OHDSI Cohort Definition and Phenotyping

Jon Duke, Chris Knoll, Nigam Shah, Juan Banda

Introductions

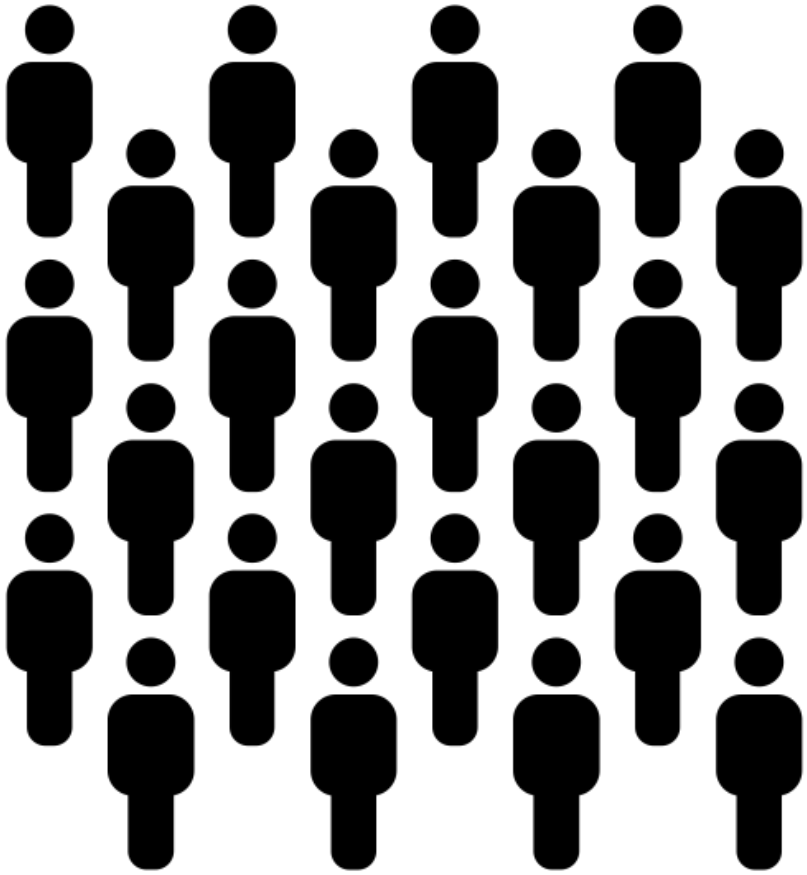
What You Will Learn Today

- What are phenotypes and what they have to do with observational data
- OHDSI Approach of Phenotyping
- Basics of rule-based phenotypes
- Basics of probabilistic phenotypes

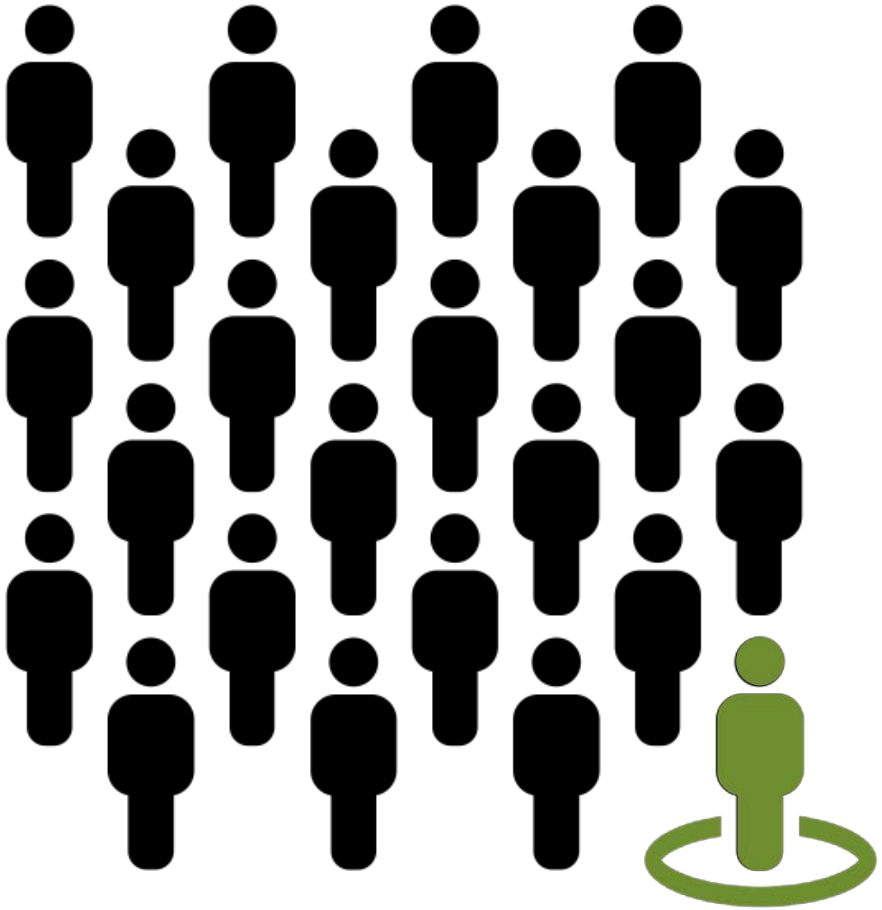
What You *Might* Learn Today

- A bit about the OHDSI cohort definition tools
- A bit about OHDSI R packages
- A bit about the OMOP vocabularies

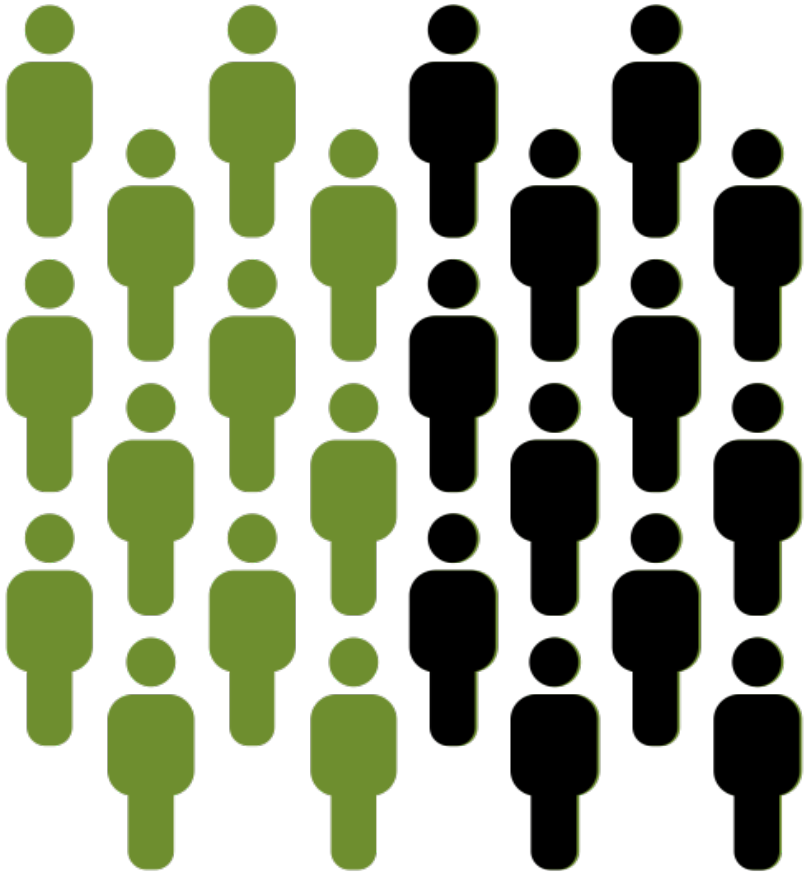
Let’s Start with People



Let’s Start with People



Let’s Start with People



Find People of Interest

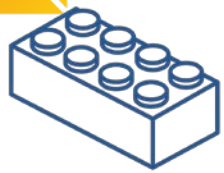
- One of the things we do at the beginning of any study involving observational data is find the people want to study
  - People who have the condition of interest
  - People who have had the intervention we want to study

What tools do we have at our disposal to identify these patients?

