OHDSI Cohort Definition and Phenotyping

Chris Knoll, Juan Banda
Gowtham Rao, Ray Chen,
Joel Swerdel

Disclosures

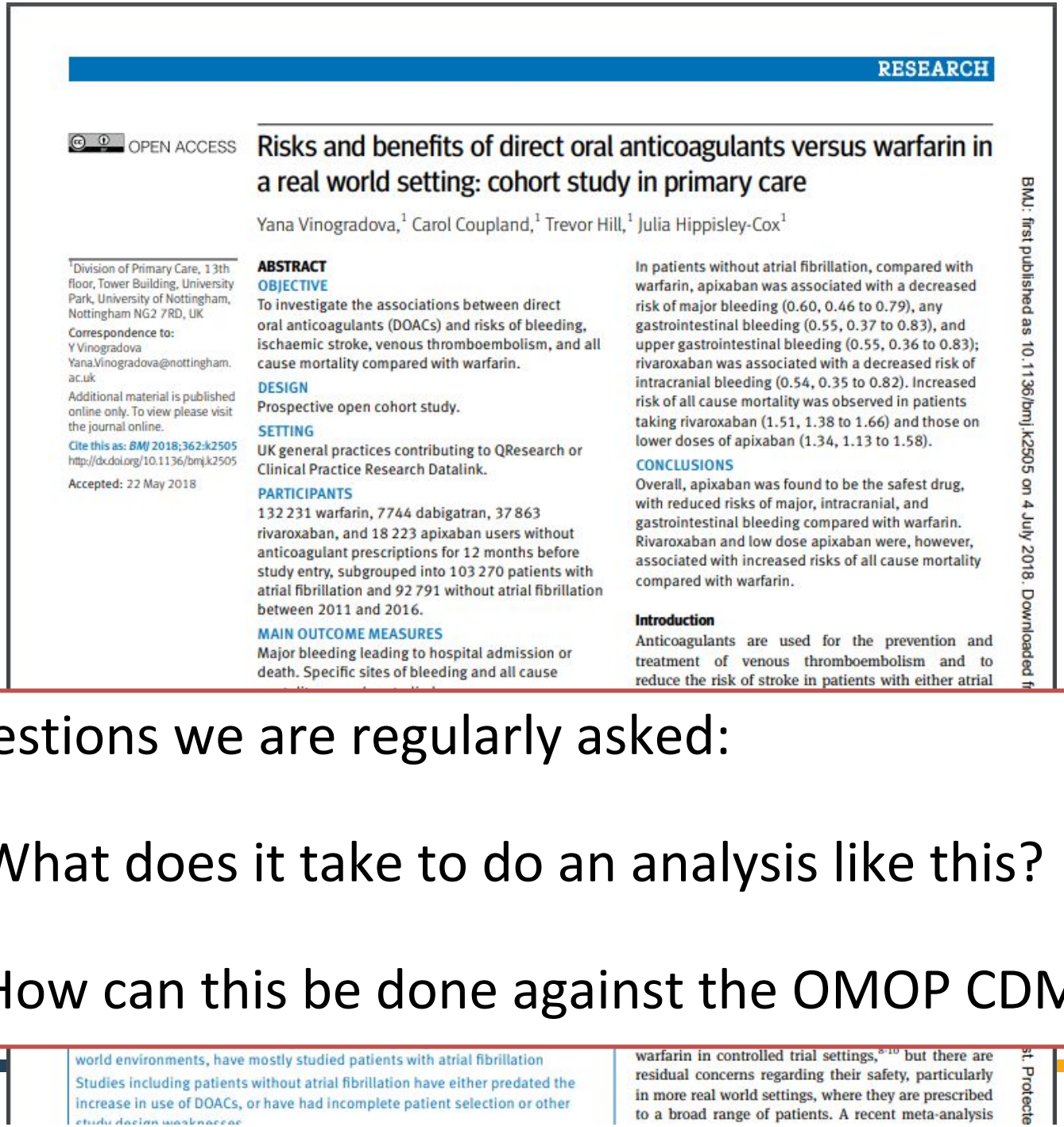
Christopher Knoll and Joel Swerdel are employees of Janssen Research and Development, and shareholders of Johnson & Johnson.

Agenda

Time	Section	
8:45 AM - 9:00 AM		Introductions
9:00 AM - 10:00 AM	Overview	Phenotyping using the Common Data Model. Christopher Knoll
10:00 AM - 10:30 AM	Student Activity	Student Ex. 1 and Break Gowtham Rao
10:30 AM - 12:00 PM	Rule-Based Phenotyping in Atlas	Atlas Walkthrough: Vocabulary Searching, Concept Set Creation, Cohort Definitions Christopher Knoll, Gowtham Rao
12:00 PM - 1:00 PM	Lunch Student Activity	Students will consider their own cohort definition designs by specifying cohort entry events, inclusion criteria, cohort exit, and cohort eras.
1:00 PM - 2:00 PM	Probabilistic Phenotyping in Aphrodite	Using Aphrodite to learn models for phenotyping. Juan Banda
2:00 PM - 2:30 PM	Phenotype Evaluation	Evaluating the performance of phenotype algorithms. Joel Swerdel
2:30 PM - 3:30 PM	Student Workgroup	Student led cohort definition design
3:30 PM - 4:00 PM	Advanced Atlas	Advanced techniques in atlas Christopher Knoll
4:00 PM - 4:45 PM	Applications of Cohorts	Cohorts used in prediction, population Level estimation, characterization
4:45 PM	Final Review	

Phenotyping using the Common Data Model

Christopher Knoll



RESEARCH

OPEN ACCESS

Risks and benefits of direct oral anticoagulants versus warfarin in a real world setting: cohort study in primary care

Yana Vinogradova,[1] Carol Coupland,[1] Trevor Hill,[1] Julia Hippisley-Cox[1]

[1]Division of Primary Care, 13th floor, Tower Building, University Park, University of Nottingham, Nottingham NG2 7RD, UK

Correspondence to:
Y Vinogradova
Yana.Vinogradova@nottingham.ac.uk

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2018;362:k2505
http://dx.doi.org/10.1136/bmj.k2505

Accepted: 22 May 2018

ABSTRACT

OBJECTIVE
To investigate the associations between direct oral anticoagulants (DOACs) and risks of bleeding, ischaemic stroke, venous thromboembolism, and all cause mortality compared with warfarin.

