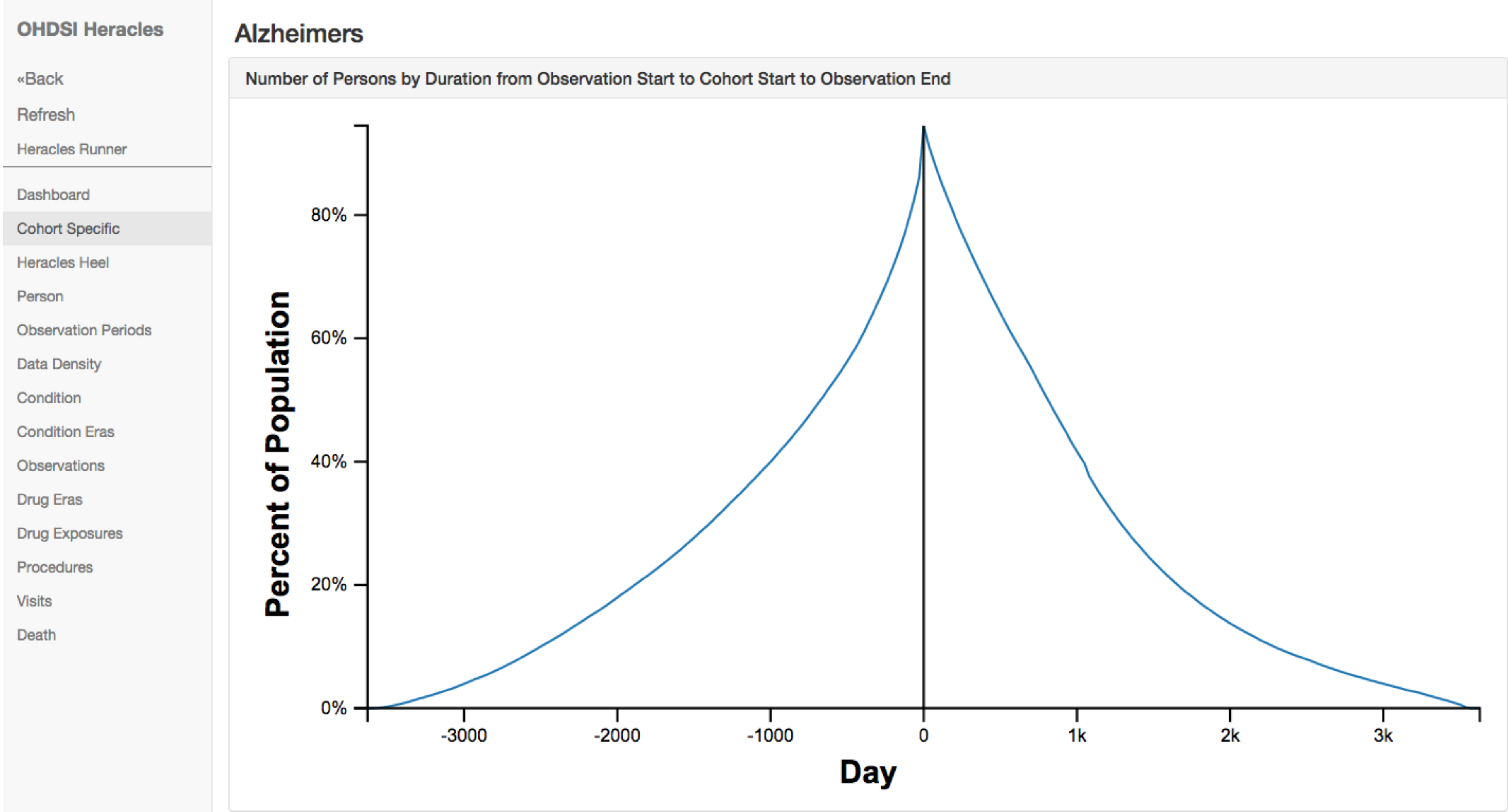
Alzheimers – Patients with **Alz**heimers and other organic dementias



Alzheimers

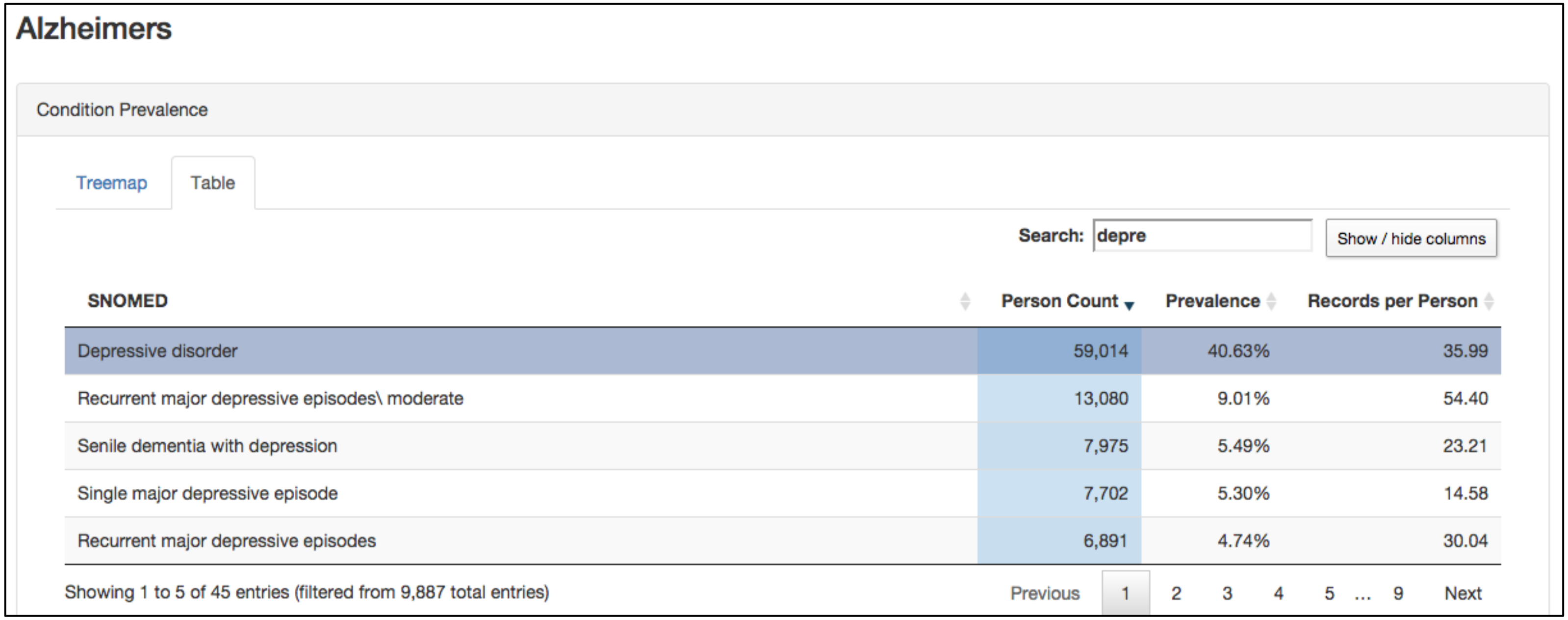
Source: INPC

Number of Persons: 145,246



Alzheimers

Number of Persons by Duration from Observation Start to Cohort Start to Observation End



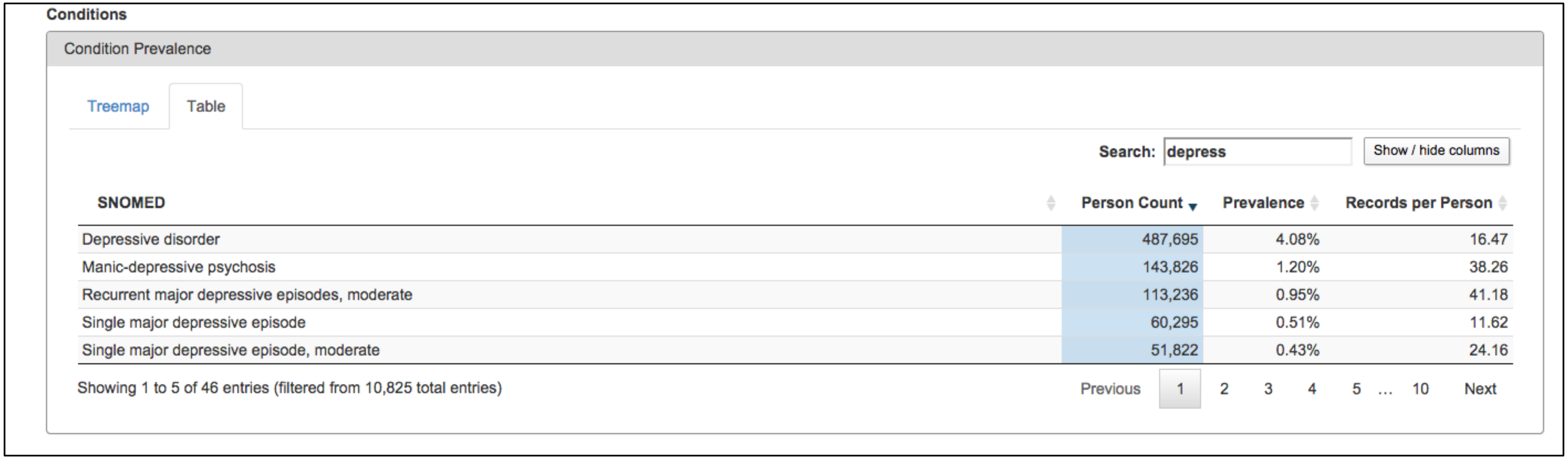
Alzheimers

Condition Prevalence

Search: depre

SNOMED	Person Count	Prevalence	Records per Person
Depressive disorder	59,014	40.63%	35.99
Recurrent major depressive episodes\ moderate	13,080	9.01%	54.40
Senile dementia with depression	7,975	5.49%	23.21
Single major depressive episode	7,702	5.30%	14.58
Recurrent major depressive episodes	6,891	4.74%	30.04

Showing 1 to 5 of 45 entries (filtered from 9,887 total entries)



Conditions

Condition Prevalence

Search: depress

SNOMED	Person Count	Prevalence	Records per Person
Depressive disorder	487,695	4.08%	16.47
Manic-depressive psychosis	143,826	1.20%	38.26
Recurrent major depressive episodes, moderate	113,236	0.95%	41.18
Single major depressive episode	60,295	0.51%	11.62
Single major depressive episode, moderate	51,822	0.43%	24.16

Showing 1 to 5 of 46 entries (filtered from 10,825 total entries)

HERACLES = Specialist

- Can limit to specific analyses (e.g., just procedures)
- Can target specific concepts (e.g., a drug class, a particular condition)
- Can window on cohort-specific date ranges

HERACLES

- Designed to be run many times per CDM
  - New cohorts
  - New target areas of interest
- Official release in April
  - Both ACHILLES and HERACLES are part of a suite of OHDSI tools available on GitHub