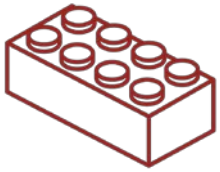
Data

- Patrick's Figure Here

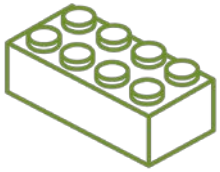
Data are Like Lego Bricks for Phenotypng



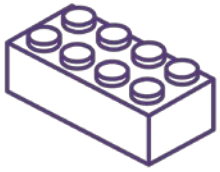
Conditions



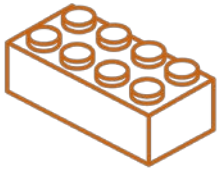
Drugs



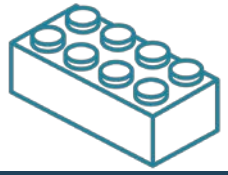
Procedures



Measurements



Observations



Visits

For Example

If a patient has had a diagnosis of diabetes

They're in!

For Example

If a patient has taken metformin in the past 12 months

They're in!

For Example

If a patient had HbA1c > 7.0

They're in!

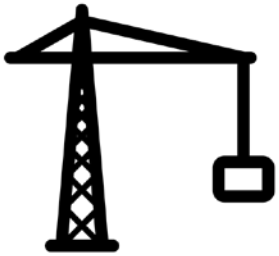
A good way to think about it…

- A phenotype is a way to represent a person with a condition or exposure using data in an electronic health record
- Thus phenotypes are an important foundation of describing the methods of an observational research study

How are people currently describing phenotypes in research publications?

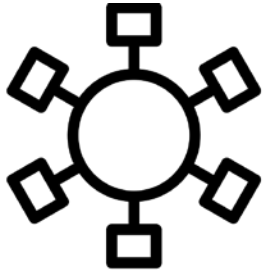
An OHDSI Approach to Phenotyping

Thoughtful
Design



Standardized
Implementation

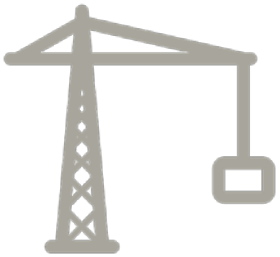
Reproducible
Evaluation



Portable
Dissemination

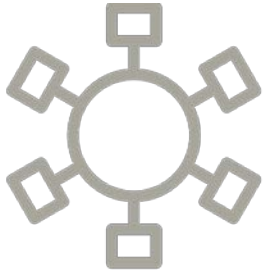
An OHDSI Approach to Phenotyping

Thoughtful
Design



Standardized
Implementation

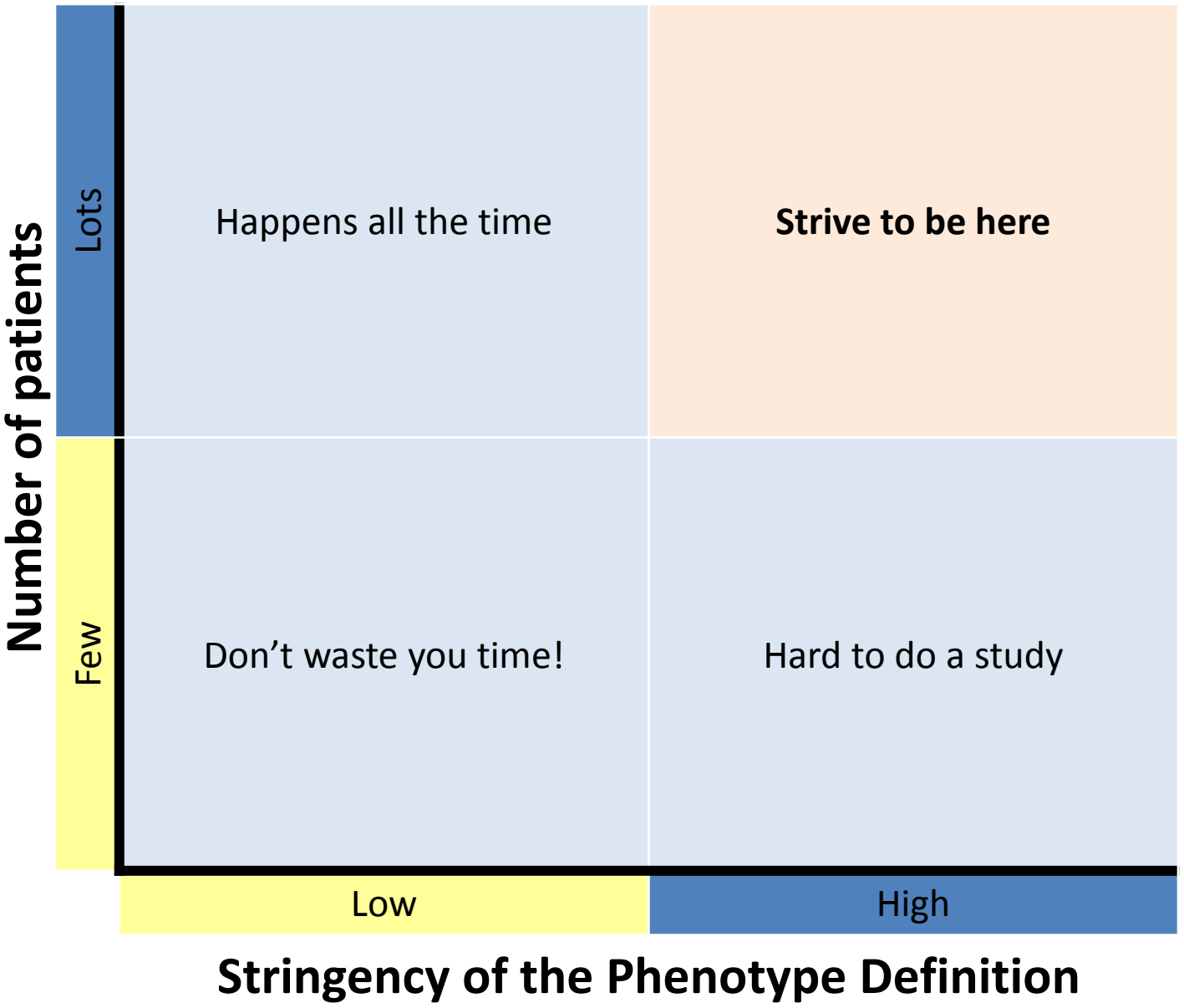
Reproducible
Evaluation



Portable
Dissemination

Basics of Phenotype Design

- What are the building blocks (data domains) you want to use to find your cases?
- Which of these is more important to you:
  - Finding all the eligible patients?
  - Getting only the ones you are confident about?


Number of patients
Lots
Few
Happens all the time
Strive to be here
Don’t waste you time!
Hard to do a study
Low
High
Stringency of the Phenotype Definition

What data types should go into a definition?

- There's no right answer. But here are some valid ones
  - Use everything you can get
  - Use the lowest common denominator so you can share
  - Use something in between

Two Approaches to Phenotyping

Rule-Based Phenotyping

Probabilistic Phenotyping

Steps in Rule-Based Phenotyping

- Primary Events (Start Date)
- Qualifying Criteria
- Exit Criteria (End Date)

Primary Events

- Cohort definitions can have lots of rules
- But the primary event is the bouncer
  - Have to clear this bar for the rest of the rules to come into play
- Besides being the first rule, the primary event is critical because it sets the *index date*

Index Date

- The patient's index date (aka cohort start date) is determined by when they satisfy the primary event
- The cohort start date can be limited to just first time a patient meets it or you can count every time they meet it
- Subsequent criteria are very commonly tied relative to the index date

Qualifying Criteria

- All the other criteria you wish you require of your cohort members
  - Noting that it is still the primary event that will mark their point of entry in the cohort
  - Can have AND or OR logic
  - Can apply the same filters as primary event
  - Temporal limitations relative to index

Exit Criteria

- Defines the end date of the individual in the cohort

Design Principles

- Phenotype design should take into consideration your goals and the nature of the study