DESIGN
Prospective open cohort study.

SETTING
UK general practices contributing to QResearch or Clinical Practice Research Datalink.

PARTICIPANTS
132 231 warfarin, 7744 dabigatran, 37 863 rivaroxaban, and 18 223 apixaban users without anticoagulant prescriptions for 12 months before study entry, subgrouped into 103 270 patients with atrial fibrillation and 92 791 without atrial fibrillation between 2011 and 2016.

MAIN OUTCOME MEASURES
Major bleeding leading to hospital admission or death. Specific sites of bleeding and all cause

In patients without atrial fibrillation, compared with warfarin, apixaban was associated with a decreased risk of major bleeding (0.60, 0.46 to 0.79), any gastrointestinal bleeding (0.55, 0.37 to 0.83), and upper gastrointestinal bleeding (0.55, 0.36 to 0.83); rivaroxaban was associated with a decreased risk of intracranial bleeding (0.54, 0.35 to 0.82). Increased risk of all cause mortality was observed in patients taking rivaroxaban (1.51, 1.38 to 1.66) and those on lower doses of apixaban (1.34, 1.13 to 1.58).

CONCLUSIONS
Overall, apixaban was found to be the safest drug, with reduced risks of major, intracranial, and gastrointestinal bleeding compared with warfarin. Rivaroxaban and low dose apixaban were, however, associated with increased risks of all cause mortality compared with warfarin.

Introduction

Anticoagulants are used for the prevention and treatment of venous thromboembolism and to reduce the risk of stroke in patients with either atrial

world environments, have mostly studied patients with atrial fibrillation
Studies including patients without atrial fibrillation have either predated the increase in use of DOACs, or have had incomplete patient selection or other

warfarin in controlled trial settings,[8-10] but there are residual concerns regarding their safety, particularly in more real world settings, where they are prescribed to a broad range of patients. A recent meta-analysis

st. Protecte

Questions we are regularly asked:

1) What does it take to do an analysis like this?

2) How can this be done against the OMOP CDM?

Artifacts of Observational Study

eTable 5 ***All patients****: Rates per 1000py of outcomes by drug and adjusted hazard ratios with reference to warfarin in QResearch and CPRD*

eTable 2 Supplementary data for Figure2: Number of patients starting anticoagulants by study year

	QResearch	CPRD

CPRD	
x adjusted rate)00py (95%CI)	Adjusted hazard ratios (95%CI)[a]

eTable 3 ***Patients with atrial fibrillation****: Characteristics of patients in the QResearch and CPRD cohorts, percentage (numbers).*

eTable 10 ***Patients with and without atrial fibrillation****: Comorbidities, previous events and other medications, percentage (number) of patients on lower and higher doses for DOAC's.*

	With atrial fibrillation				Without atrial fibrillation			
	QResearch		CPRD		QResearch		CPRD	
Total N	Lower dose	Higher dose	Lower dose	Higher dose	Lower dose	Higher dose	Lower dose	Higher dose
Comorbidities at baseline								
Alcohol dependence	2.1 (171)	3.4 (646)	1.9 (32)	3.4 (128)	2.2 (146)	3.6 (738)	2.5 (32)	3.6 (88)
Bleeding disorders	1.2 (101)	0.9 (175)	1.8 (30)	1.2 (43)	1.2 (82)	1.1 (226)	1.5 (20)	1.7 (41)
Cancer (any)	15.4 (1245)	12.1 (2333)	14.1 (231)	11.9 (442)	14.5 (968)	12.5 (2591)	14.1 (184)	13.3 (321)
Chronic liver disease or pancreatitis	1.5 (122)	1.3 (247)	1.3 (22)	1.5 (54)	1.2 (82)	1.3 (264)	1.3 (17)	1.7 (42)
Chronic obstructive pulmonary disease	11.2 (906)	8.9 (1710)	10.1 (166)	8.9 (331)	8.5 (568)	8.5 (1754)	6.9 (90)	8.6 (209)
Chronic renal disease	4.2 (336)	0.7 (128)	4.1 (68)	0.9 (34)	3.1 (209)	1.0 (213)	3.3 (43)	1.1 (26)
Congestive cardiac failure	16.7 (1355)	9.7 (1873)	17.0 (280)	9.8 (364)	9.1 (608)	5.1 (1060)	6.5 (85)	5.5 (132)
Coronary heart disease	28.9 (2335)	20.3 (3901)	27.4 (451)	20.3 (754)	20.8 (1393)	14.7 (3053)	17.6 (230)	15.0 (363)
Diabetes	20.1 (1629)	17.5 (3362)	19.9 (327)	17.2 (637)	18.0 (1208)	16.0 (3315)	16.2 (212)	13.8 (333)
Dyspepsia	19.2 (1557)	18.4 (3542)	27.3 (449)	25.6 (949)	17.8 (1193)	17.3 (3588)	26.4 (345)	24.3 (588)
Falls or hip fracture (within 180 days)	12.3 (995)	5.7 (1098)	9.5 (157)	4.9 (183)	17.9 (1195)	5.5 (1132)	7.9 (103)	4.9 (118)
Hip or knee operation (within 180 days)	1.3 (102)	0.7 (128)	2.4 (39)	1.3 (49)	22.6 (1513)	2.4 (496)	21.9 (286)	3.6 (88)
Hypertension	66.2 (5354)	57.0 (10958)	69.6 (1144)	58.5 (2171)	52.2 (3497)	41.2 (8539)	49.5 (646)	42.0 (1014)
Ischaemic stroke*	26.1 (2115)	17.0 (3265)	29.1 (478)	19.8 (736)	19.0 (1272)	12.9 (2680)	15.6 (204)	14.4 (348)
Oesophageal varices	0.1 (9)	0.1 (19)	<5	<5	0.1 (4)	0.1 (24)	<5	<5
Peptic ulcer	9.4 (757)	6.9 (1322)	10.0 (164)	8.4 (310)	7.3 (486)	5.8 (1198)	7.8 (102)	6.5 (158)
Valvular heart disease	12.6 (1022)	8.2 (1572)	9.9 (162)	6.3 (233)	7.0 (471)	4.5 (934)	4.8 (63)	3.9 (94)
Venous thromboembolism*	5.1 (412)	4.9 (940)	8.3 (137)	5.9 (219)	13.2 (883)	37.4 (7747)	27.3 (356)	46.0 (1112)