Population-level Estimation

Patrick Ryan, PhD

Janssen Research and Development

25 March 2015

Questions OHDSI Seeks to Answer from Observational Data

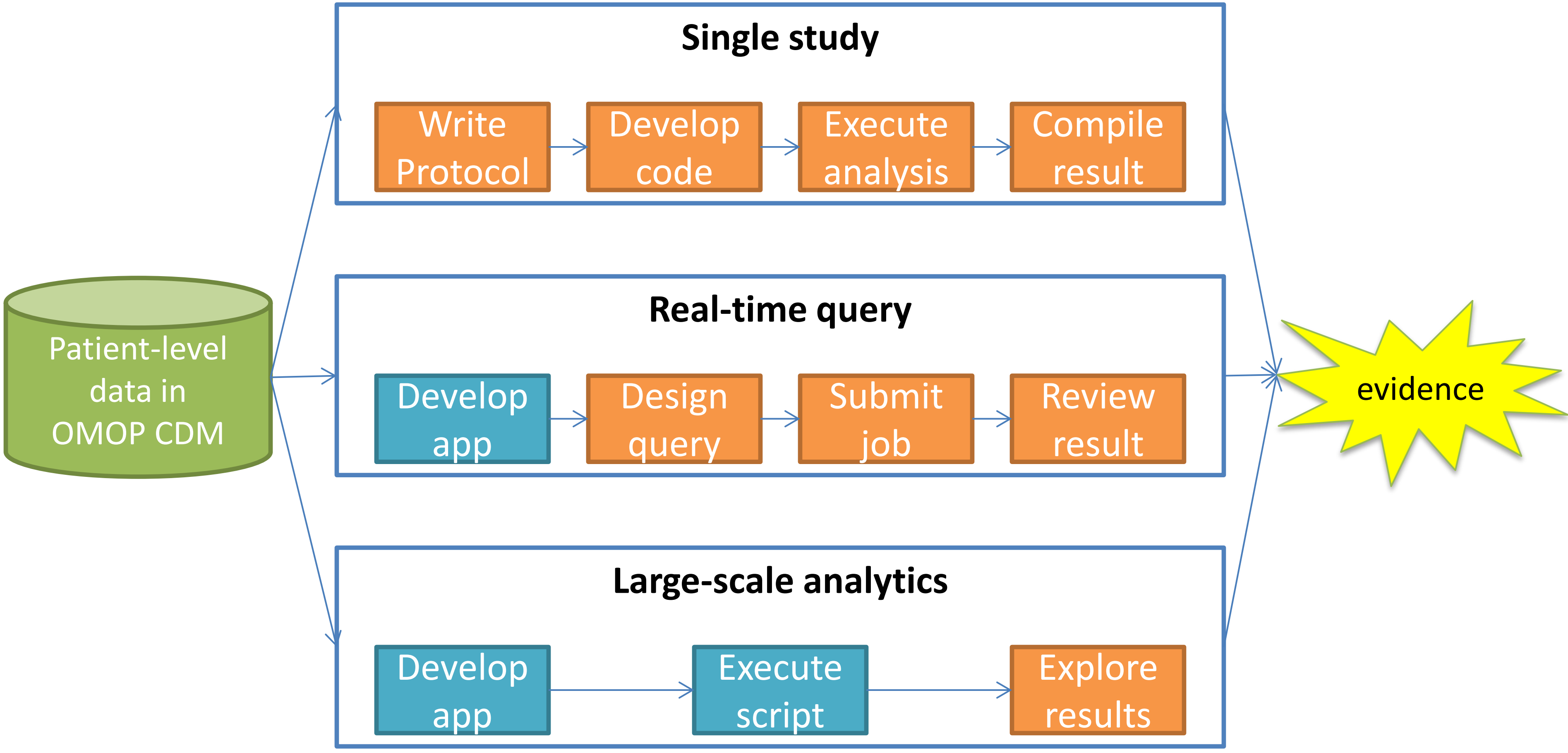
- Clinical characterization:
  - Natural history: Who are the patients who have diabetes? Among those patients, who takes metformin?
  - Quality improvement: what proportion of patients with diabetes experience disease-related complications?
- Population-level estimation
  - Safety surveillance: Does metformin cause lactic acidosis?
  - Comparative effectiveness: Does metformin cause lactic acidosis more than glyburide?
- Patient-level prediction
  - Given everything you know about me and my medical history, if I start taking metformin, what is the chance that I am going to have lactic acidosis in the next year?

Opportunities for Standardization in the Evidence Generation Process

Protocol

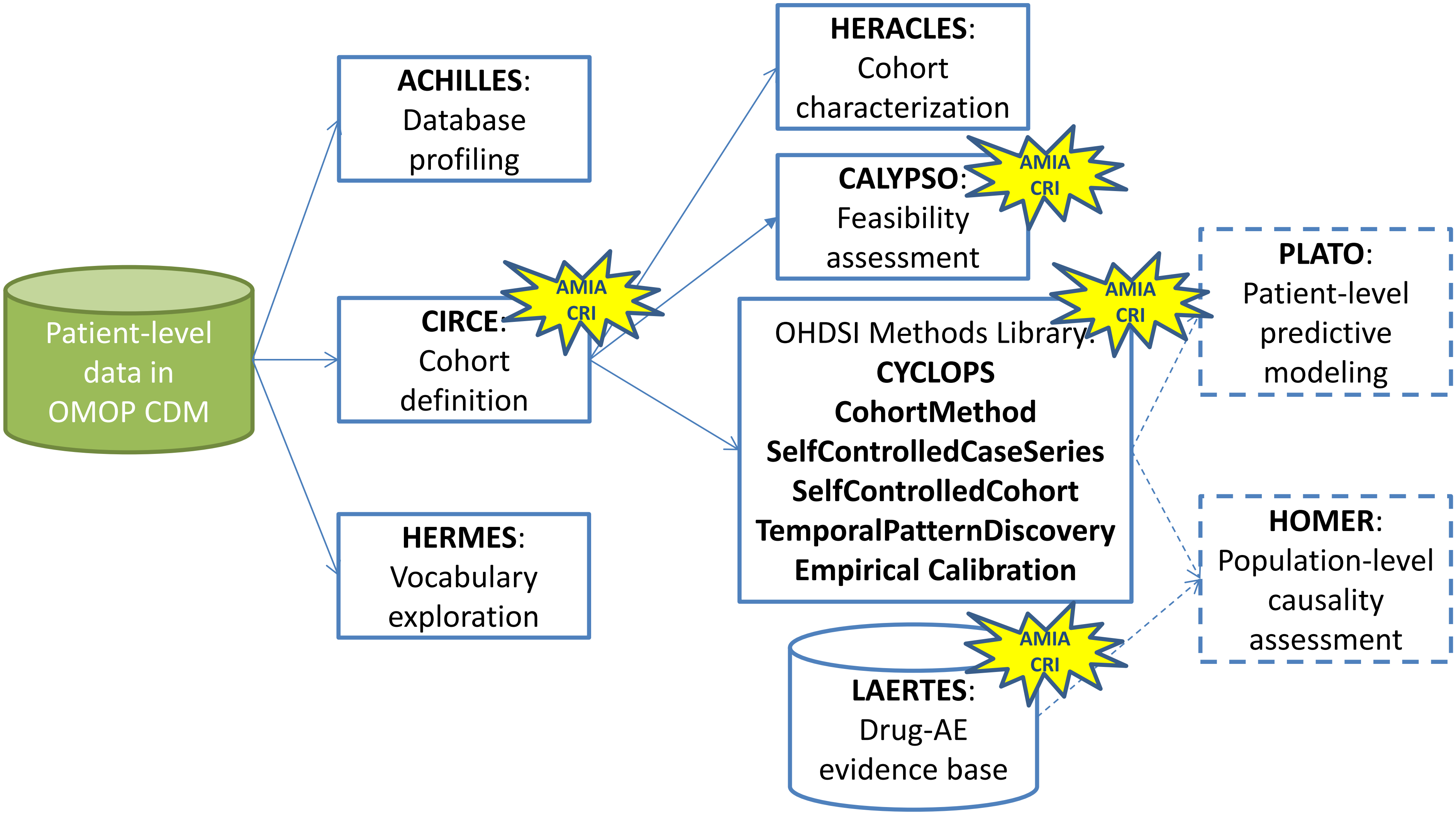
- **Data structure** : tables, fields, data types
- **Data content** : vocabulary to codify clinical domains
- **Data semantics** : conventions about meaning
- **Cohort definition** : algorithms for identifying the set of patients who meet a collection of criteria for a given interval of time
- **Covariate construction** : logic to define variables available for use in statistical analysis
- **Analysis** : collection of decisions and procedures required to produce aggregate summary statistics from patient-level data
- **Results reporting** : series of aggregate summary statistics presented in tabular and graphical form

~~Data~~ Evidence Sharing Paradigms



🟦 One-time 🟧 Repeated

Standardized Large-scale Analytics Tools Under Development within OHDSI



Patient-level data in OMOP CDM

ACHILLES: Database profiling

CIRCE: Cohort definition

HERMES: Vocabulary exploration

HERACLES: Cohort characterization

CALYPSO: Feasibility assessment

OHDSI Methods Library:
CYCLOPS
CohortMethod
SelfControlledCaseSeries
SelfControlledCohort
TemporalPatternDiscovery
Empirical Calibration