New User of a Drug

A drug exposure of metformin

With 0 exposures of metformin prior

Using the earliest event per person

Diagnosis with Confirmation

A condition occurrence of **hypertension**

With 2 condition occurrences of **hypertension** within 1 year after index

Condition validated by Procedure

A condition occurrence of **cataract**

With procedure for **cataract removal**
within 2 weeks before and after

More Stringent Definitions

A condition occurrence of **diabetes**

With drug exposure of **oral DM meds**
within 90 days after index

Within measurement **HbA1c** > 7.0
within 90 days before and after index

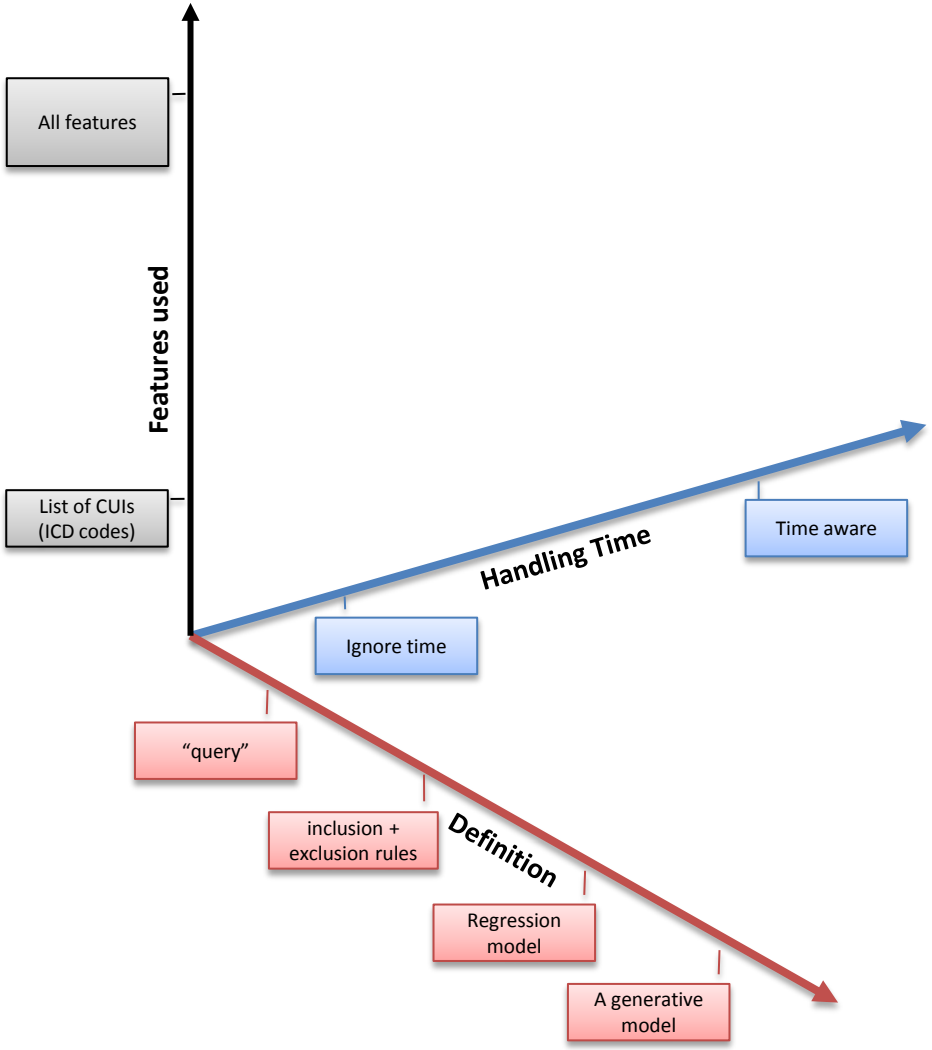

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Break

Probabilistic Phenotyping

Electronic phenotyping

- Identifying a set of patients:
  - For observational research
  - For clinical trial eligibility,
  - As Numerators or denominators of quality metrics
  - For whom a decision support reminder should “fire”
  - Who are “similar” based on whom a clinical decision should be based.
  - Who progress along similar paths
- The main problems:
  - the need for a gold standard
  - poor portability across sites and studies



Two approaches to phenotyping

- Rule based, expert-consensus definitions
  - Exemplified by www.phekb.org
  - Implemented by ATLAS www.ohdsi.org/web/atlas/
- Probabilistic phenotyping
  - Relatively new
  - APHRODITE, ANCHOR learning
  - https://github.com/OHDSI/Aphrodite

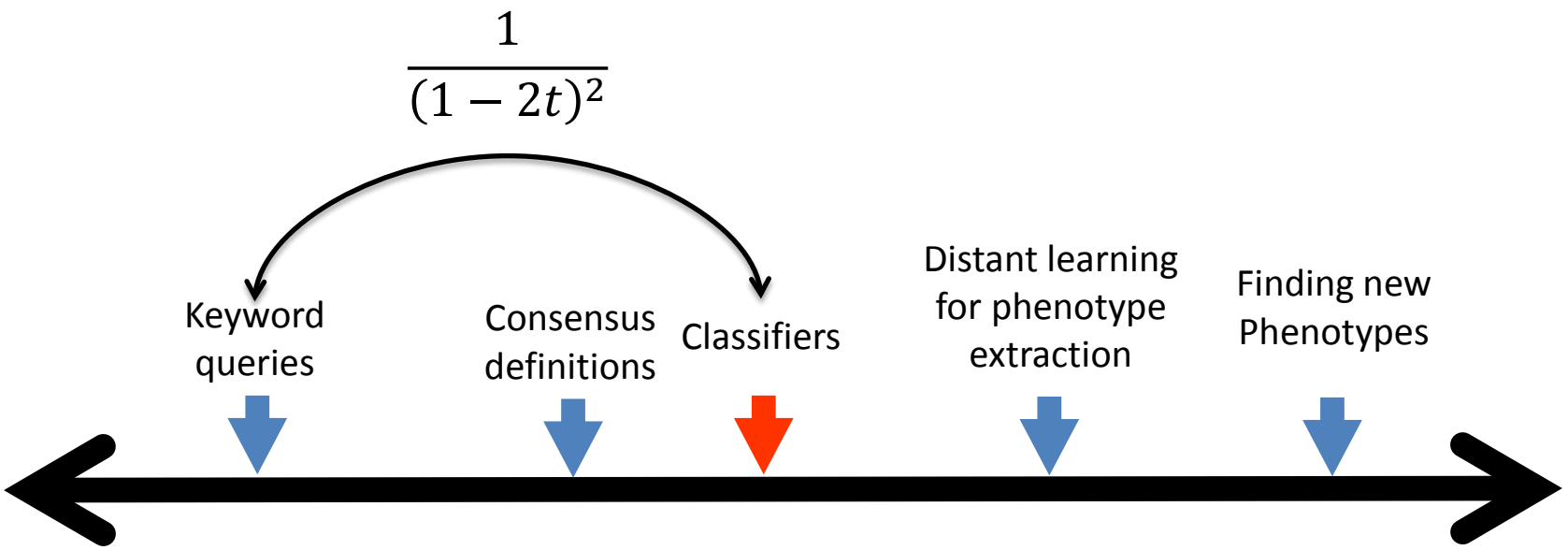
Probabilistic phenotyping

- The core idea is to learn from a set of labeled examples (i.e. supervised learning)

- Broad themes
  - Automated feature selection
  - Reduce the number of training samples
  - Probability of phenotype as a continuous trait

- APHRODITE aims to create large training datasets for “cheap” and still learn a good phenotype model.