The common building block of all observational analysis: cohorts

Required inputs:

Target cohort:
- Person
- cohort start date
- cohort end date

Comparator cohort:
- Person
- cohort start date
- cohort end date

Outcome cohort:
- Person
- cohort start date
- cohort end date

Desired outputs:

Clinical characterization
- Baseline summary of exposures (treatment utilization)

Clinical characterization
- Baseline summary of outcome (disease natural history)

Incidence summary
- Proportion/rate of outcome occurring during time-at-risk for exposure

Population-level effect estimation
- Relative risk (HR, OR, IRR) of outcome occurring during time-at-risk for exposure

Patient-level prediction
- Probability of outcome occurring during time-at-risk for each patient in population

OHDSI's definition of 'cohort'

A cohort is a set of persons who satisfy one or more inclusion criteria (a phenotype) for a duration of time

Objective consequences based on this cohort definition:

- One person may belong to multiple cohorts
- One person may belong to the same cohort at multiple different time periods
- One person may not belong to the same cohort multiple times during the same period of time
- One cohort may have zero or more members
- A codeset is NOT a cohort…

  …logic for how to use the codeset in a criteria is required

Defining ‘phenotype’

Journal of the American Medical Informatics Association, 0(0), 2017, 1–6
doi: 10.1093/jamia/ocx110
Perspective

Perspective

High-fidelity phenotyping: richness and freedom from bias

George Hripcsak[1] and David J Albers[1]

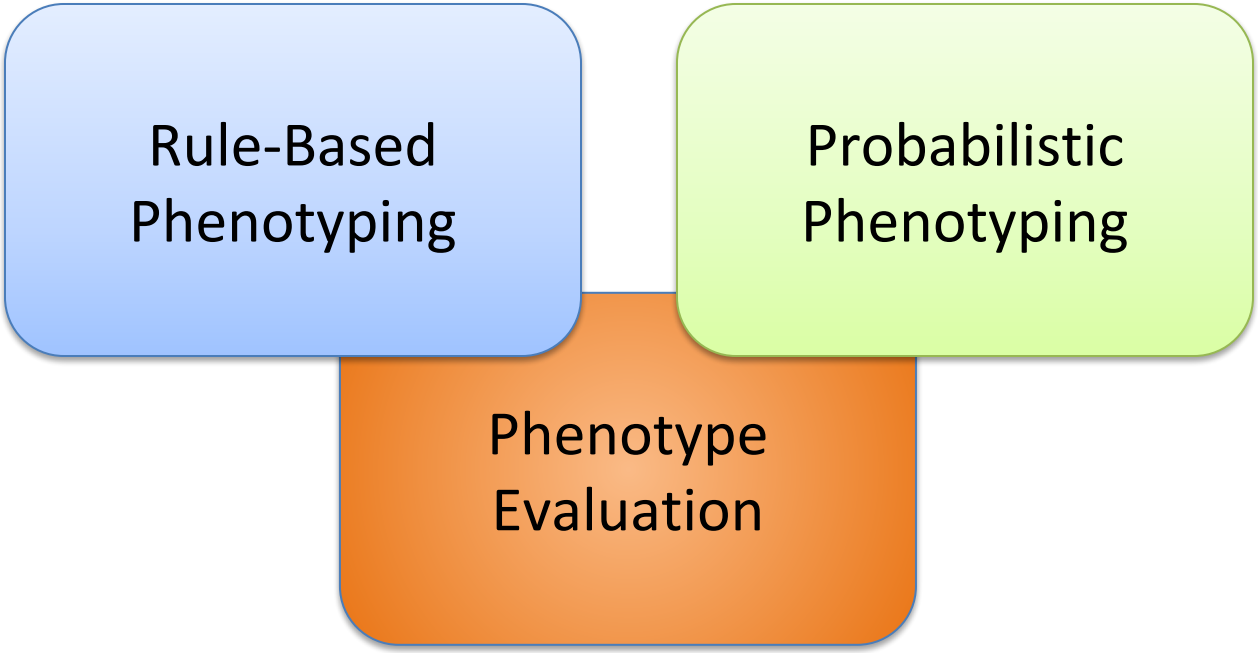
[1]Department of Biomedical Informatics, Columbia University Medical Center, New York, NY, USA

Corresponding Author: George Hripcsak, Department of Biomedical Informatics, Columbia University Medical Center, 622 W 168th St, PH20, New York, NY 10032, USA. E-mail: hripcsak@columbia.edu. Phone: (212) 305-5712