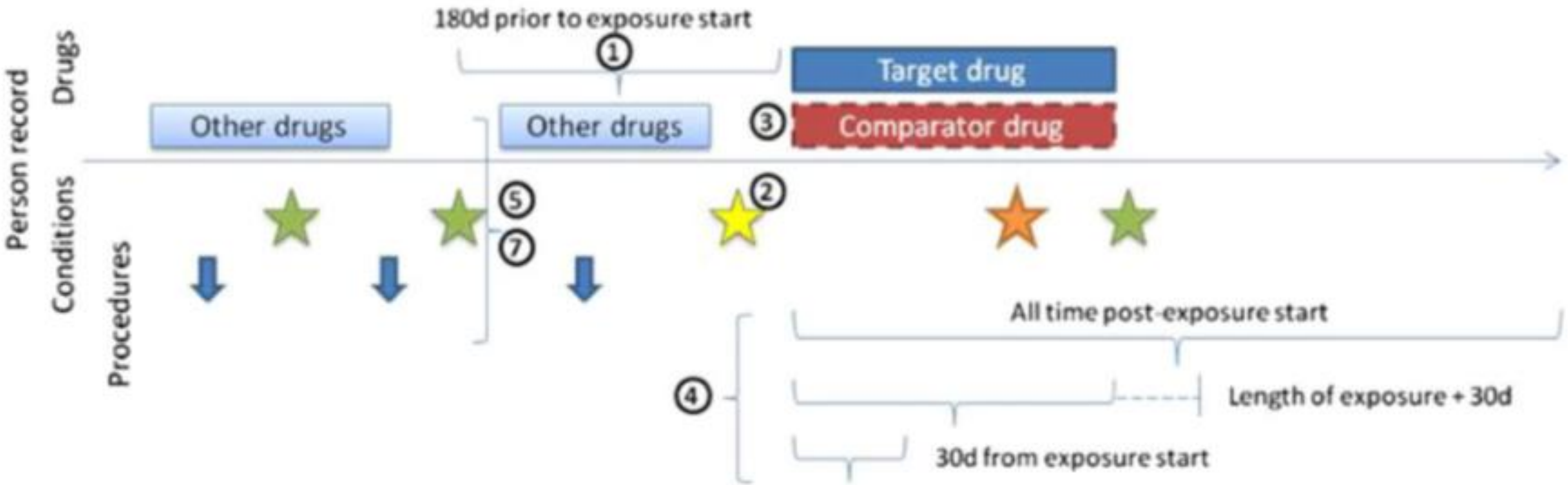
LAERTES: Drug-AE evidence base

PLATO: Patient-level predictive modeling

HOMER: Population-level causality assessment

http://github.com/OHDSI

Standardizing Analytic Decisions in Cohort Studies

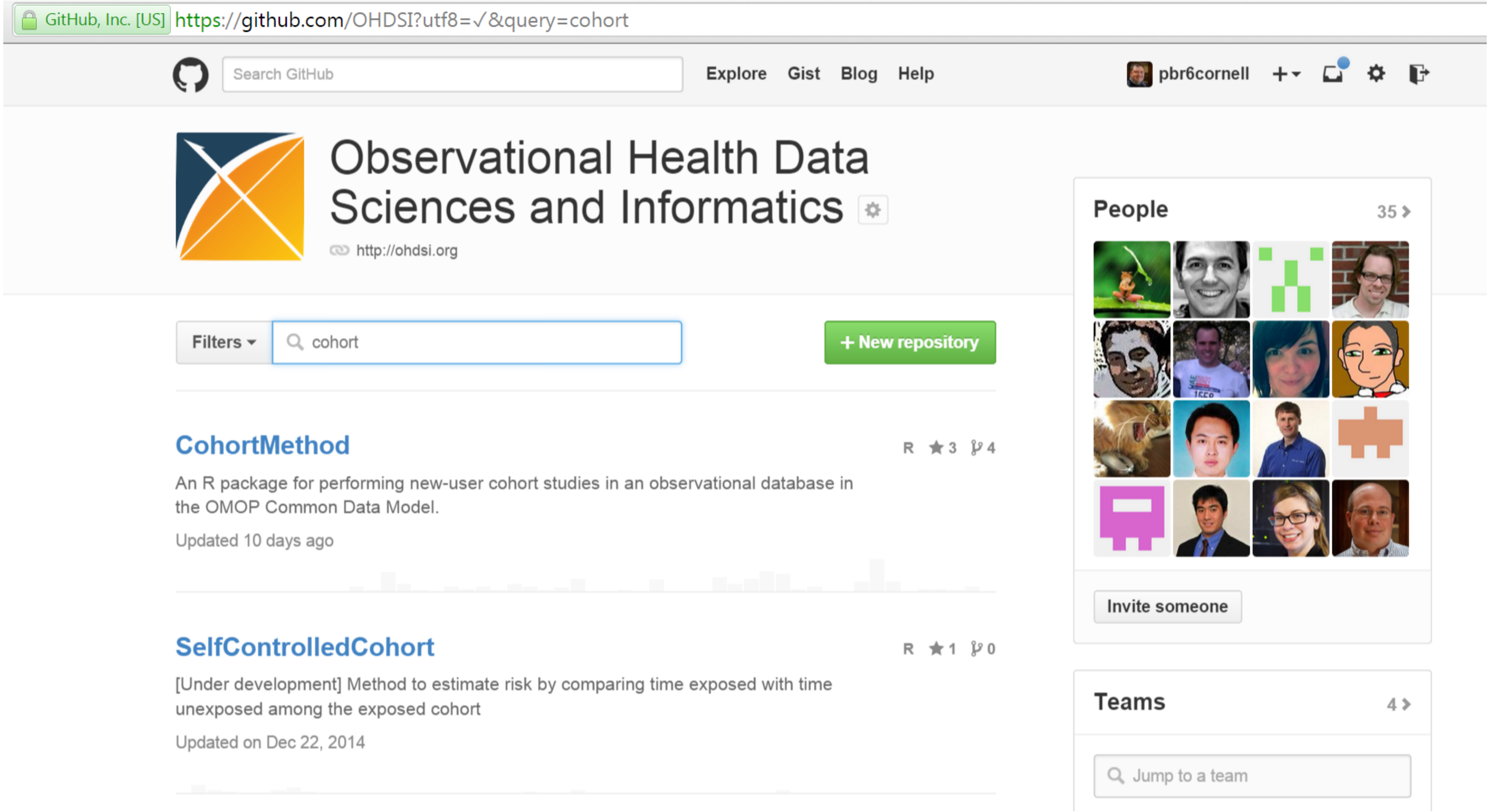


Decisions a researcher needs to make

→ parameters a standardized analytic routine needs to accommodate:

1. Washout period length
2. Nesting cohorts within indication
3. Comparator population
4. Time-at-risk
5. Propensity score covariate selection strategy
6. Covariate eligibility window
7. Propensity score adjustment strategy (trimming, stratification, matching)
8. Outcome model

Standardized Analytics to Enable Reproducible Research



GitHub, Inc. [US] https://github.com/OHDSI?utf8=✓&query=cohort

Observational Health Data Sciences and Informatics

http://ohdsi.org

CohortMethod R ★3 4

An R package for performing new-user cohort studies in an observational database in the OMOP Common Data Model.

Updated 10 days ago

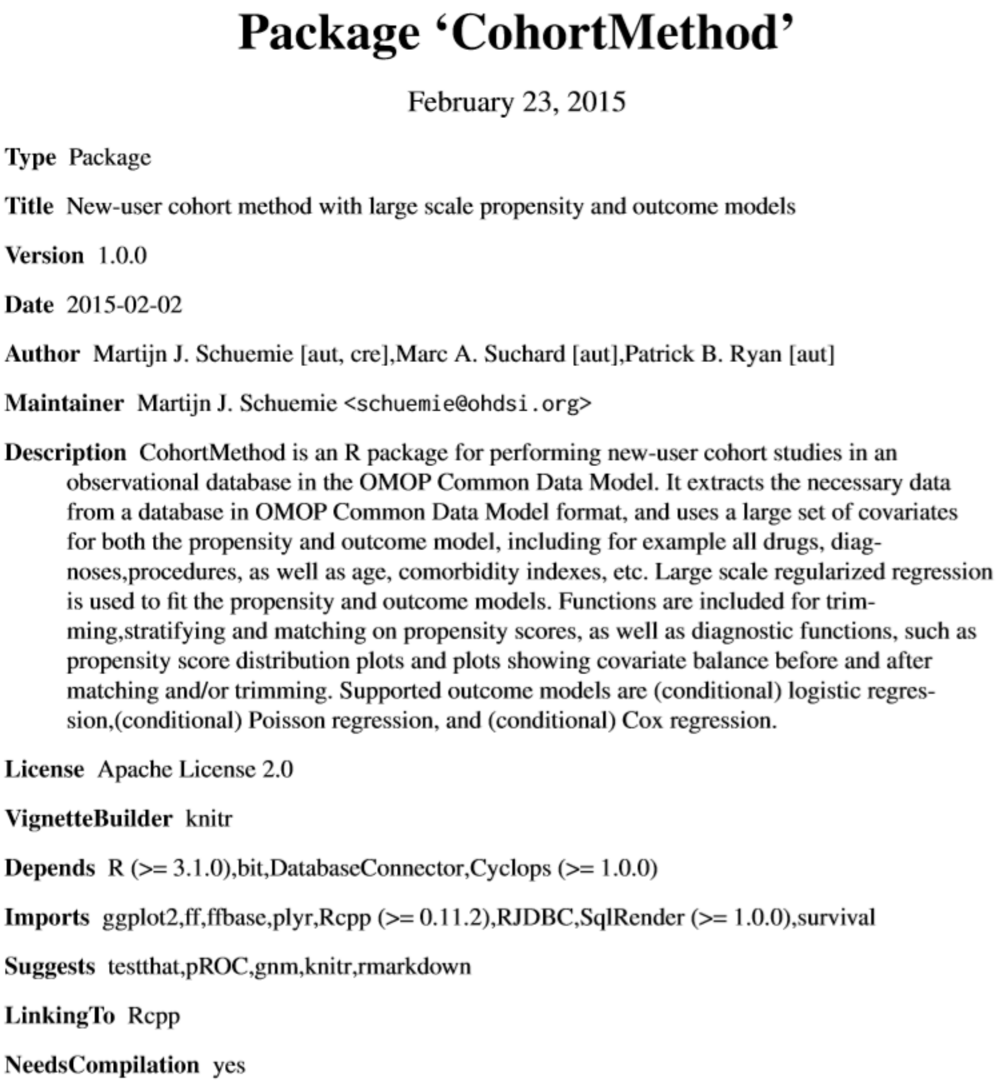
SelfControlledCohort R ★1 0

[Under development] Method to estimate risk by comparing time exposed with time unexposed among the exposed cohort

Updated on Dec 22, 2014

http://github.com/OHDSI

Open-source Large-scale Analytics through R



Package 'CohortMethod'

February 23, 2015

Type Package

Title New-user cohort method with large scale propensity and outcome models

Version 1.0.0

Date 2015-02-02

Author Martijn J. Schuemie [aut, cre],Marc A. Suchard [aut],Patrick B. Ryan [aut]

Maintainer Martijn J. Schuemie <schuemie@ohdsi.org>

Description CohortMethod is an R package for performing new-user cohort studies in an observational database in the OMOP Common Data Model. It extracts the necessary data from a database in OMOP Common Data Model format, and uses a large set of covariates for both the propensity and outcome model, including for example all drugs, diagnoses,procedures, as well as age, comorbidity indexes, etc. Large scale regularized regression is used to fit the propensity and outcome models. Functions are included for trimming,stratifying and matching on propensity scores, as well as diagnostic functions, such as propensity score distribution plots and plots showing covariate balance before and after matching and/or trimming. Supported outcome models are (conditional) logistic regression,(conditional) Poisson regression, and (conditional) Cox regression.

License Apache License 2.0

VignetteBuilder knitr

Depends R (>= 3.1.0),bit,DatabaseConnector,Cyclops (>= 1.0.0)

Imports ggplot2,ff,ffbase,plyr,Rcpp (>= 0.11.2),RJDBC,SqlRender (>= 1.0.0),survival

Suggests testthat,pROC,gnm,knitr,rmarkdown

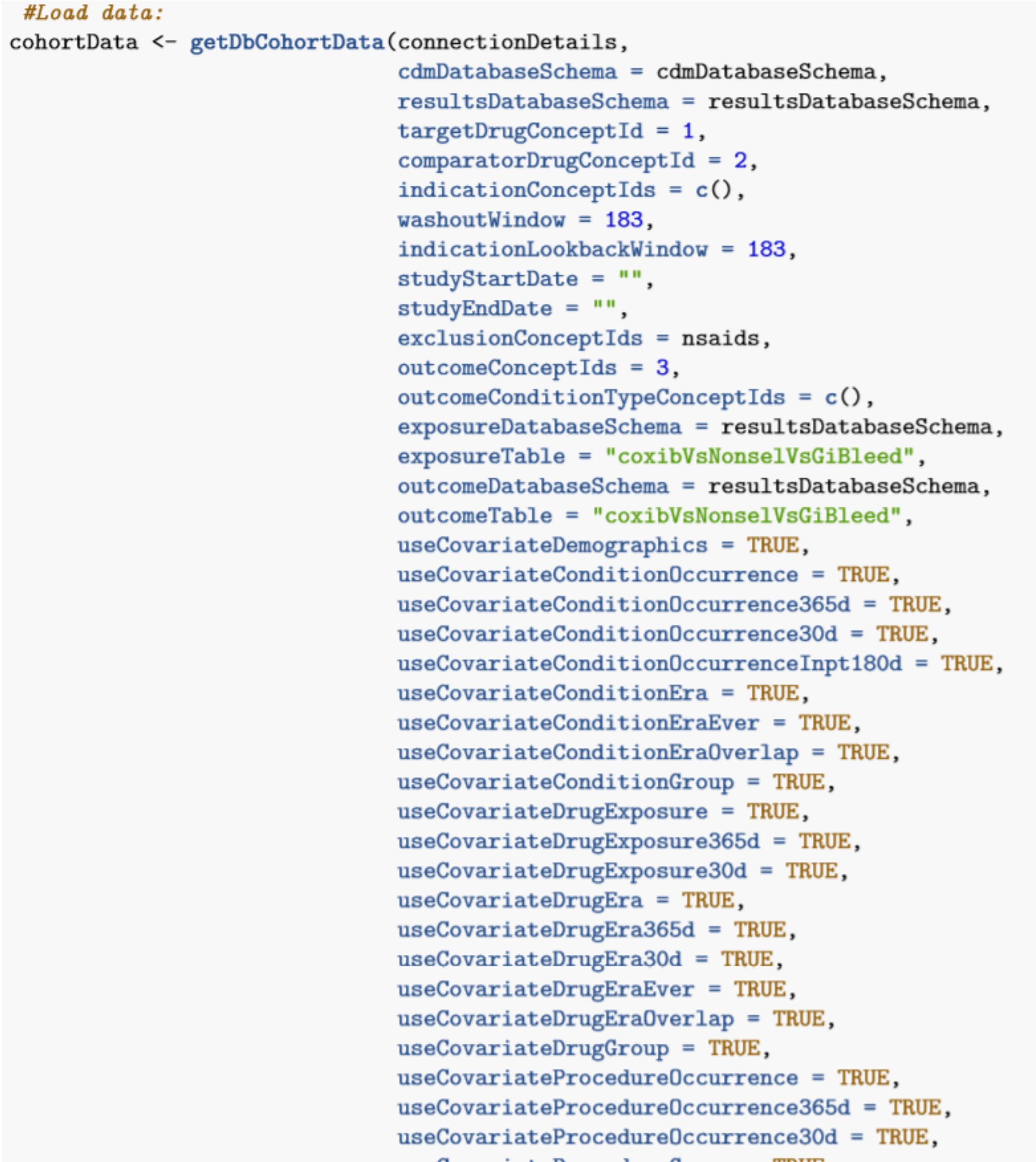
LinkingTo Rcpp

NeedsCompilation yes

Why is this a novel approach?