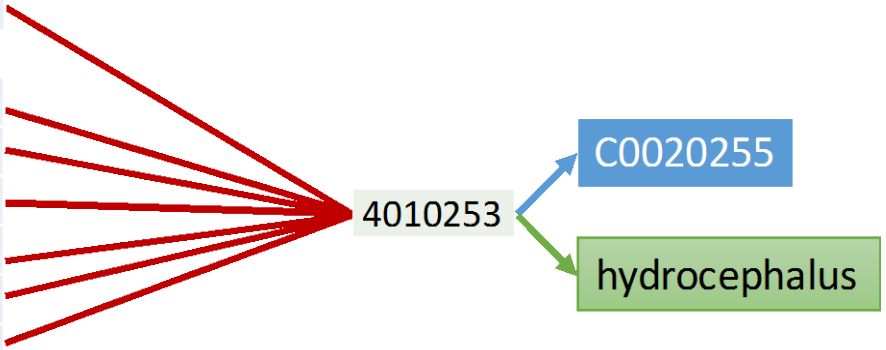
Learning using imperfect labels



Error rate in labeling	Sample size
10 %	1.56 x
20 %	2.77 x
30 %	6.25 x
40 %	25 x

“noisy labeling” to create training data

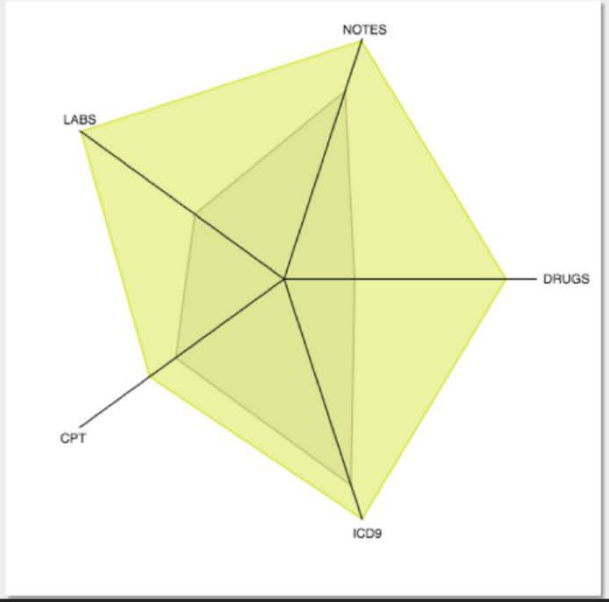
2933	hydrocephalus
90773	water on the brain
3107785	hydrocephalus [disease/finding]
42612	hydrocephaly
2889104	hydrocephalus (disorder)
1936514	hydrocephalus, unspecified
2001177	hydrocephalus nos



tid	cui	str	Note freq	syn	Medline freq	% noun
2933	C0020255	hydrocephalus	29,634	NNS	19,541	64.61
42612	C0020255	hydrocephaly	113	NN	275	49.81
90773	C0020255	water on the brain	8	ROOT	1	50

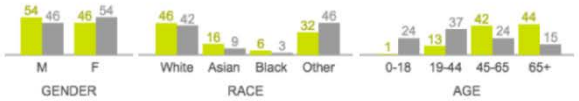
Assumption: “long mention” is a reliable indicator of presence

23,668 patients 37.118 s

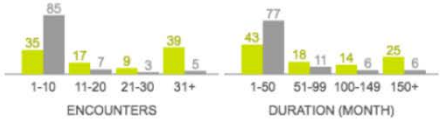


DOWNLOAD DATASET

Cohort is **male**, **white** and **65+ years old**



Mostly **1-10 encounters**, lasting **1-50 months**



QUERY

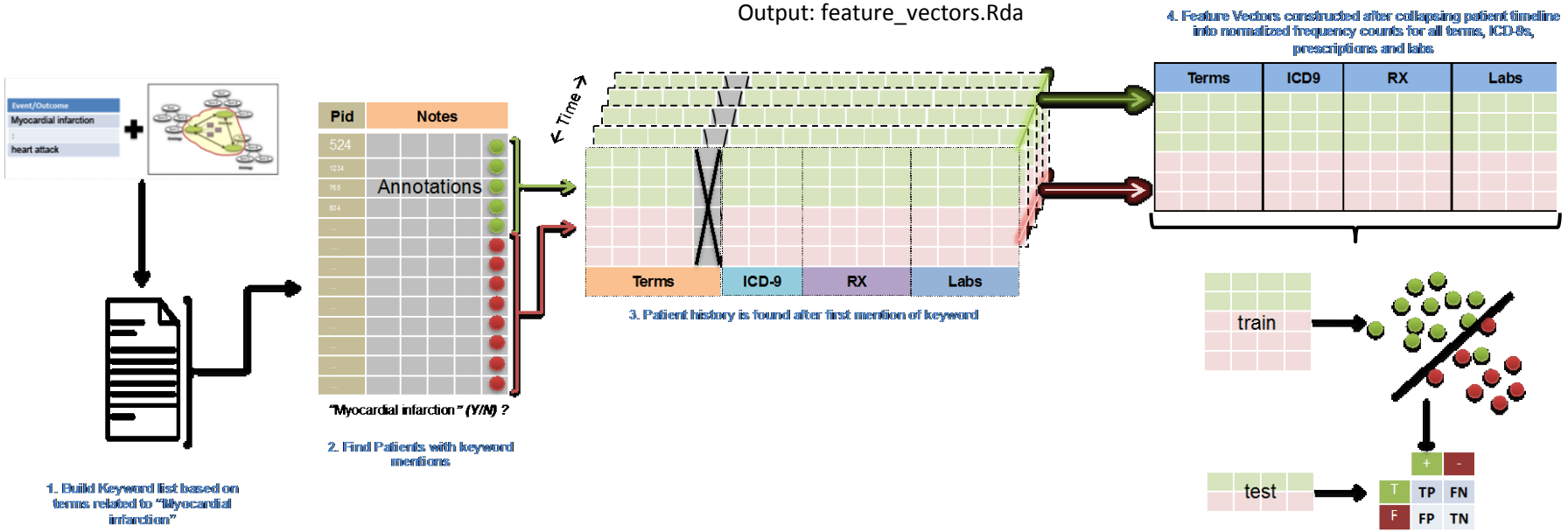
```
//CASE
var age_range = AGE(18 years, MAX)
var dx = UNION(ICD9=250.30, ICD9=250.20, ICD9=250.90, ICD9=250.80,
ICD9=250.70, ICD9=250.60, ICD9=250.50, ICD9=250.40, ICD9=250.00,
ICD9=250.32, ICD9=250.22, ICD9=250.92, ICD9=250.82, ICD9=250.72,
ICD9=250.62, ICD9=250.52, ICD9=250.42, ICD9=250.02)
var rx_noninsulin = UNION(RX=2404, RX=4821, RX=4815, RX=25789,
RX=73044, RX=274332, RX=6809, RX=84108, RX=33738, RX=16681,
RX=30009, RX=593411, RX=60548, RX=10633, RX=10635)
var gluc = UNION(LABS("GLU", "HIGH"), LABS("UGLU", "HIGH"),
LABS("GLUF", "HIGH"), LABS("GLUCSF", "HIGH"), LABS("GLT2",
"HIGH"), LABS("GTT1", "HIGH"), LABS("GLT1", "HIGH"))
var a1c = LABS("A1C", "HIGH")
var rx_insulin = UNION(RX=253182, RX=139953, RX=253181, RX=352385,
RX=314684, RX=86009, RX=51428, RX=139825)
var case1 = INTERSECT(HISTORY OF($dx), HISTORY OF($rx_noninsulin))
var case2 = INTERSECT(HISTORY OF($dx), HISTORY OF($gluc), HISTORY
OF($a1c), UNION(NO HISTORY OF($rx_noninsulin), HISTORY
OF($rx_insulin)))
var case3 = INTERSECT(HISTORY OF($gluc), HISTORY OF($a1c), HISTORY
OF($rx_insulin))
var hackathon.t2dm_vandy = UNION($case1, $case2, $case3)

$hackathon.t2dm_vandy
```

VARS: MINE GROUP

- a1c
- age_range
- case1
- case2
- case3
- dx
- gluc
- hackathon.t2dm_vandy
- rx_insulin
- rx_noninsulin