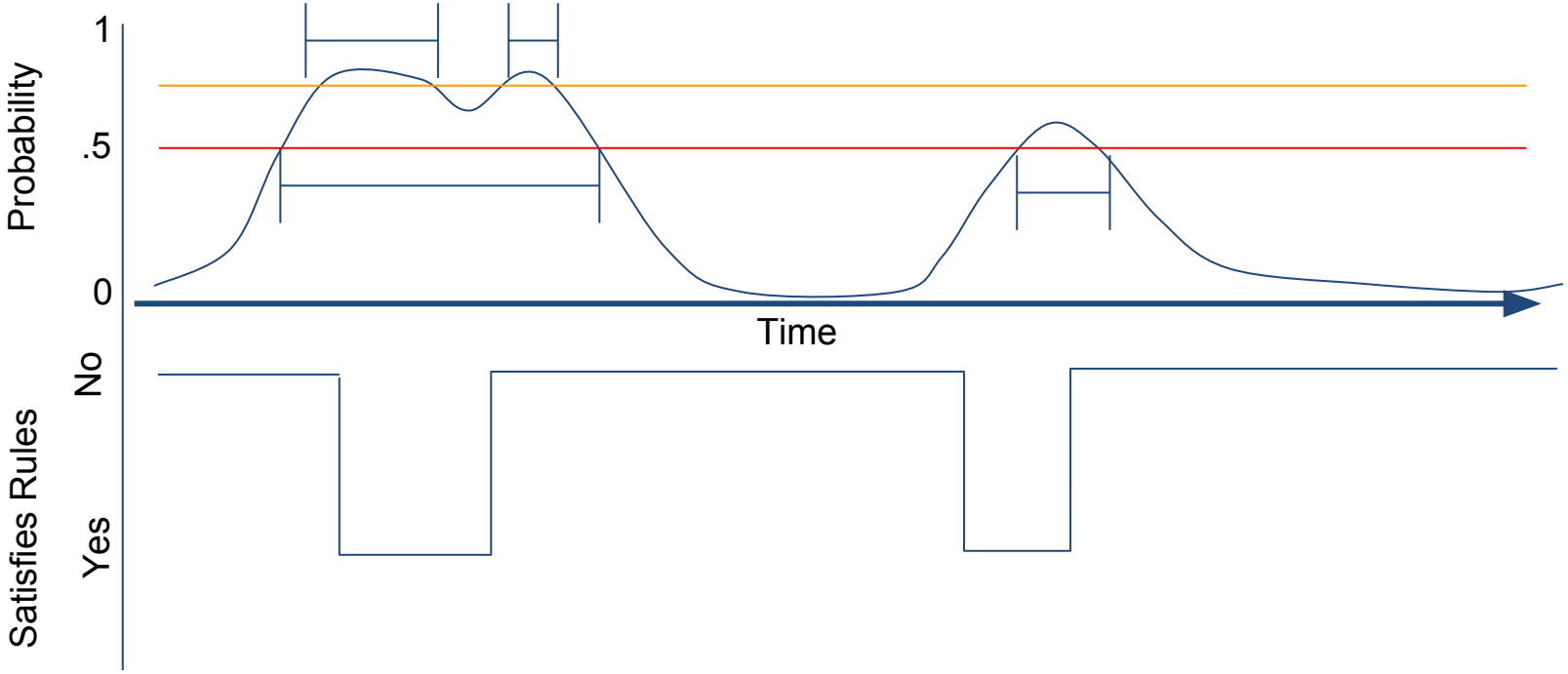
Received 16 June 2017; Revised 7 August 2017; Editorial Decision 6 September 2017; Accepted 13 September 2017

- A phenotype is a specification of an observable, potentially changing state of an organism (as distinguished from the genotype, derived from genetic makeup).
- The term phenotype can be applied to patient characteristics inferred from electronic health record (EHR) data.
- The goal is to draw conclusions about a target concept based on raw EHR data, claims data, or other clinically relevant data.
- Phenotype algorithms – ie, algorithms that identify or characterize phenotypes – may be generated by domain experts and knowledge engineers, or through diverse forms of machine learning to generate novel representations of data.

Two Approaches to Phenotyping



Probabilistic vs. Rule Based over time



Data are building blocks for Phenotypng

- Conditions
- Drugs
- Procedures
- Measurements
- Observations
- Visits

Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance

RECEIVED 8 January 2015
REVISED 14 July 2015
ACCEPTED 15 July 2015
PUBLISHED ONLINE FIRST 2 September 2015

Wei-Qi Wei[1], Pedro L Teixeira[1], Huan Mo[1], Robert M Cronin[1,2], Jeremy L Warner[1,2], Joshua C Denny[1,2]