- Large-scale analytics, scalable to 'big data' problems in healthcare:
  - millions of patients
  - millions of covariates
  - millions of questions
- End-to-end analysis, from CDM through evidence
  - No longer de-coupling 'informatics' from 'statistics' from 'epidemiology'

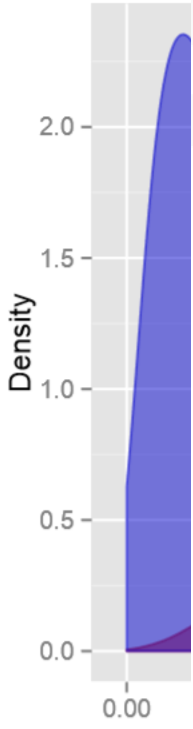
Standardize Covariate Construction



```
#Load data:
cohortData <- getDbCohortData(connectionDetails,
                              cdmDatabaseSchema = cdmDatabaseSchema,
                              resultsDatabaseSchema = resultsDatabaseSchema,
                              targetDrugConceptId = 1,
                              comparatorDrugConceptId = 2,
                              indicationConceptIds = c(),
                              washoutWindow = 183,
                              indicationLookbackWindow = 183,
                              studyStartDate = "",
                              studyEndDate = "",
                              exclusionConceptIds = nsaids,
                              outcomeConceptIds = 3,
                              outcomeConditionTypeConceptIds = c(),
                              exposureDatabaseSchema = resultsDatabaseSchema,
                              exposureTable = "coxibVsNonselVsGiBleed",
                              outcomeDatabaseSchema = resultsDatabaseSchema,
                              outcomeTable = "coxibVsNonselVsGiBleed",
                              useCovariateDemographics = TRUE,
                              useCovariateConditionOccurrence = TRUE,
                              useCovariateConditionOccurrence365d = TRUE,
                              useCovariateConditionOccurrence30d = TRUE,
                              useCovariateConditionOccurrenceInpt180d = TRUE,
                              useCovariateConditionEra = TRUE,
                              useCovariateConditionEraEver = TRUE,
                              useCovariateConditionEraOverlap = TRUE,
                              useCovariateConditionGroup = TRUE,
                              useCovariateDrugExposure = TRUE,
                              useCovariateDrugExposure365d = TRUE,
                              useCovariateDrugExposure30d = TRUE,
                              useCovariateDrugEra = TRUE,
                              useCovariateDrugEra365d = TRUE,
                              useCovariateDrugEra30d = TRUE,
                              useCovariateDrugEraEver = TRUE,
                              useCovariateDrugEraOverlap = TRUE,
                              useCovariateDrugGroup = TRUE,
                              useCovariateProcedureOccurrence = TRUE,
                              useCovariateProcedureOccurrence365d = TRUE,
                              useCovariateProcedureOccurrence30d = TRUE,
```

Standardize Model Diagnostics

```
plotPs(ps, scale = "preference")
```

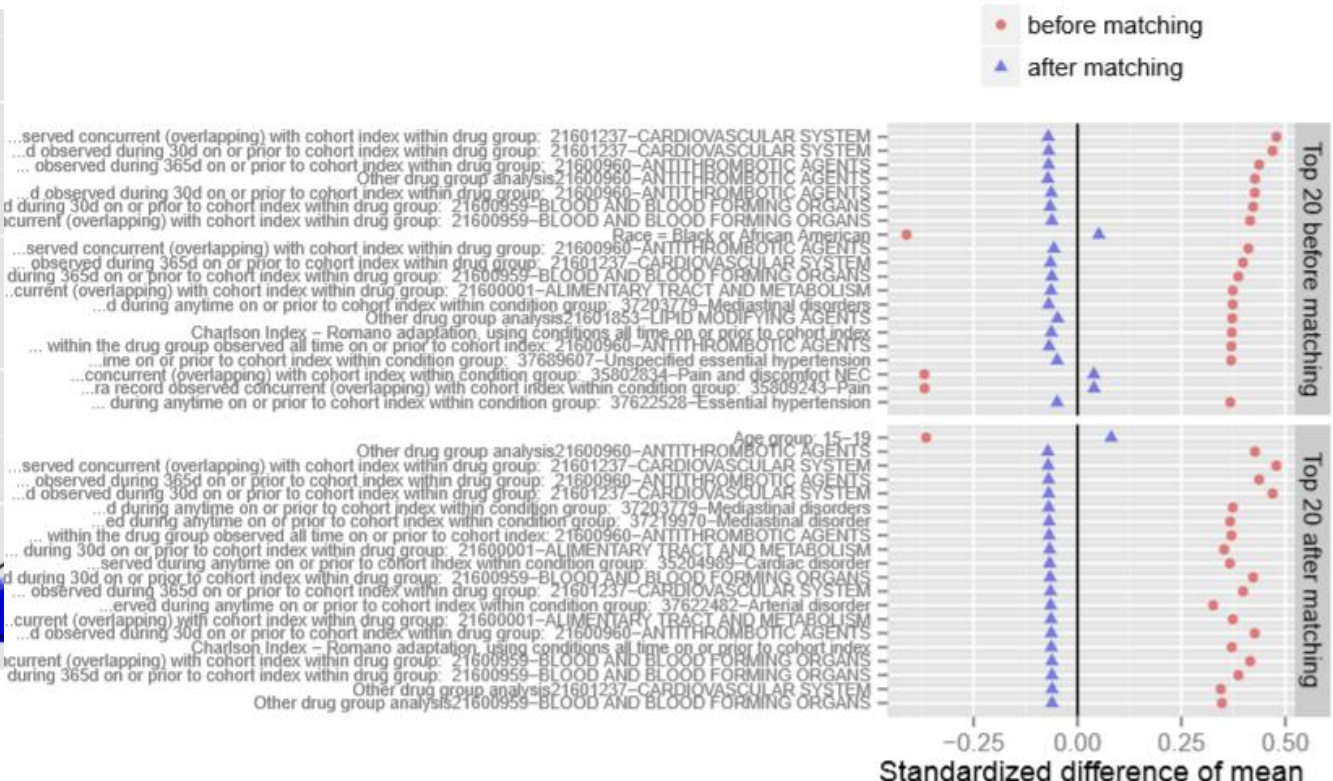


```
balance <- computeCovariateBalance(strata, cohortData, outcomeConceptId = 3)
```

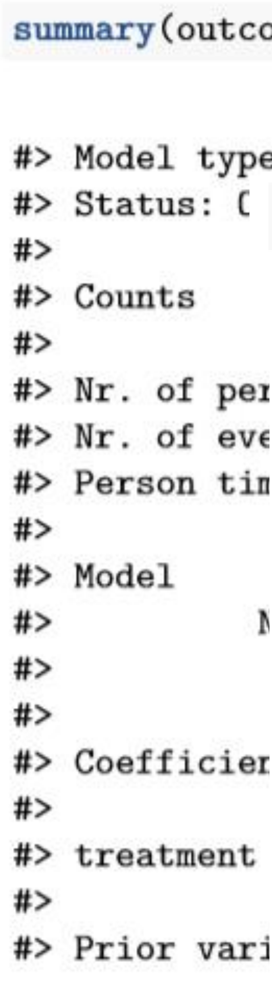
```
plotCovariateBalanceScatterPlot(balance)
```

After matching
0.5
0.4
0.3
0.2
0.1
0.0
0.0

```
plotCovariateBalanceOfTopVariables(balance)
```



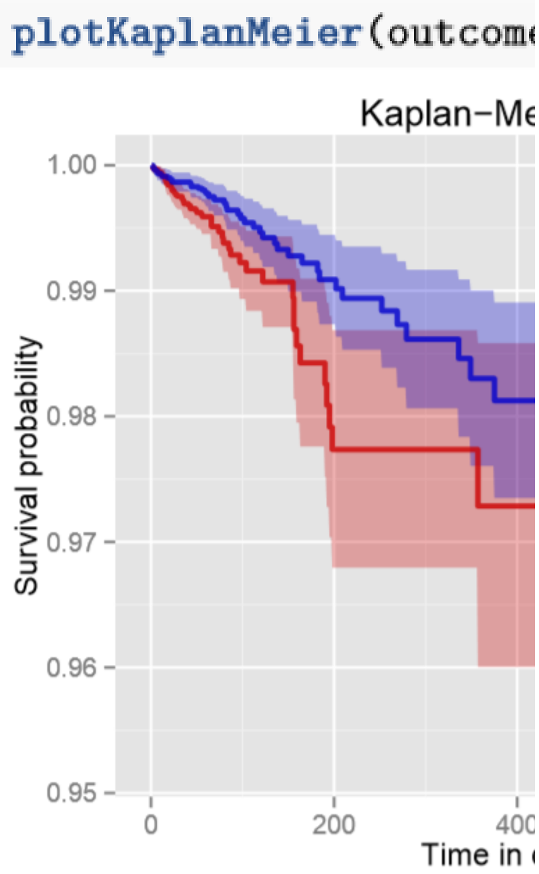
Standardize Analysis and Results Reporting