XPRESS- EXtraction of Phenotypes from clinical Records using Silver Standards

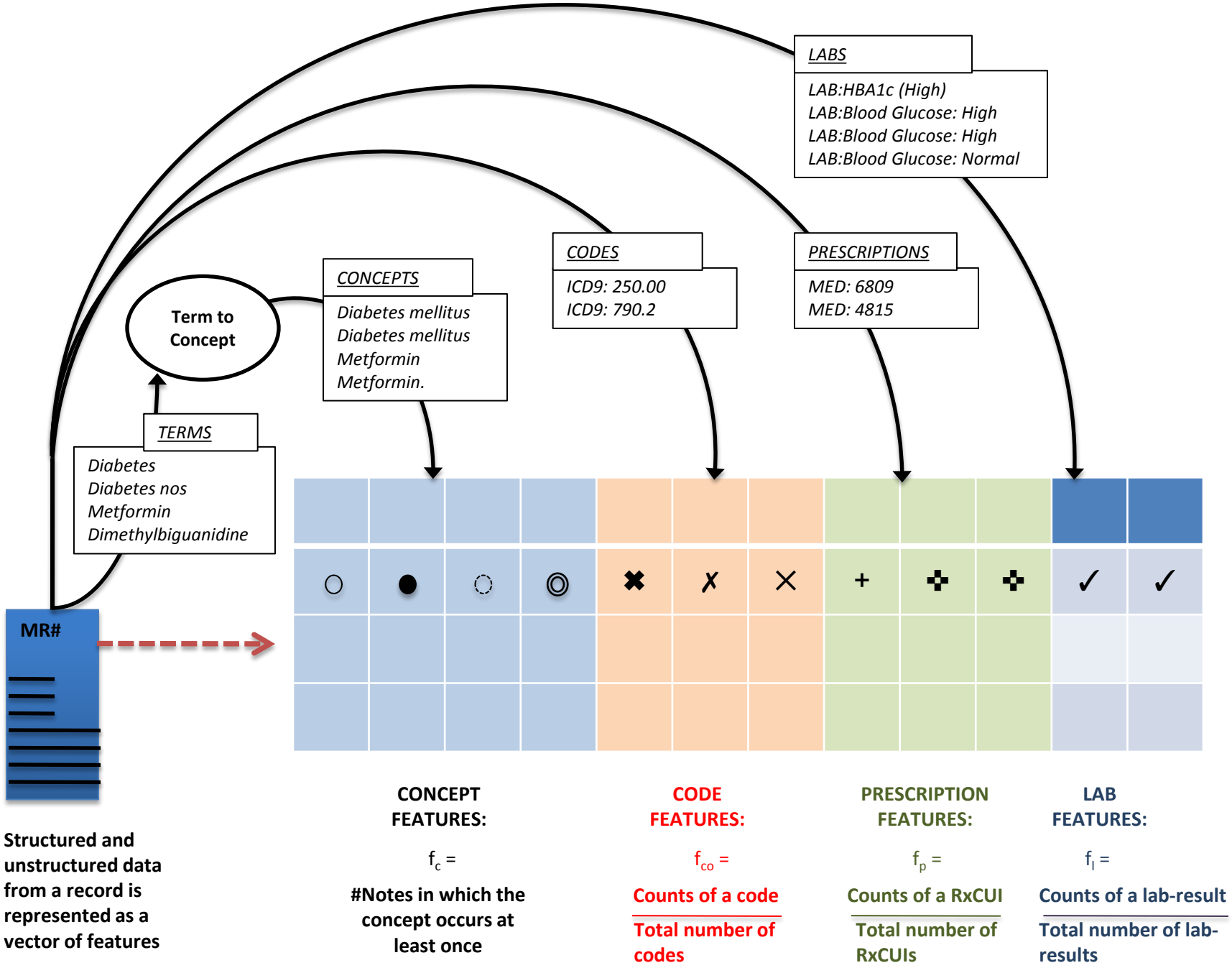
Input: getPatients.R -- config.R, keywords.tsv, ignore.tsv
Output: feature_vectors.Rda



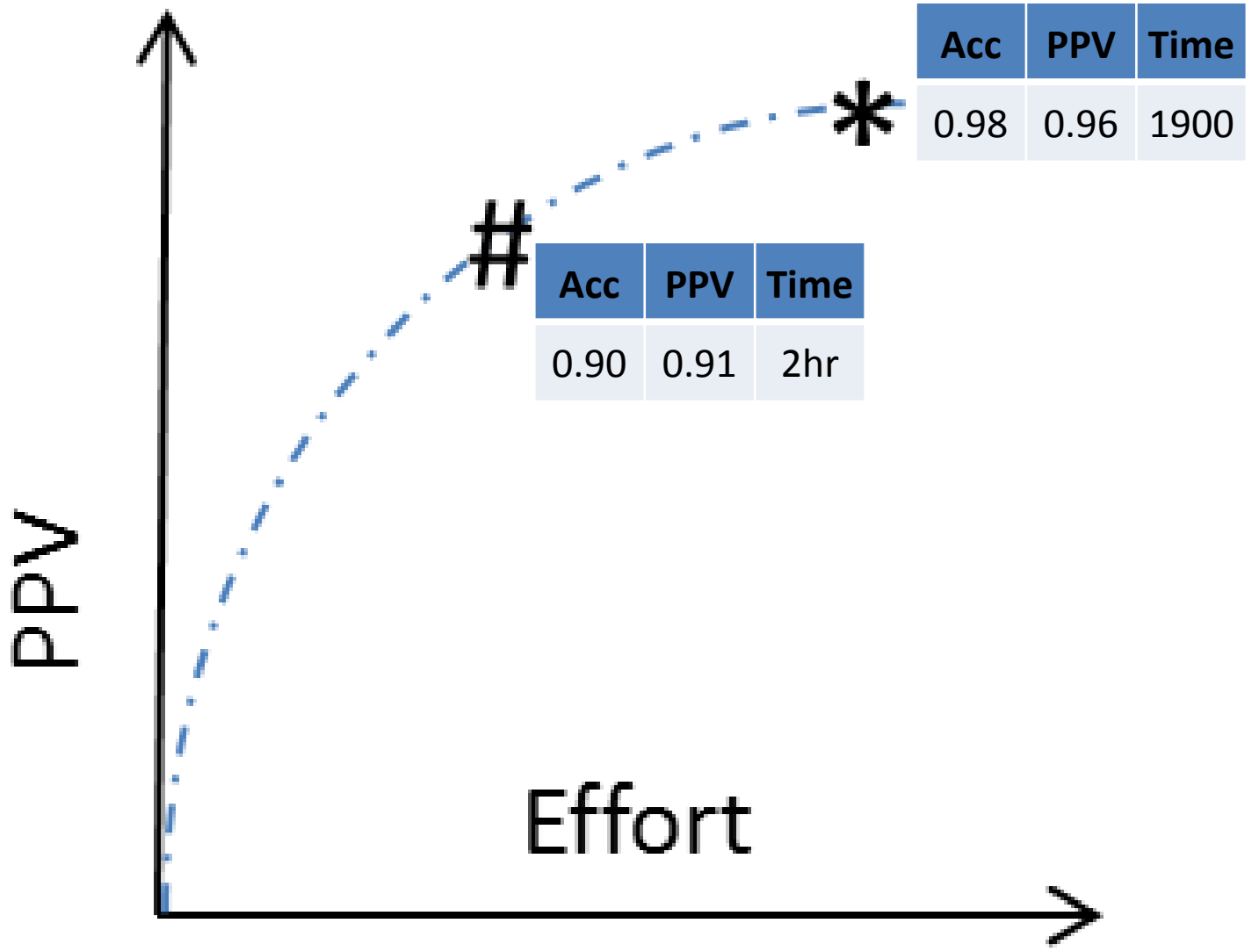
Input: config.R – with term search settings
Output: keywords.tsv and ignore.tsv

Phenotype	AUC	Sens.	Spec.	PPV
DM	0.95	91 %	83 %	83 %
MI	0.91	89 %	91 %	91 %
FH	0.90	76.5%	93.6%	~20%
Celiac	0.75	40 %	90 %	~4 %

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


LABS
LAB:HBA1c (High)
LAB:Blood Glucose: High
LAB:Blood Glucose: High
LAB:Blood Glucose: Normal
CODES
ICD9: 250.00
ICD9: 790.2
PRESCRIPTIONS
MED: 6809
MED: 4815
CONCEPTS
Diabetes mellitus
Diabetes mellitus
Metformin
Metformin.
Term to Concept
TERMS
Diabetes
Diabetes nos
Metformin
Dimethylbiguanidine
MR#
Structured and unstructured data from a record is represented as a vector of features
CONCEPT FEATURES:
f_c =
#Notes in which the concept occurs at least once
CODE FEATURES:
f_{co} =
Counts of a code / Total number of codes
PRESCRIPTION FEATURES:
f_p =
Counts of a RxCUI / Total number of RxCUIs
LAB FEATURES:
f_l =
Counts of a lab-result / Total number of lab-results

Effort precision trade off



http://github.com/OHDSI/Aphrodite

- Build phenotype models in 5 easy steps!