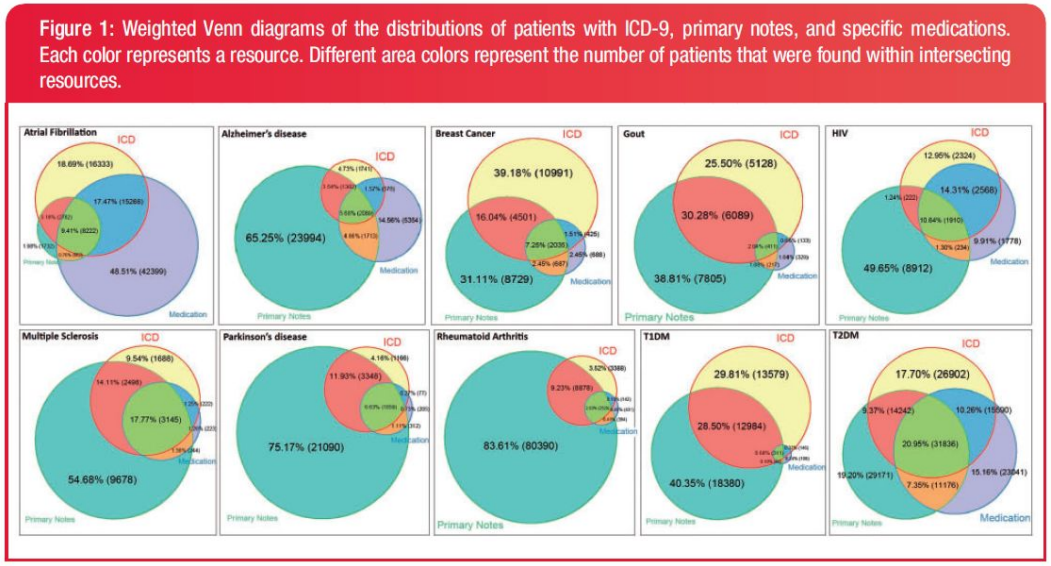
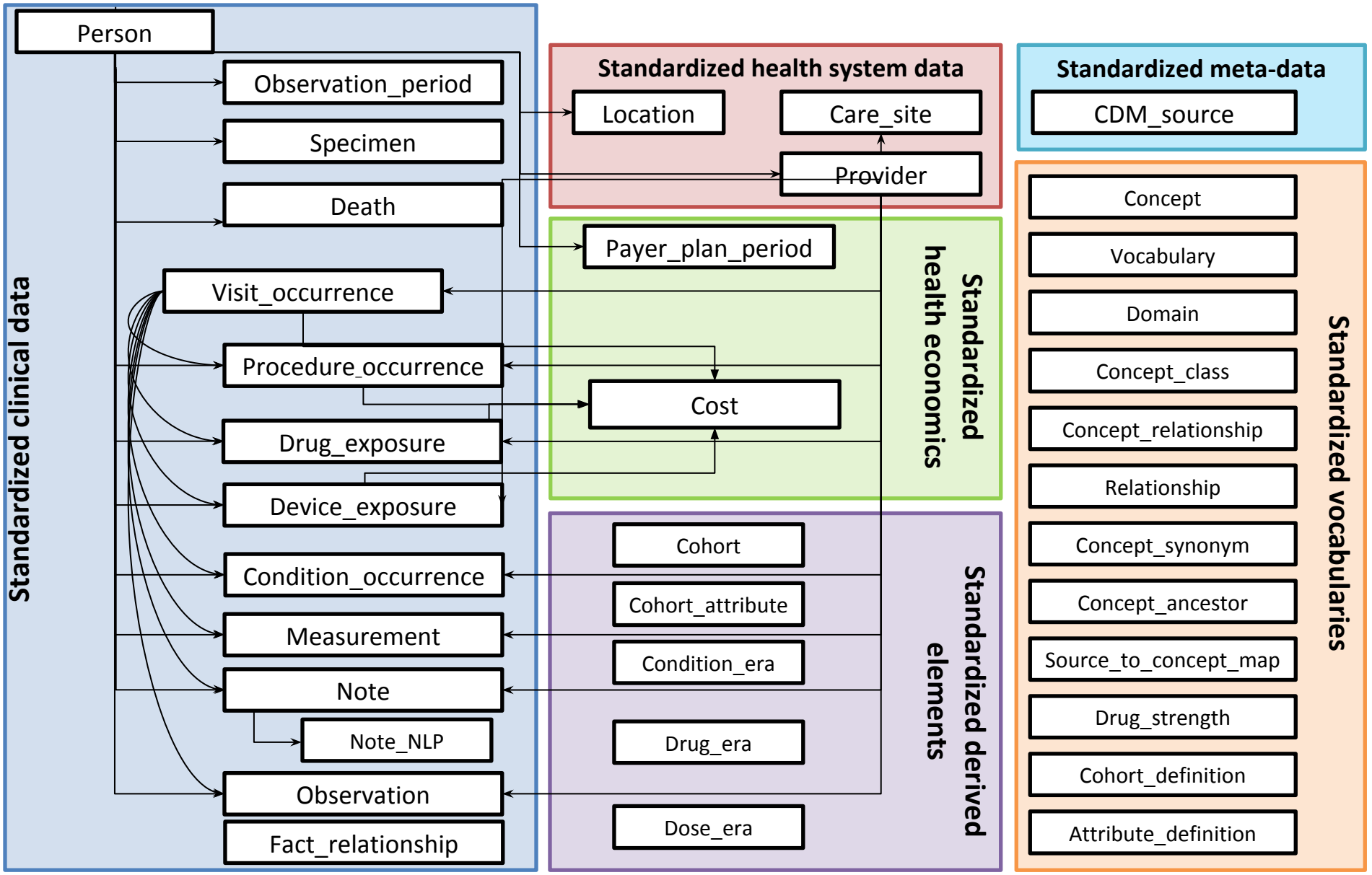


Figure 1: Weighted Venn diagrams of the distributions of patients with ICD-9, primary notes, and specific medications. Each color represents a resource. Different area colors represent the number of patients that were found within intersecting resources.

OMOP CDM Version 5



Constructing Cohorts with SQL

```
# Acute myocardial infarction

AMISql <- "
with events as (
  SELECT PERSON_ID, CONDITION_START_DATE as start_date, CONDITION_END_DATE as end_date
  FROM CMSDESynPUF23m.condition_occurrence co
  WHERE co.condition_concept_id in (
    select descendant_concept_id from CMSDESynPUF23m.concept_ancestor where ancestor_concept_id = 312327 -- Acute MI
  )
)
select top 10 * from events
ORDER BY PERSON_ID, START_DATE; -- not required, just to organize the results
";

conn <- DatabaseConnector::connect(connectionDetails = connectionDetails);
amiEvents <-doQuery(conn, AMISql);
DatabaseConnector::disconnect(conn);

print(amiEvents, row.names=FALSE);
```

```
> print(amiEvents, row.names=FALSE);
 PERSON_ID START_DATE   END_DATE
         7 2009-10-09 2009-10-10
         7 2009-10-09 2009-10-10
         7 2009-10-09 2009-10-10
         7 2009-11-05 2009-11-05
         7 2009-11-05 2009-11-05
         9 2010-05-04 2010-05-04
         9 2010-05-04 2010-05-04
         9 2010-05-04 2010-05-04
         9 2010-05-04 2010-05-04
         9 2010-05-04 2010-05-04
```

FAIL!

Constructing Cohorts with SQL

```
# Maybe select DISTINCT will work?

AMISql <- "
with events as (
  SELECT DISTINCT PERSON_ID, CONDITION_START_DATE as start_date, CONDITION_END_DATE as end_date
  FROM CMSDESynPUF23m.condition_occurrence co
  WHERE co.condition_concept_id in (
    select descendant_concept_id from CMSDESynPUF23m.concept_ancestor where ancestor_concept_id = 312327
  )
)
select top 10 * from events
ORDER BY PERSON_ID, start_date; -- not required, just to organize the results
";

conn <- DatabaseConnector::connect(connectionDetails = connectionDetails);
amiEvents <-doQuery(conn, AMISql);
DatabaseConnector::disconnect(conn);

print(amiEvents);
```

```
> print(amiEvents);
   PERSON_ID START_DATE   END_DATE
1          7 2009-10-09 2009-10-10
2          7 2009-11-05 2009-11-05
3          9 2010-05-04 2010-05-04
4          9 2010-06-04 2010-06-05
5         13 2010-03-19 2010-03-19
6         14 2008-05-13 2008-05-15
7         30 2008-07-13 2008-07-13
8         30 2008-10-31 2008-10-31
9         35 2008-12-06 2008-12-06
10        44 2009-05-06 2009-05-06
>
```

PASS?