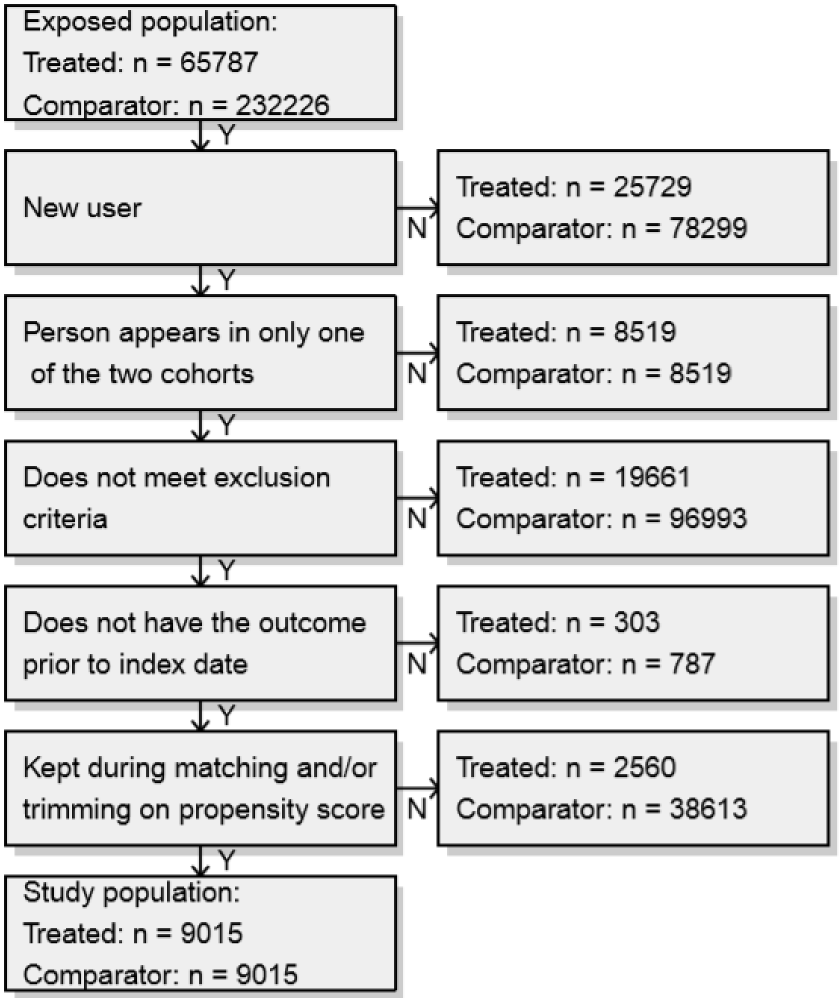
```
summary(outcomeModel)
```

```
#> Model type: cox
#> Status: (
#>
#> Counts
#>
#> Nr. of per
#> Nr. of eve
#> Person tin
#>
#> Model
#>           N
#>
#>
#> Coefficier
#>
#> treatment
#>
#> Prior vari
```

```
plotKaplanMeier(outcome
```



```
drawAttritionDiagram(outcomeModel)
```



Exposed population:
Treated: n = 65787
Comparator: n = 232226

Y

New user

N

Treated: n = 25729
Comparator: n = 78299

Y

Person appears in only one of the two cohorts

N

Treated: n = 8519
Comparator: n = 8519

Y

Does not meet exclusion criteria

N

Treated: n = 19661
Comparator: n = 96993

Y

Does not have the outcome prior to index date

N

Treated: n = 303
Comparator: n = 787

Y

Kept during matching and/or trimming on propensity score

N

Treated: n = 2560
Comparator: n = 38613

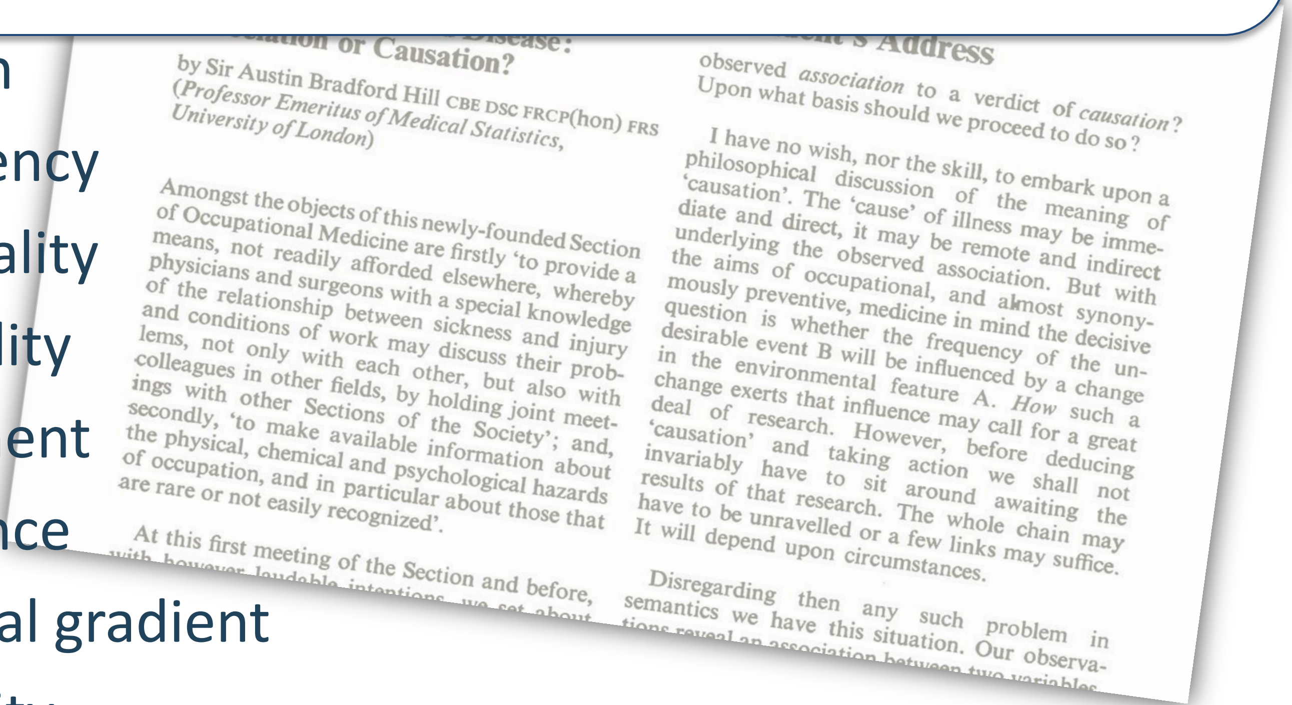
Y

Study population:
Treated: n = 9015
Comparator: n = 9015

To Go Forward, We Must Go Back

“What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?”

- Strength
- Consistency
- Temporality
- Plausibility
- Experiment
- Coherence
- Biological gradient
- Specificity
- Analogy



or Causation?

by Sir Austin Bradford Hill CBE DSC FRCP(hon) FRS
(*Professor Emeritus of Medical Statistics, University of London*)

Amongst the objects of this newly-founded Section of Occupational Medicine are firstly ‘to provide a means, not readily afforded elsewhere, whereby physicians and surgeons with a special knowledge of the relationship between sickness and injury and conditions of work may discuss their problems, not only with each other, but also with colleagues in other fields, by holding joint meetings with other Sections of the Society’; and, secondly, ‘to make available information about the physical, chemical and psychological hazards of occupation, and in particular about those that are rare or not easily recognized’.

At this first meeting of the Section and before, with however laudable intentions, we set about

Address

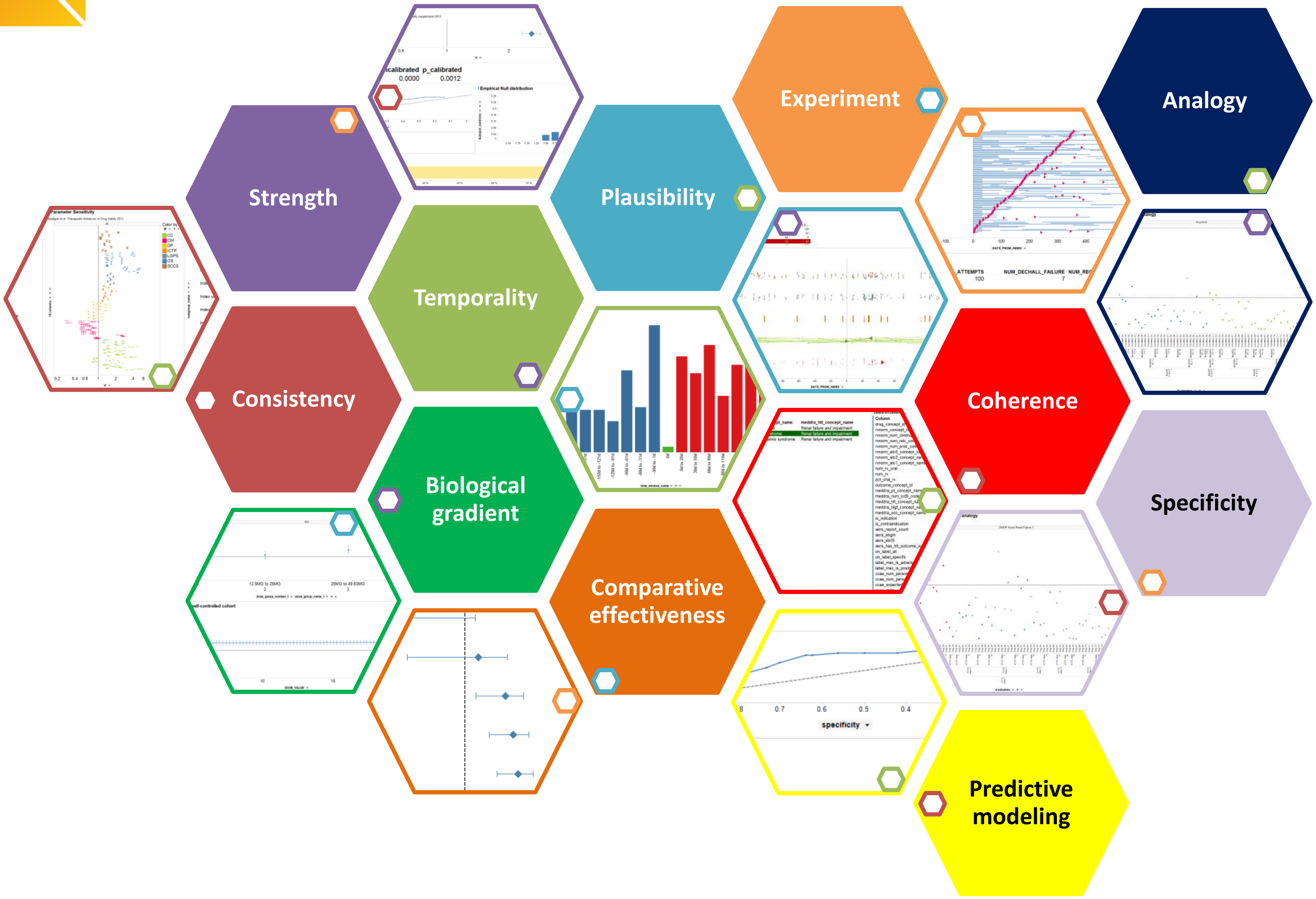
observed *association* to a verdict of *causation*? Upon what basis should we proceed to do so?

I have no wish, nor the skill, to embark upon a philosophical discussion of the meaning of ‘causation’. The ‘cause’ of illness may be immediate and direct, it may be remote and indirect underlying the observed association. But with the aims of occupational, and almost synonymously preventive, medicine in mind the decisive question is whether the frequency of the undesirable event B will be influenced by a change in the environmental feature A. *How* such a change exerts that influence may call for a great deal of research. However, before deducing ‘causation’ and taking action we shall not invariably have to sit around awaiting the results of that research. The whole chain may have to be unravelled or a few links may suffice. It will depend upon circumstances.

Disregarding then any such problem in semantics we have this situation. Our observations reveal an association between two variables

Austin Bradford Hill, “The Environment and Disease: Association or Causation?,” *Proceedings of the Royal Society of Medicine*, 58 (1965), 295-300.