- Designed and Implemented using OHDSI CDMv5 and Vocabulary 5

```
          Reference
Prediction  F  T
         F 86 15
         T  1 72

               Accuracy : 0.908
                 95% CI : (0.855, 0.9465)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8161
 Mcnemar's Test P-Value : 0.001154

            Sensitivity : 0.8276
            Specificity : 0.9885
         Pos Pred Value : 0.9863
         Neg Pred Value : 0.8515
             Prevalence : 0.5000
         Detection Rate : 0.4138
   Detection Prevalence : 0.4195
      Balanced Accuracy : 0.9080

       'Positive' Class : T


Model Details

glmnet

 526 samples
1932 predictors
   2 classes: 'F', 'T'
```

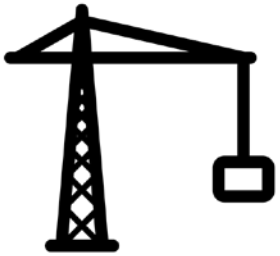
Tutorial Video: http://tinyurl.com/use-aphrodite

Unsolved questions

- Do we share learned models, or do we share the modeling building workflow?
- How do we share the model or the workflow?
- CDM v5 extensions to make it all work
  - ✓Term mentions from clinical notes
  - Time in all tables
  - Consistent ICD/CPT mappings to SNOMED