Constructing Cohorts with SQL

```
conflictSql <- "
with events as (
  SELECT DISTINCT PERSON_ID, CONDITION_START_DATE as start_date, CONDITION_END_DATE as end_date
  FROM CMSDESynPUF23m.condition_occurrence co
  WHERE co.condition_concept_id in (select descendant_concept_id from CMSDESynPUF23m.concept_ancestor where ancestor_concept_id = 312327)
),
conflicts as (
  select e1.rn, e1.person_id, e1.start_date, e1.end_date, e2.rn as conflict_rn, e2.start_date as conflict_start, e2.end_date as conflict_end
  FROM (
    SELECT row_number() over (partition by person_id order by start_date) as rn, person_id, start_date, end_date
    FROM events
  ) e1
  JOIN (
    SELECT row_number() over (partition by person_id order by start_date) as rn, person_id, start_date, end_date
    FROM events
  ) e2 on e1.person_id = e2.person_id
  where e1.rn < e2.rn
    AND not (e2.start_date > e1.end_date OR e1.start_date > e2.end_date) -- this will identify overlaps
)
select top 10 * from conflicts
ORDER BY PERSON_ID, START_DATE; -- not required, just to organize the results
";

conn <- DatabaseConnector::connect(connectionDetails = connectionDetails);
conflicts <-doQuery(conn, conflictSql);
DatabaseConnector::disconnect(conn);

print(conflicts);
```

```
> print(conflicts);
   RN PERSON_ID START_DATE   END_DATE CONFLICT_RN CONFLICT_START CONFLICT_END
1   1       505 2008-05-10 2008-05-19           2     2008-05-19   2008-05-22
2   1       569 2008-12-15 2008-12-16           2     2008-12-16   2008-12-16
3   1      5099 2008-06-08 2008-06-30           2     2008-06-23   2008-07-02
4   2      5610 2009-03-26 2009-04-30           3     2009-04-22   2009-04-22
5   2      7654 2008-05-21 2008-05-23           3     2008-05-22   2008-05-22
6   1      8115 2010-03-12 2010-03-13           2     2010-03-13   2010-03-13
7   1     15777 2010-10-24 2010-10-26           2     2010-10-24   2010-10-24
8   1     20546 2009-03-10 2009-03-30           2     2009-03-13   2009-03-13
9   2     21364 2008-05-01 2008-05-07           3     2008-05-02   2008-05-02
10  5     21873 2008-07-08 2008-07-13           6     2008-07-09   2008-07-10
>
```

FAIL!

Constructing Cohorts with SQL

```
conflictSql <- "
with events as (
  select PERSON_ID, START_DATE, END_DATE FROM (
    SELECT row_number() over (partition by person_id order by start_date) as rn,
      PERSON_ID, CONDITION_START_DATE as start_date, CONDITION_END_DATE as end_date
    FROM CMSDESynPUF23m.condition_occurrence co
    WHERE co.condition_concept_id in (select descendant_concept_id from CMSDESynPUF23m.concept_ancestor where ancestor_concept_id = 312327)
  ) allEvents
  WHERE allEvents.rn = 1
),
conflicts as (
  select e1.rn, e1.person_id, e1.start_date, e1.end_date, e2.rn as conflict_rn, e2.start_date as conflict_start, e2.end_date as conflict_end
  FROM (
    SELECT row_number() over (partition by person_id order by start_date) as rn, person_id, start_date, end_date
    FROM events
  ) e1
  JOIN (
    SELECT row_number() over (partition by person_id order by start_date) as rn, person_id, start_date, end_date
    FROM events
  ) e2 on e1.person_id = e2.person_id
  where e1.rn < e2.rn
    AND not (e2.start_date > e1.end_date OR e1.start_date > e2.end_date) -- this will identify overlaps
)
select top 10 * from conflicts
ORDER BY PERSON_ID, START_DATE; -- not required, just to organize the results
";

conn <- DatabaseConnector::connect(connectionDetails = connectionDetails);
conflicts <-doQuery(conn, conflictSql);
DatabaseConnector::disconnect(conn);

print(conflicts);
> print(conflicts);
[1] RN              PERSON_ID      START_DATE      END_DATE        CONFLICT_RN     CONFLICT_START CONFLICT_END
<0 rows> (or 0-length row.names)
>
```

PASS! But this is only the beginning!

Constructing Cohorts with SQL

```
# Looking for continous Observation for 365d prior to start.

cohortSql <- "
with cohort as (
  select allEvents.PERSON_ID, allEvents.START_DATE, allEvents.END_DATE FROM (
    SELECT row_number() over (partition by person_id order by start_date) as rn,
      PERSON_ID, CONDITION_START_DATE as start_date, CONDITION_END_DATE as end_date
    FROM CMSDESynPUF23m.condition_occurrence co
    WHERE co.condition_concept_id in (select descendant_concept_id from CMSDESynPUF23m.concept_ancestor where ancestor_concept_id = 312327)
  ) allEvents
  JOIN CMSDESynPUF23m.observation_period op on allEvents.person_id = op.person_id
  WHERE allEvents.rn = 1
    AND allEvents.start_date >= op.observation_period_start_date and allEvents.end_date < op.observation_period_end_date
    AND datediff(d, op.observation_period_start_date, allEvents.start_date) > 365
)
select count(*) from cohort
";

conn <- DatabaseConnector::connect(connectionDetails = connectionDetails);
cohortCount <-doQuery(conn, cohortSql);
DatabaseConnector::disconnect(conn);
```

Continuous Observation: the time between a single observation period start and observation period end. When we want '365d continuous observation prior', we are saying that there must be at least 365 days between the observation period start and the cohort event.