HOMER Implementation of Hill's Viewpoints



Concluding Thoughts

- We need to build informatics solutions to enable reliable, scalable evidence generation for population-level estimation
- Open-source large-scale analytics on a common data platform are required to facilitate efficient, transparent, and reproducible science
- A multi-disciplinary, community approach can greatly accelerate the research and development of shares solutions

Personalized Risk Prediction

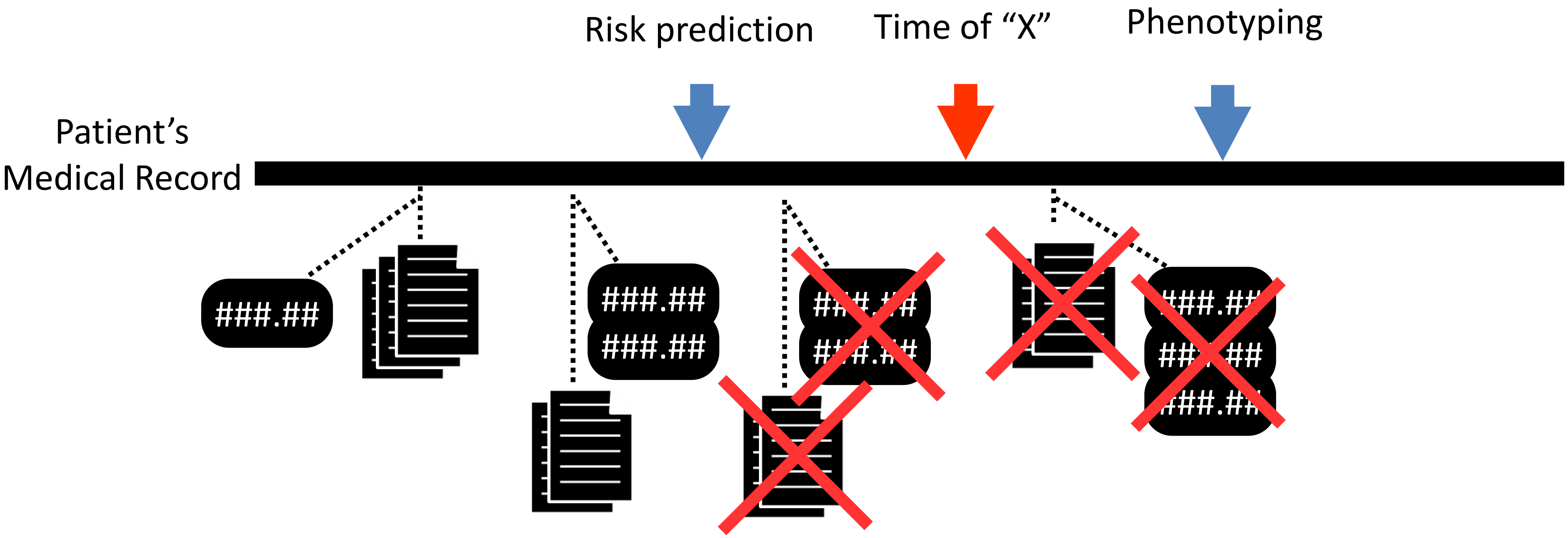
Nigam H. Shah MBBS, PhD
Assistant Professor
Dept. of Medicine (Biomedical Informatics)
Stanford University

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

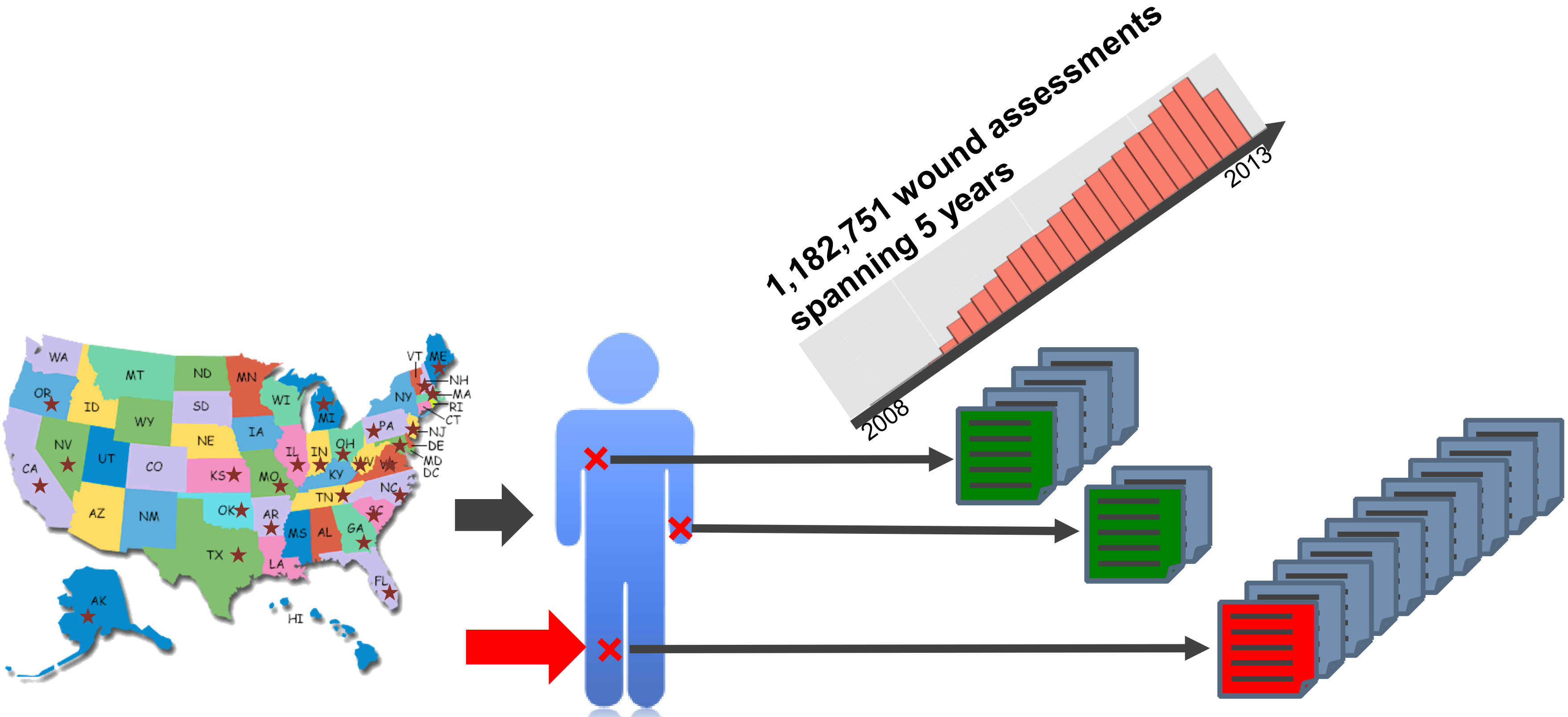
STANFORD
SCHOOL OF MEDICINE

Stanford University Medical Center

Phenotyping and Risk Prediction



Dataset and Prediction Task



1,182,751 wound assessments spanning 5 years

68 Healogics Wound Care Centers in 26 states

- Center Code

59,958 patients

- Age
- Gender
- Insurance
- Zip code

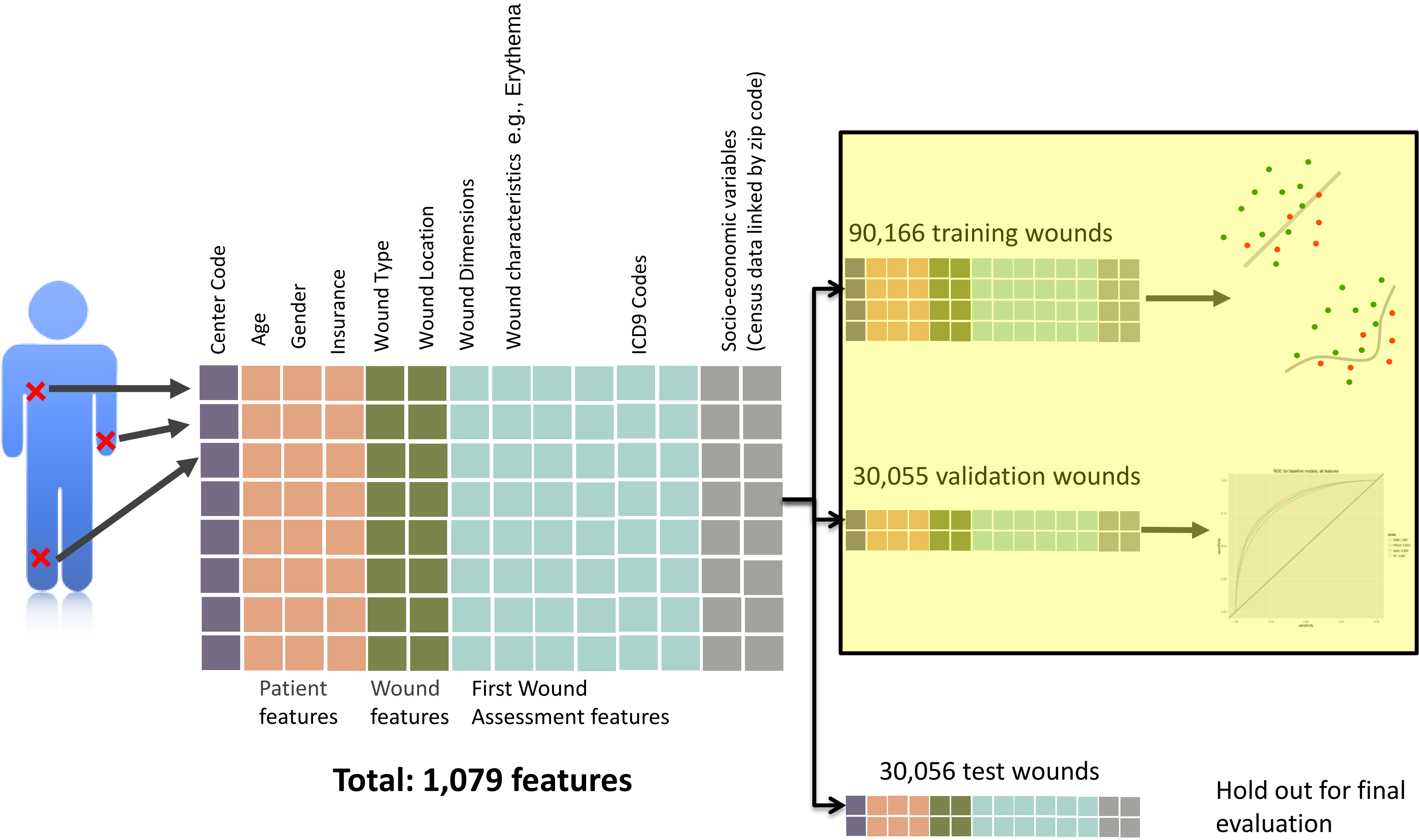
180,716 wounds

- Wound type
- Wound location

Each wound assessment:

- Dimensions
- Edema
- Erythema
- Rubor
- Other wound qualities

Setup and Feature Engineering



Center Code

Age

Gender

Insurance

Wound Type

Wound Location

Wound Dimensions

Wound characteristics e.g., Erythema

ICD9 Codes

Socio-economic variables (Census data linked by zip code)