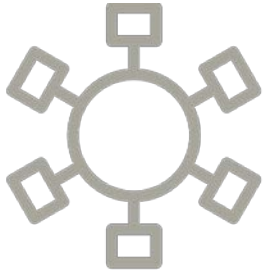
An OHDSI Approach to Phenotyping

Thoughtful
Design



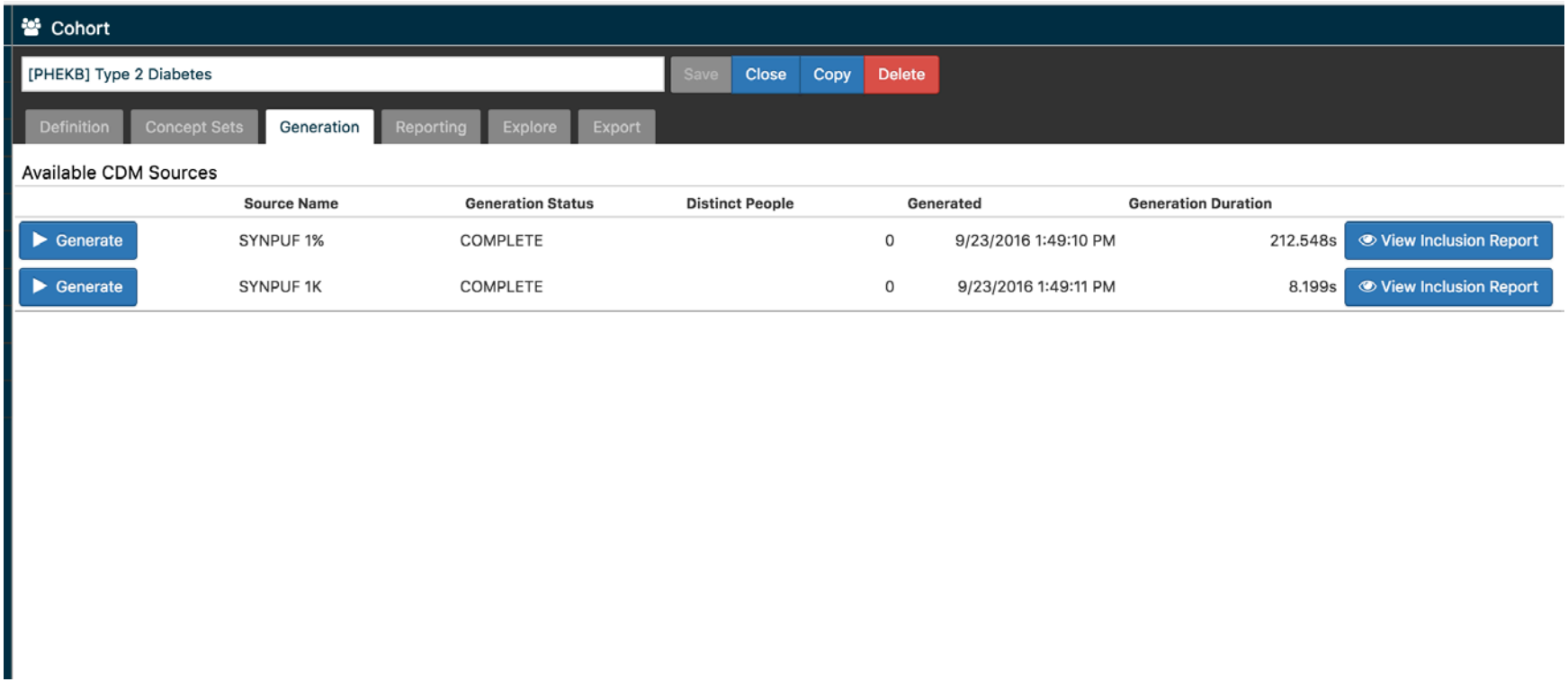
Standardized
Implementation

Reproducible
Evaluation



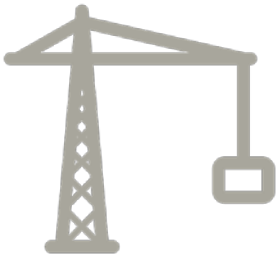
Portable
Dissemination

Implement via Atlas



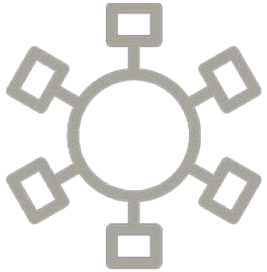
An OHDSI Approach to Phenotyping

Thoughtful
Design



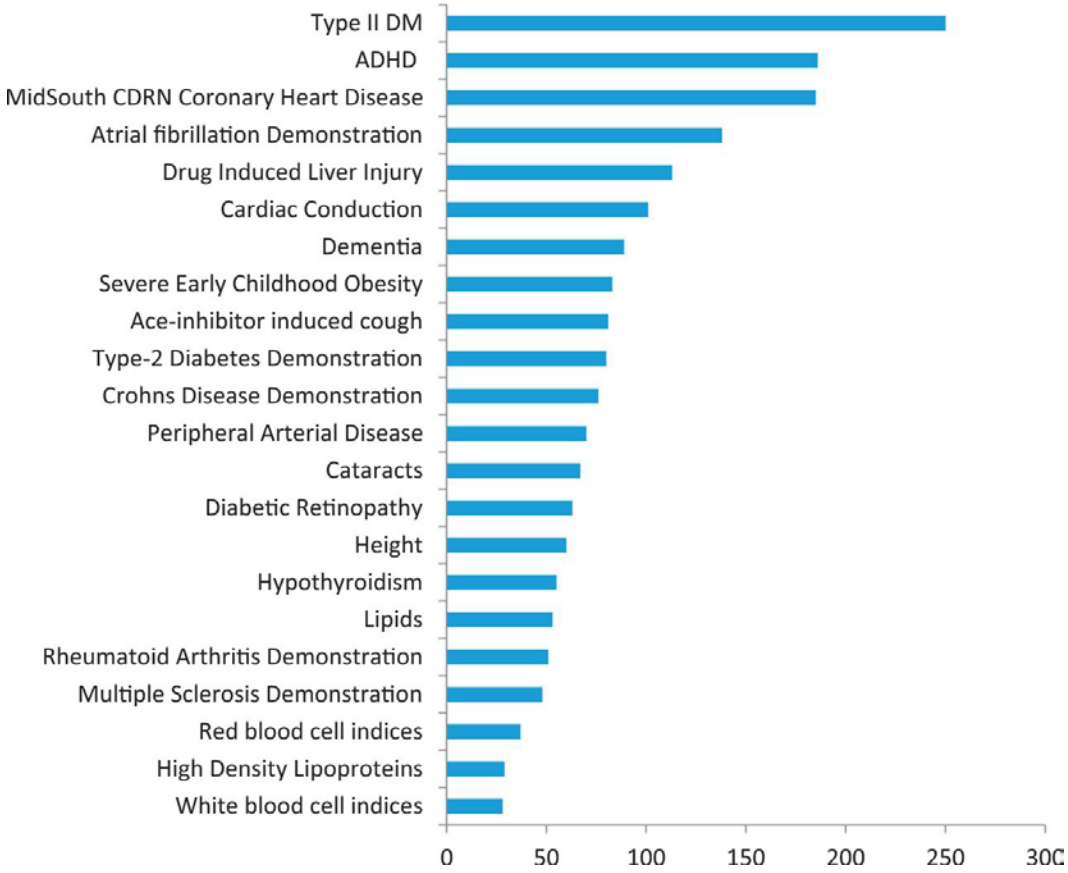
Standardized
Implementation

Reproducible
Evaluation

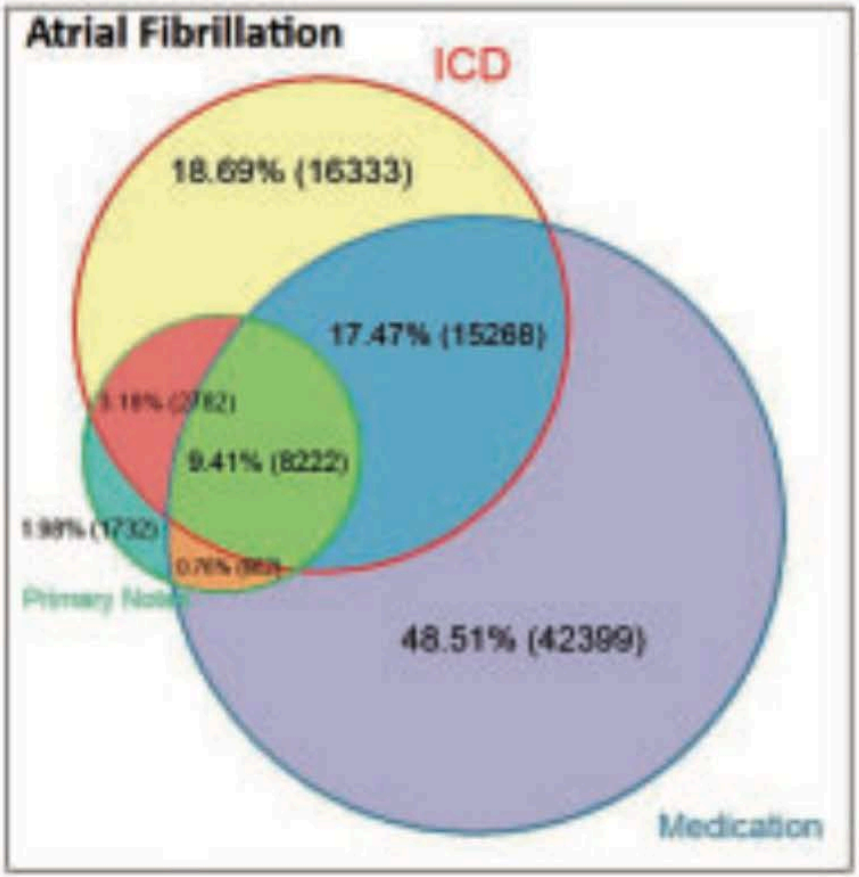


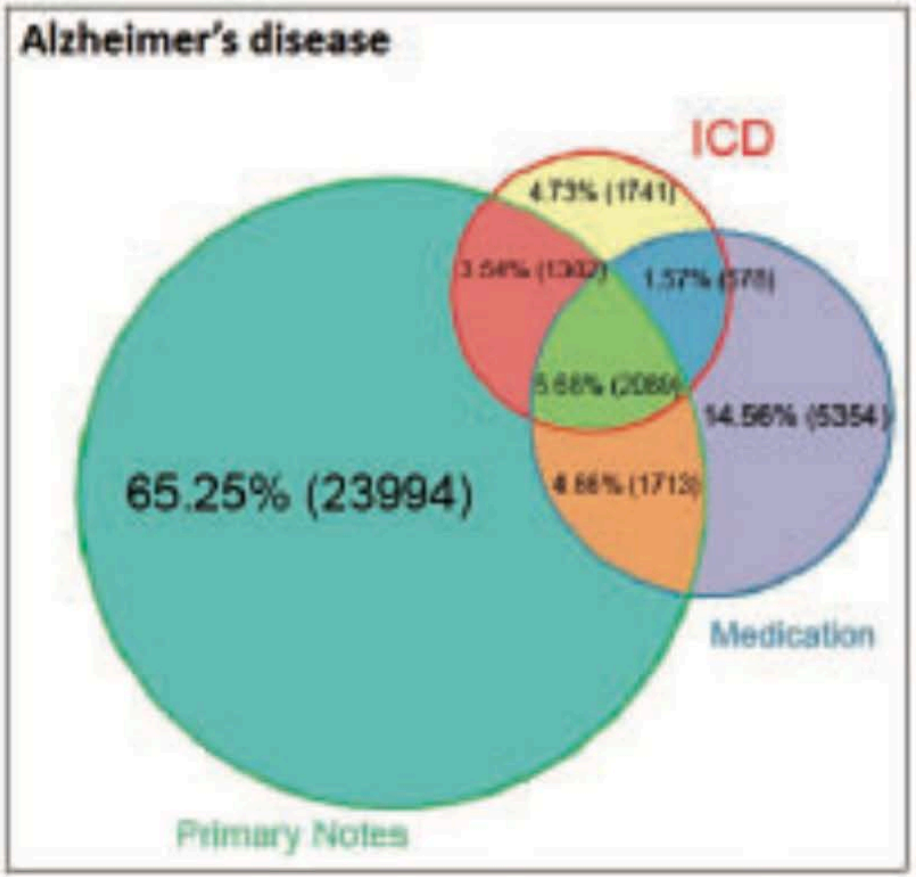
Portable
Dissemination

PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability



Jacqueline C Kirby et al. J Am Med Inform Assoc 2016;jamia.ocv202



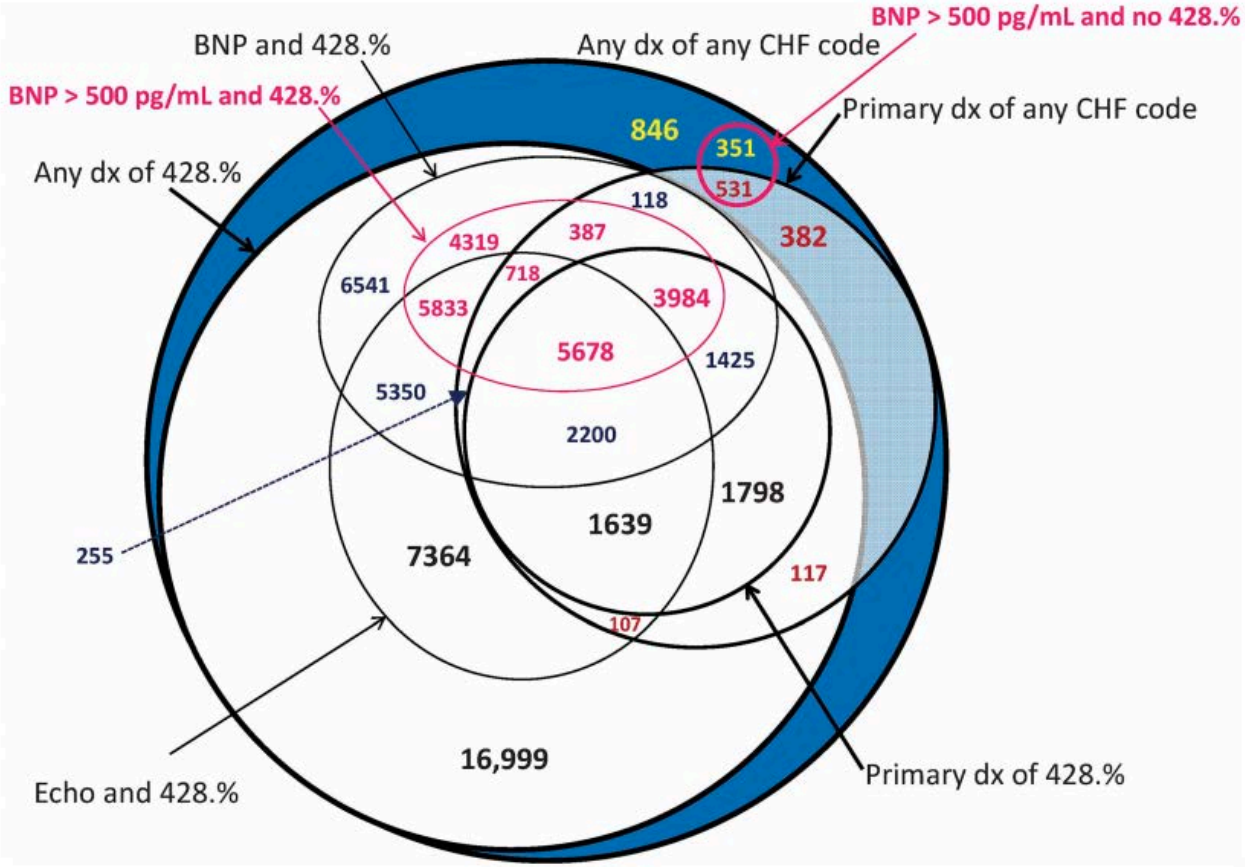


Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. Journal of the American Medical Informatics Association. 2016 Apr 1;23(e1):e20-7.

PheKB T2DM Evaluation

Implementation Details	Case PPV	Control PPV	Dataset/Dictionary	Dataset Validation
T2D Marshfield Implementation *Marshfield Clinic Research Foundation* *Cases: 0 Controls: 0 (Case, Control)* *Uploaded: 03/20/2012*	0.99	0.98	No datasets uploaded	
T2D Northwestern Implementation *Northwestern University* *Cases: 0 Controls: 0 (Case, Control)* *Uploaded: 03/20/2012*	0.982456	1	No datasets uploaded	
T2D Vanderbilt Implementation *Vanderbilt University* *Cases: 0 Controls: 0 (Case, Control)* *Uploaded: 03/20/2012*	1	1	No datasets uploaded	
T2DM Implementation - Columbia *Columbia University* *Cases: 293 Controls: 478 (Case, Control)* *Uploaded: 05/03/2016*			No datasets uploaded	

Highly Granular Phenotype Evaluation