Constructing Cohorts with SQL

```
# Looking for continous Observation for 365d prior to start, males

cohortSql <- "
with cohort as (
  select allEvents.PERSON_ID, allEvents.START_DATE, allEvents.END_DATE FROM (
    SELECT row_number() over (partition by person_id order by start_date) as rn,
      PERSON_ID, CONDITION_START_DATE as start_date, CONDITION_END_DATE as end_date
    FROM CMSDESynPUF23m.condition_occurrence co
    WHERE co.condition_concept_id in (select descendant_concept_id from CMSDESynPUF23m.concept_ancestor where ancestor_concept_id = 312327)
  ) allEvents
  JOIN CMSDESynPUF23m.observation_period op on allEvents.person_id = op.person_id
  JOIN CMSDESynPUF23m.person p on allEvents.person_id = p.person_id
  WHERE allEvents.rn = 1
    AND allEvents.start_date >= op.observation_period_start_date and allEvents.end_date < op.observation_period_end_date
    AND datediff(d, op.observation_period_start_date, allEvents.start_date) > 365
    AND p.gender_concept_id = 8507
)
select count(*) from cohort
";

conn <- DatabaseConnector::connect(connectionDetails = connectionDetails);
cohortCount <-doQuery(conn, cohortSql);
DatabaseConnector::disconnect(conn);
```

Checking by Gender/Sex: Stored in the Person table's gender_concept_id column. Standard concept_ids are used to specify the gender value.

Constructing Cohorts with SQL

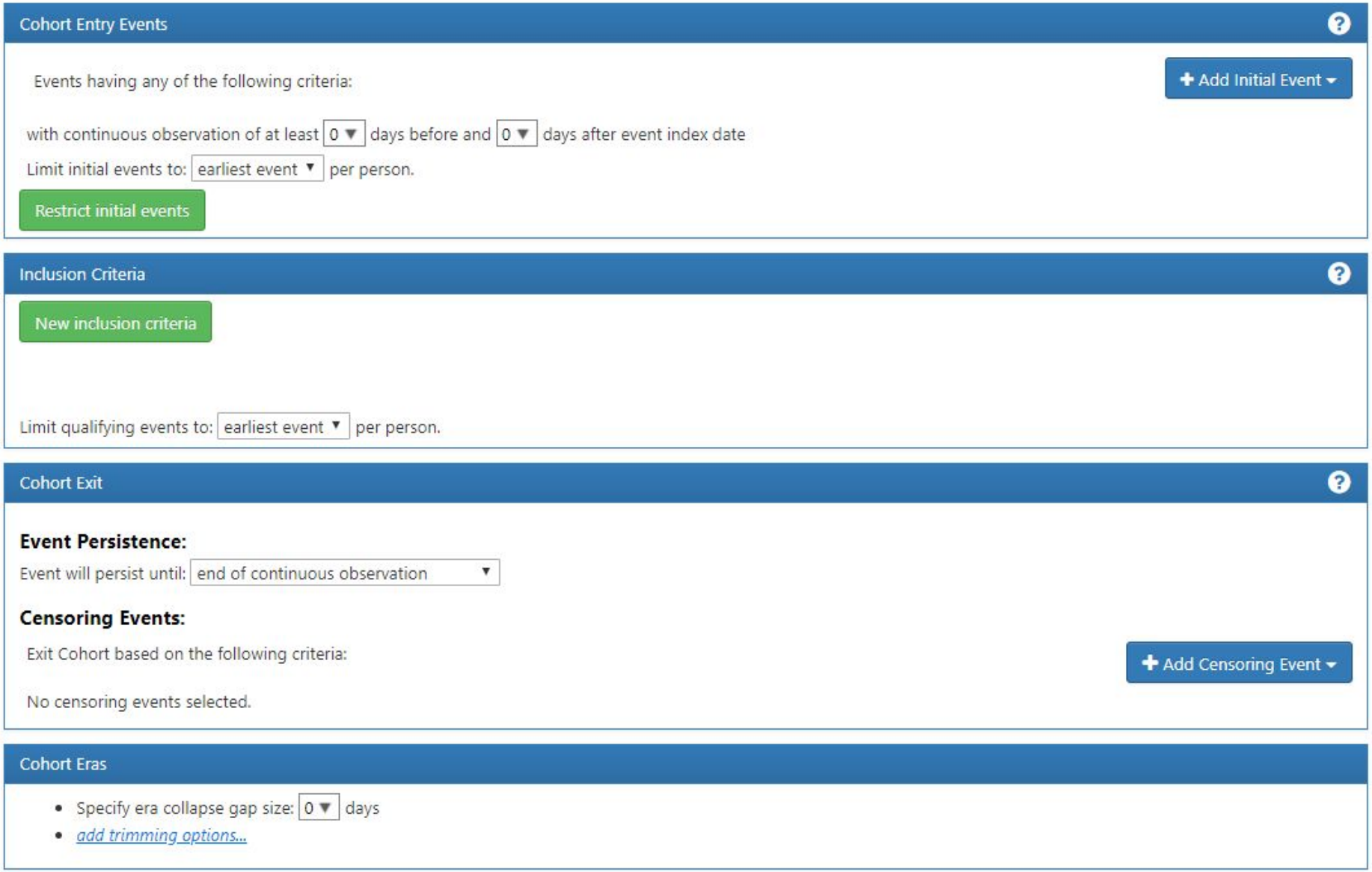
```
# Looking for continous Observation for 365d prior to start, males, hypertension in past 183d

cohortSql <- "
with cohort as (
  select allEvents.PERSON_ID, allEvents.START_DATE, allEvents.END_DATE FROM (
    SELECT row_number() over (partition by person_id order by start_date) as rn,
      PERSON_ID, CONDITION_START_DATE as start_date, CONDITION_END_DATE as end_date
    FROM CMSDESynPUF23m.condition_occurrence co
    WHERE co.condition_concept_id in (select descendant_concept_id from CMSDESynPUF23m.concept_ancestor where ancestor_concept_id = 312327)
  ) allEvents
  JOIN CMSDESynPUF23m.observation_period op on allEvents.person_id = op.person_id
  JOIN CMSDESynPUF23m.person p on allEvents.person_id = p.person_id
  WHERE allEvents.rn = 1
    AND allEvents.start_date >= op.observation_period_start_date and allEvents.end_date < op.observation_period_end_date
    AND datediff(d, op.observation_period_start_date, allEvents.start_date) > 365
    AND p.gender_concept_id = 8507
    AND EXISTS (
      select 1 from CMSDESynPUF23m.condition_occurrence co
      where co.condition_start_date >= dateadd(d,-183,allEvents.start_date)
        AND allEvents.person_id = co.person_id
        and co.condition_concept_id in (select descendant_concept_id
                                        from CMSDESynPUF23m.concept_ancestor
                                        where ancestor_concept_id = 320128) -- Essential Hypertension
    )
)
select count(*) from cohort
";

conn <- DatabaseConnector::connect(connectionDetails = connectionDetails);
cohortCount <-doQuery(conn, cohortSql);
DatabaseConnector::disconnect(conn);

print(cohortCount);
```