Patient features

Wound features

First Wound Assessment features

Total: 1,079 features

90,166 training wounds

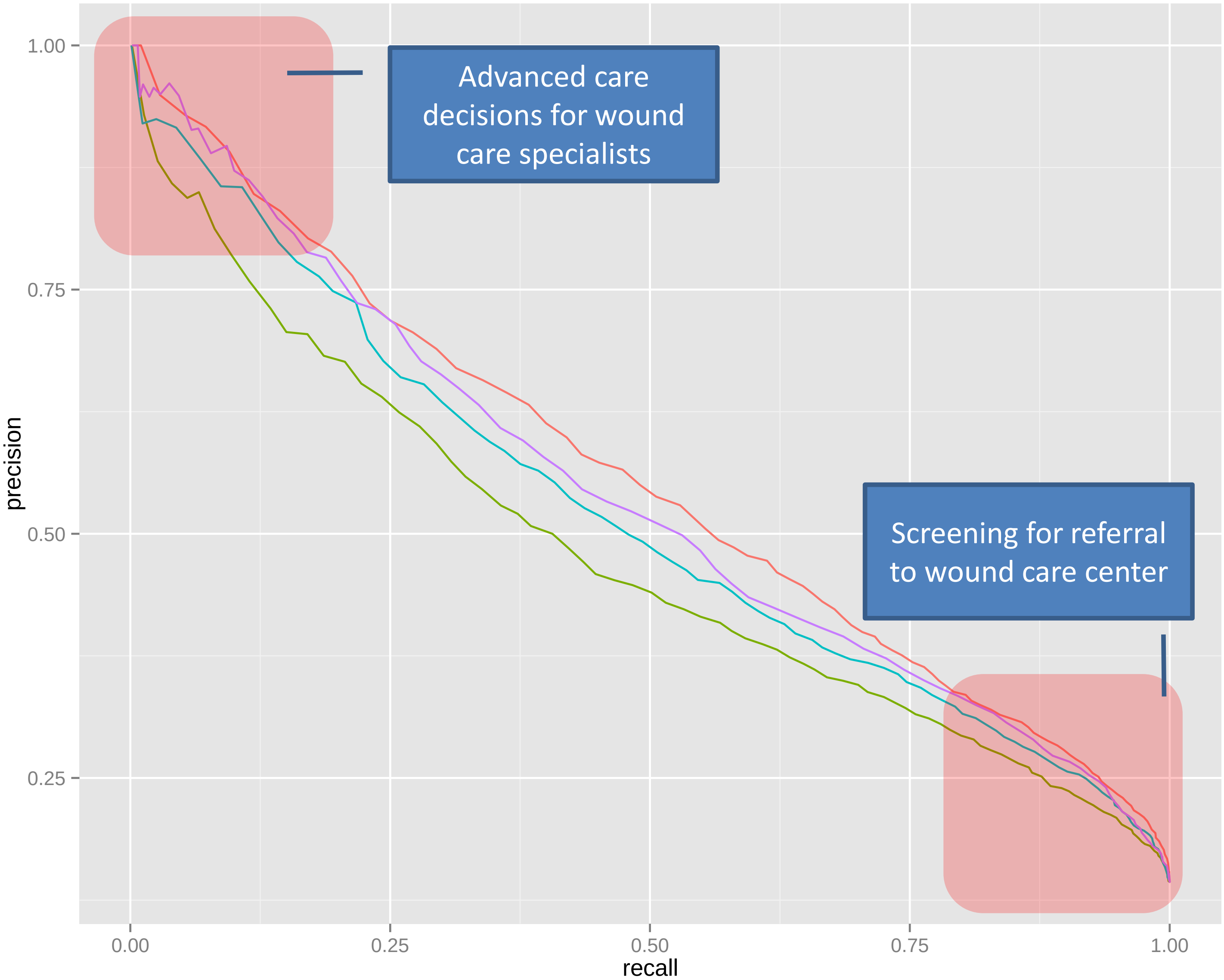
30,055 validation wounds

30,056 test wounds

Hold out for final evaluation

Summary

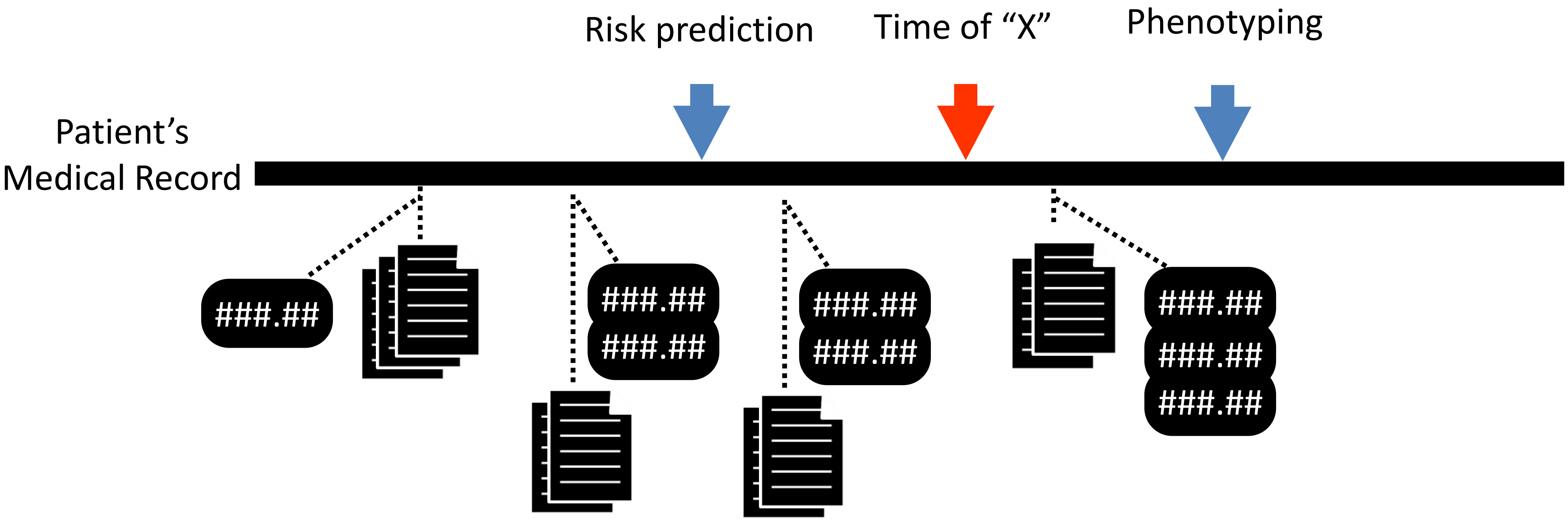
	Outlier at first visit
AUROC	0.857
Specificity	0.724
Sensitivity	0.796
Precision/ PPV	0.280



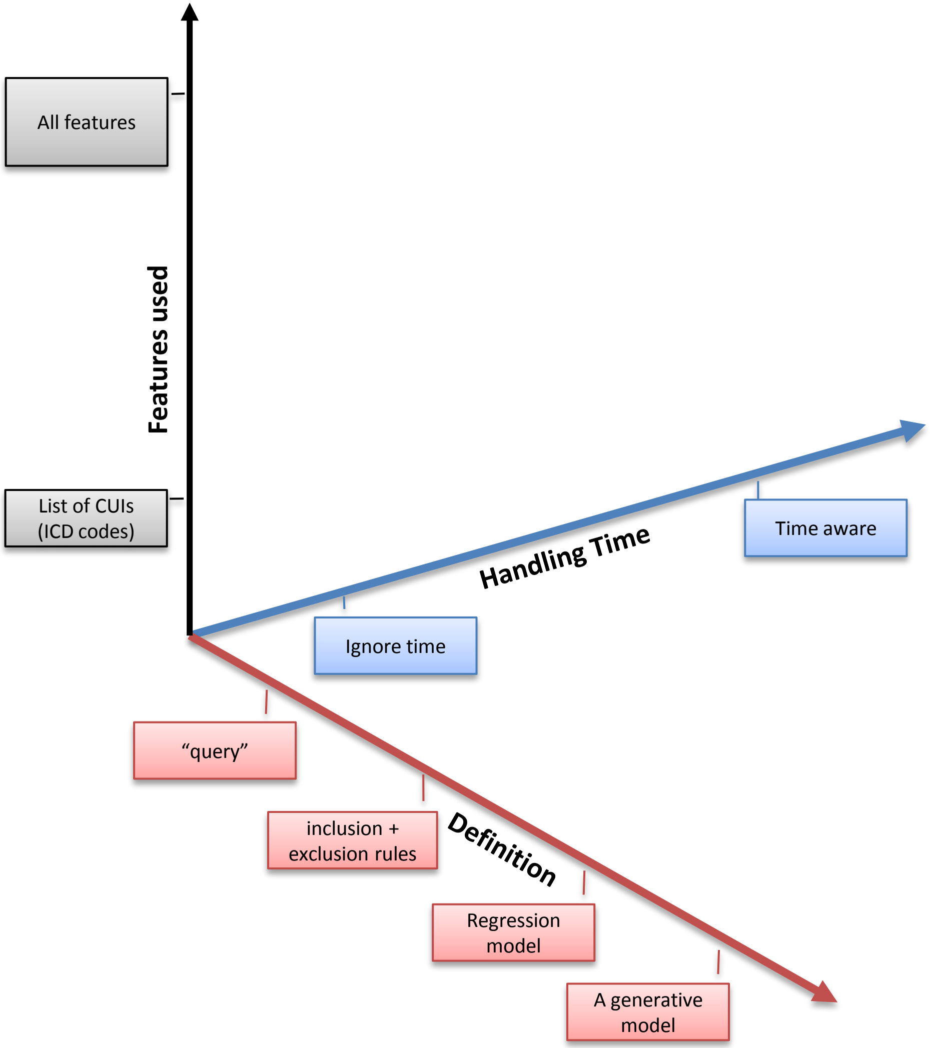
Advanced care decisions for wound care specialists