Rosenman et al. Database queries for hospitalizations for acute congestive heart failure: flexible methods and validation based on set theory. *J Am Med Inform Assoc*. 2014 Mar-Apr;21(2):345-52.

Table 3 Results for the 10 congestive heart failure (CHF) phenotype queries

Criteria to combine Venn diagram zones	N in query	Sensitivity (%)	Sensitivity, SE (%)	PPV (%)	PPV, SE (%)
Any CHF	66 942	94.3	1.3	42.8	1.5
Any dx of 428	64 832	90.9	1.3	42.5	1.5
Any dx of CHF and BNP >500 pg/mL	21 801	50.8	1.8	70.7	2.5
1^0 dx of any CHF	19 339	54.8	1.9	86.0	2.2
1^0 dx of 428	16 724	47.6	1.7	86.3	2.5
1^0 dx of any CHF and BNP >500 pg/mL	11 298	33.5	1.3	90.0	2.1
1^0 dx of 428 and BNP >500 pg/mL	9662	28.8	1.1	90.4	2.4
1^0 dx of 428 and BNP >500 pg/mL and echocardiogram	5678	16.2	0.8	86.6	3.5
1^0 dx of any CHF or BNP >500 pg/mL	29 587	71.4	2.1	73.3	2.2
1^0 dx of 428 or BNP >500 pg/mL	28 863	69.6	2.1	73.2	2.2
High BNP, no ICD-9 diagnosis for CHF					
Zone X: no ICD-9 dx of 428, but BNP >500 pg/mL	12 149	N/A	N/A	14.3	3.5

BNP, B-natriuretic peptide; PPV, positive predictive value.

Did you find these metrics in the papers you read?

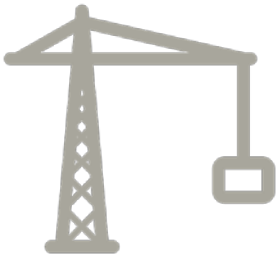
What information did the authors provide to give you confidence in the reliability of their definitions?

Phenotype Evaluation @ OHDSI

- A major initiative for the coming year
- Help wanted building our evaluation framework!

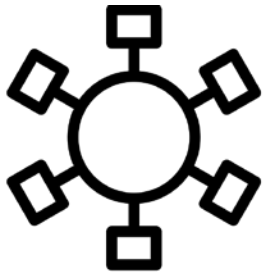
An OHDSI Approach to Phenotyping

Thoughtful
Design



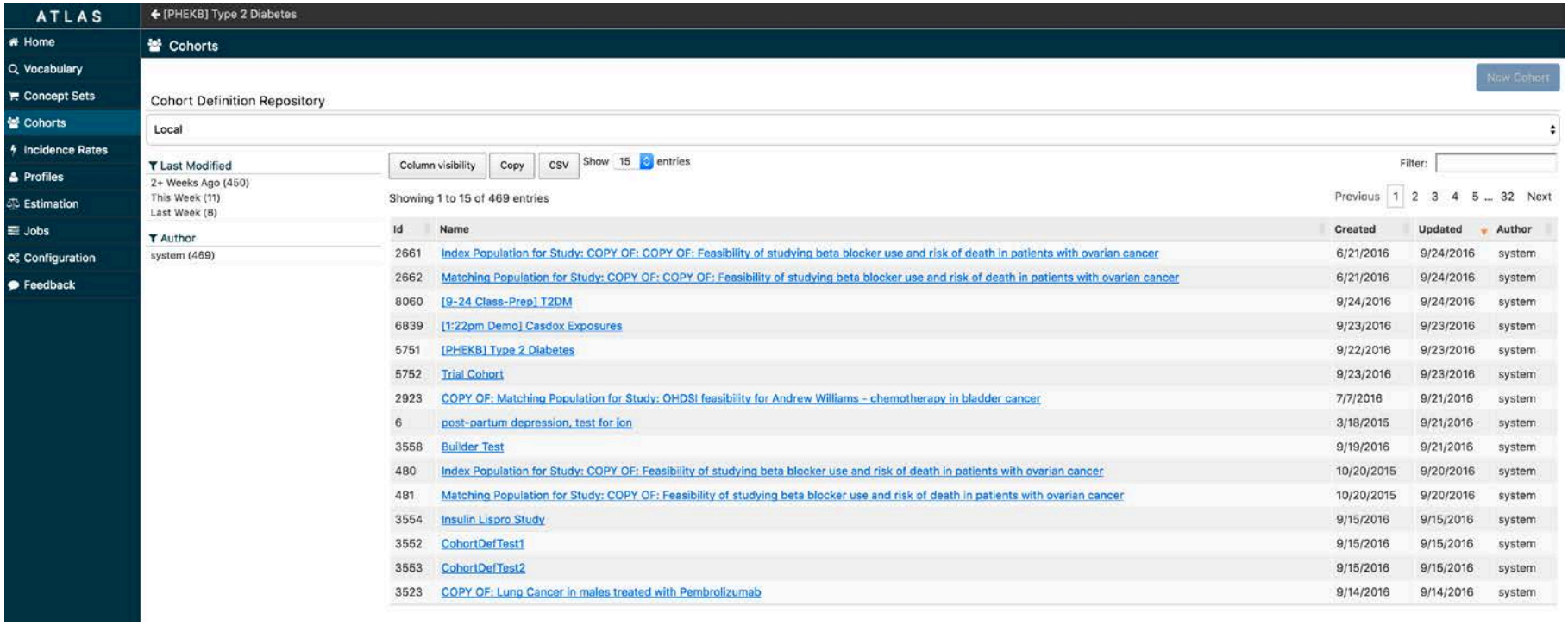
Standardized
Implementation

Reproducible
Evaluation



Portable
Dissemination

Share via OHDSI.org



Hands-On Exercises

Pair up in groups of 3, working together then we come in and help with the groups