Screening for referral to wound care center

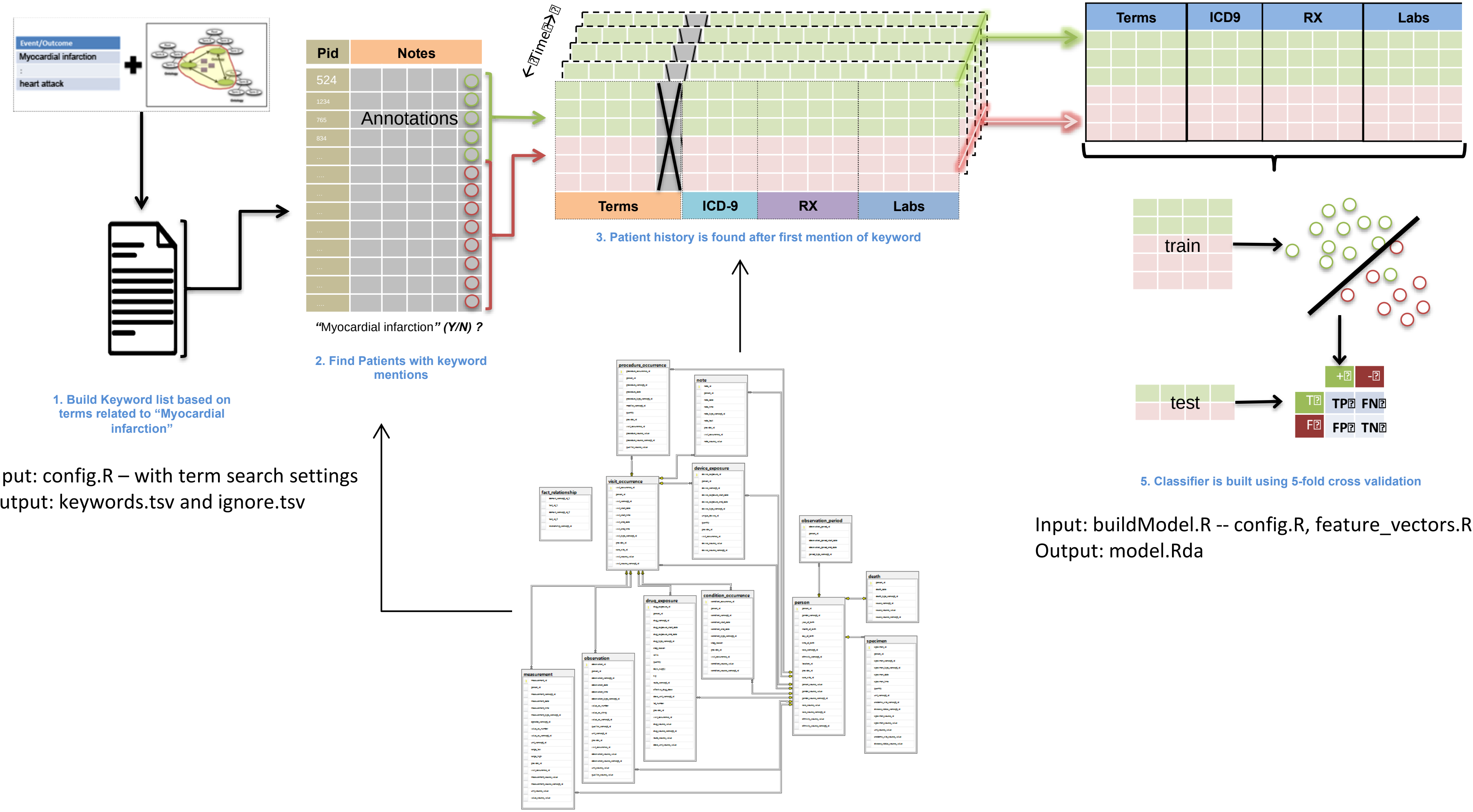
Phenotyping and Risk Prediction



Electronic Phenotyping



XPRESS- EXtraction of Phenotypes from clinical Records using Silver Standards



Input: getPatients.R -- config.R, keywords.tsv, ignore.tsv
Output: feature_vectors.Rda

1. Build Keyword list based on terms related to "Myocardial infarction"

Input: config.R – with term search settings
Output: keywords.tsv and ignore.tsv

"Myocardial infarction" *(Y/N) ?*

2. Find Patients with keyword mentions

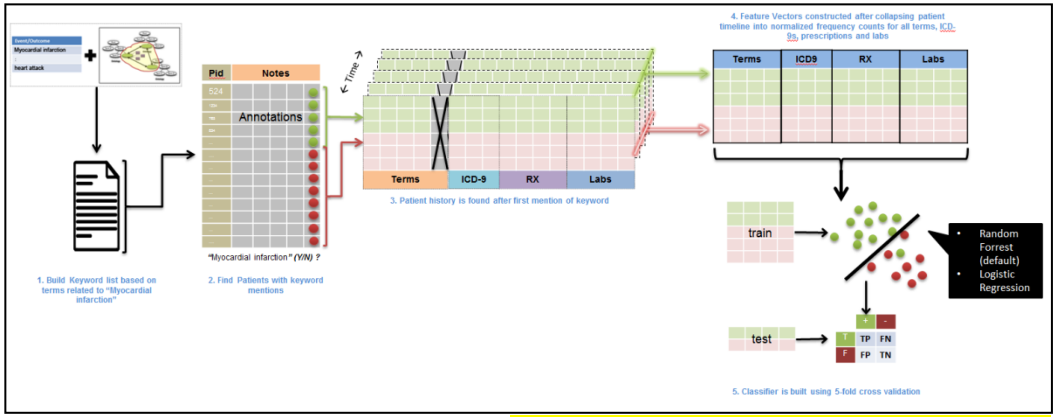
3. Patient history is found after first mention of keyword

4. Feature Vectors constructed after collapsing patient timeline into normalized frequency counts for all terms, ICD-9s, prescriptions and labs

5. Classifier is built using 5-fold cross validation

Input: buildModel.R -- config.R, feature_vectors.Rda
Output: model.Rda

XPRESS- EXtraction of Phenotypes from clinical Records using Silver Standards



1. Build Keyword list based on terms related to "Myocardial infarction"

2. Find Patients with keyword mentions

3. Patient history is found after first mention of keyword

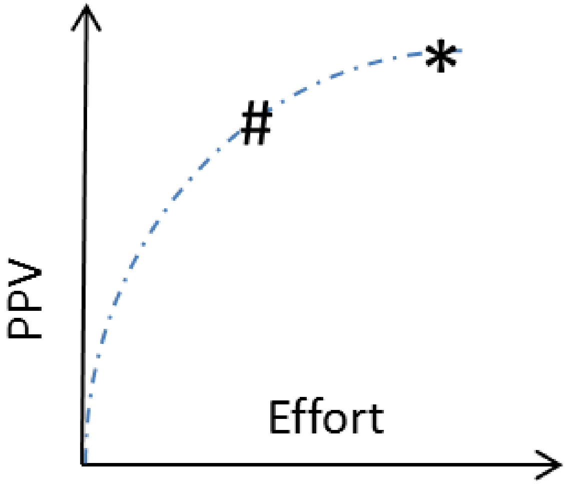
4. Feature Vectors constructed after collapsing patient timeline into normalized frequency counts for all terms, ICD-9s, prescriptions and labs

5. Classifier is built using 5-fold cross validation

- Random Forrest (default)
- Logistic Regression

predict.R
Input: config.R, model.Rda, classify.txt
Output: predictions.txt

Acute	Myocardial Infarction (OMOP)
Chronic	T2DM (PheKB)



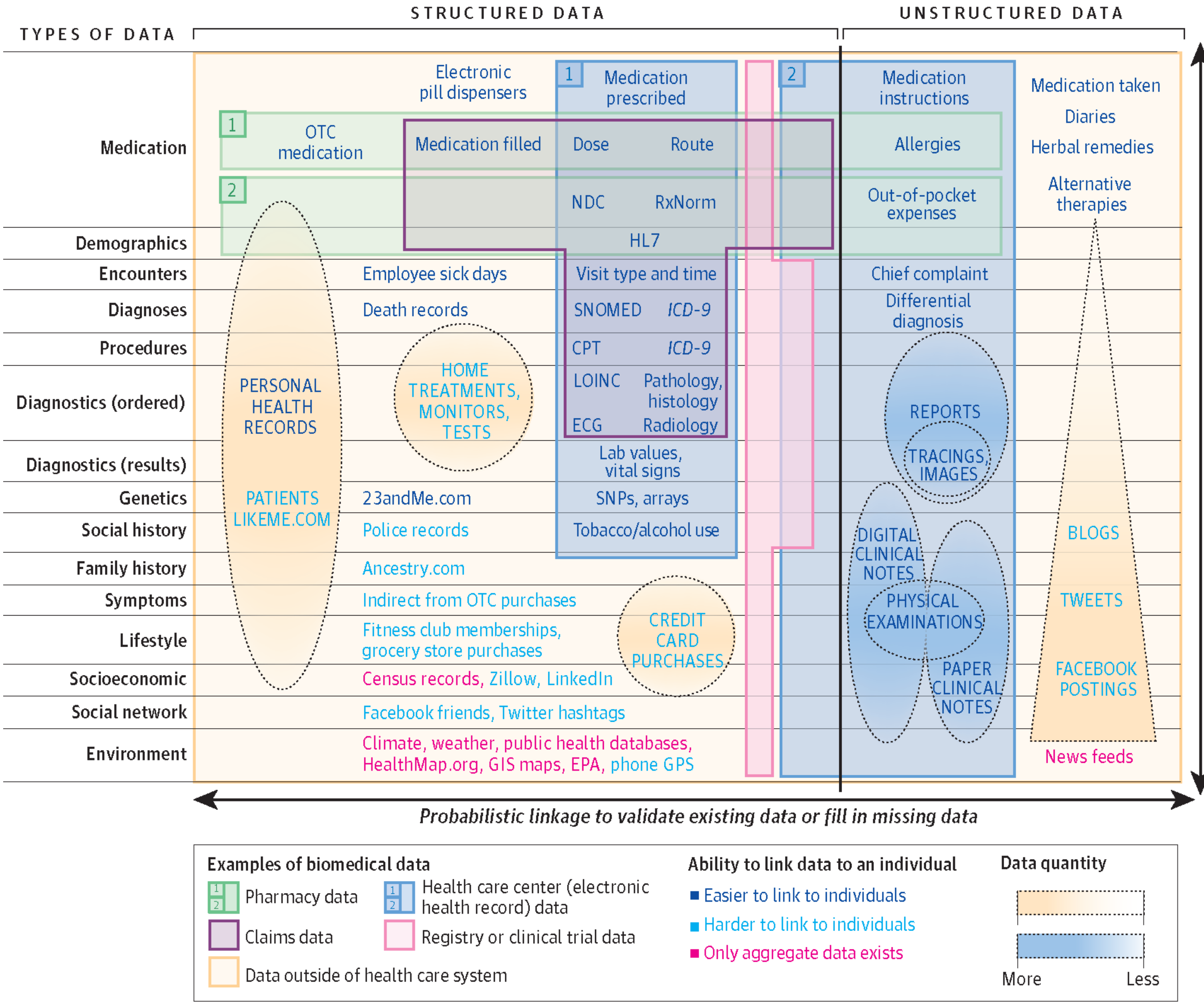
Acc	PPV	Time
0.87	0.84	?

Acc	PPV	Time
0.89	0.86	2hr

Acc	PPV	Time
0.98	0.96	1900

Acc	PPV	Time
0.89	0.90	2hr

The Sources of Features (Weber et al.)



TYPES OF DATA

STRUCTURED DATA

UNSTRUCTURED DATA

Medication: Electronic pill dispensers; 1 Medication prescribed; 2 Medication instructions; Medication taken; 1 OTC medication; Medication filled; Dose; Route; Allergies; Diaries; Herbal remedies; 2 NDC; RxNorm; Out-of-pocket expenses; Alternative therapies

Demographics: HL7

Encounters: Employee sick days; Visit type and time; Chief complaint

Diagnoses: Death records; SNOMED; *ICD-9*; Differential diagnosis

Procedures: CPT; *ICD-9*

Diagnostics (ordered): PERSONAL HEALTH RECORDS; HOME TREATMENTS, MONITORS, TESTS; LOINC; Pathology, histology; ECG; Radiology; REPORTS

Diagnostics (results): Lab values, vital signs; TRACINGS, IMAGES

Genetics: PATIENTS LIKEME.COM; 23andMe.com; SNPs, arrays

Social history: Police records; Tobacco/alcohol use; DIGITAL CLINICAL NOTES; BLOGS

Family history: Ancestry.com

Symptoms: Indirect from OTC purchases; PHYSICAL EXAMINATIONS; TWEETS

Lifestyle: Fitness club memberships, grocery store purchases; CREDIT CARD PURCHASES

Socioeconomic: Census records, Zillow, LinkedIn; PAPER CLINICAL NOTES; FACEBOOK POSTINGS

Social network: Facebook friends, Twitter hashtags

Environment: Climate, weather, public health databases, HealthMap.org, GIS maps, EPA, phone GPS; News feeds

Probabilistic linkage to obtain new types of data

Probabilistic linkage to validate existing data or fill in missing data

Examples of biomedical data

- Pharmacy data
- Health care center (electronic health record) data
- Claims data
- Registry or clinical trial data
- Data outside of health care system

Ability to link data to an individual

- ■ Easier to link to individuals
- ■ Harder to link to individuals
- ■ Only aggregate data exists

Data quantity

More — Less

Questions and Discussion

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

An Open Collaborative Approach for Rapid Evidence Generation

David K. Vawdrey, PhD
Jon D. Duke MD, MS
George Hripcsak MD, MS
Patrick Ryan PhD
Nigam H. Shah MBBS, PhD

AMIA Joint Summits on Translational Science
March 25, 2015