Inclusion Rule:
Essential Hypertension within prior 183d

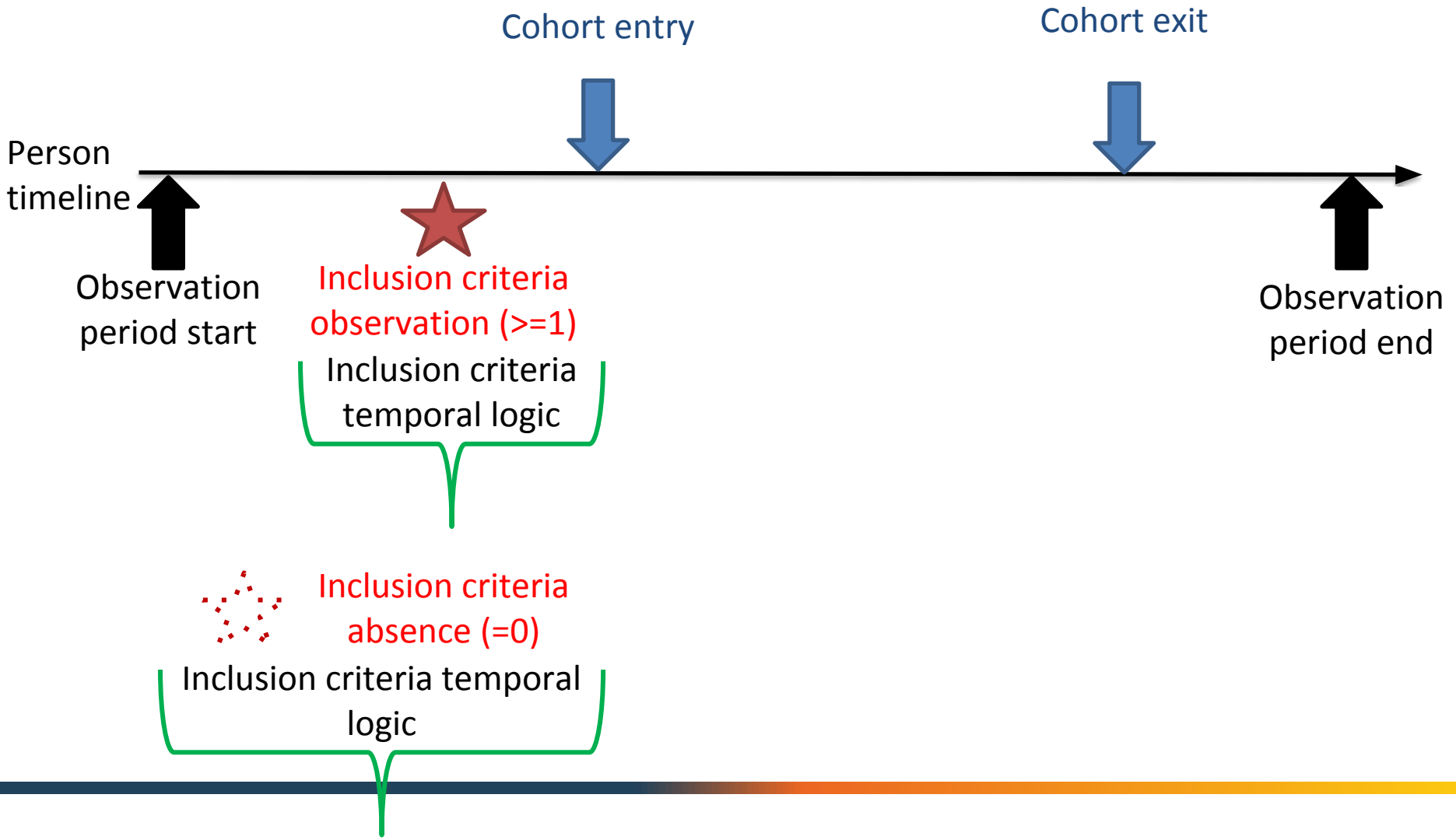
```
> print(cohortCount);
  COUNT
1  8684
>
```

Why So few?

Tools to Standardize Chort Building



Dissecting the anatomy of a cohort definition



Student Activity #1

Gowtham Rao

RESEARCH

OPEN ACCESS

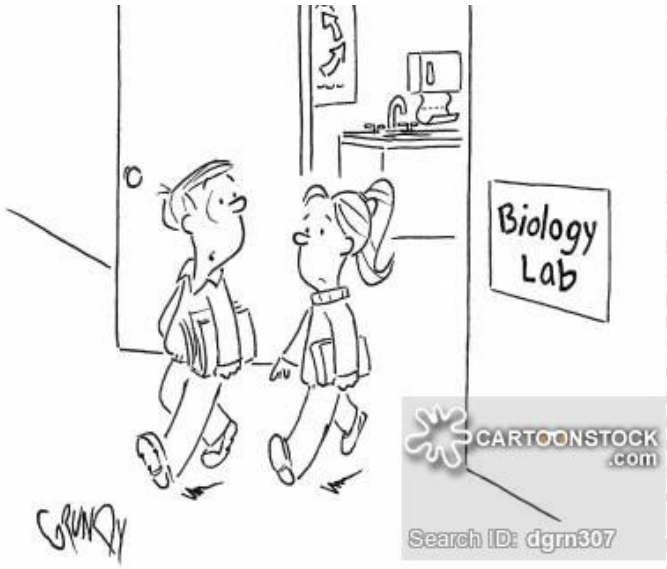
Risks and benefits of direct oral anticoagulants versus warfarin in a real world setting: cohort study in primary care

Yana Vinogradova,[1] Carol Coupland,[1] Trevor Hill,[1] Julia Hippisley-Cox[1]

10 minutes break

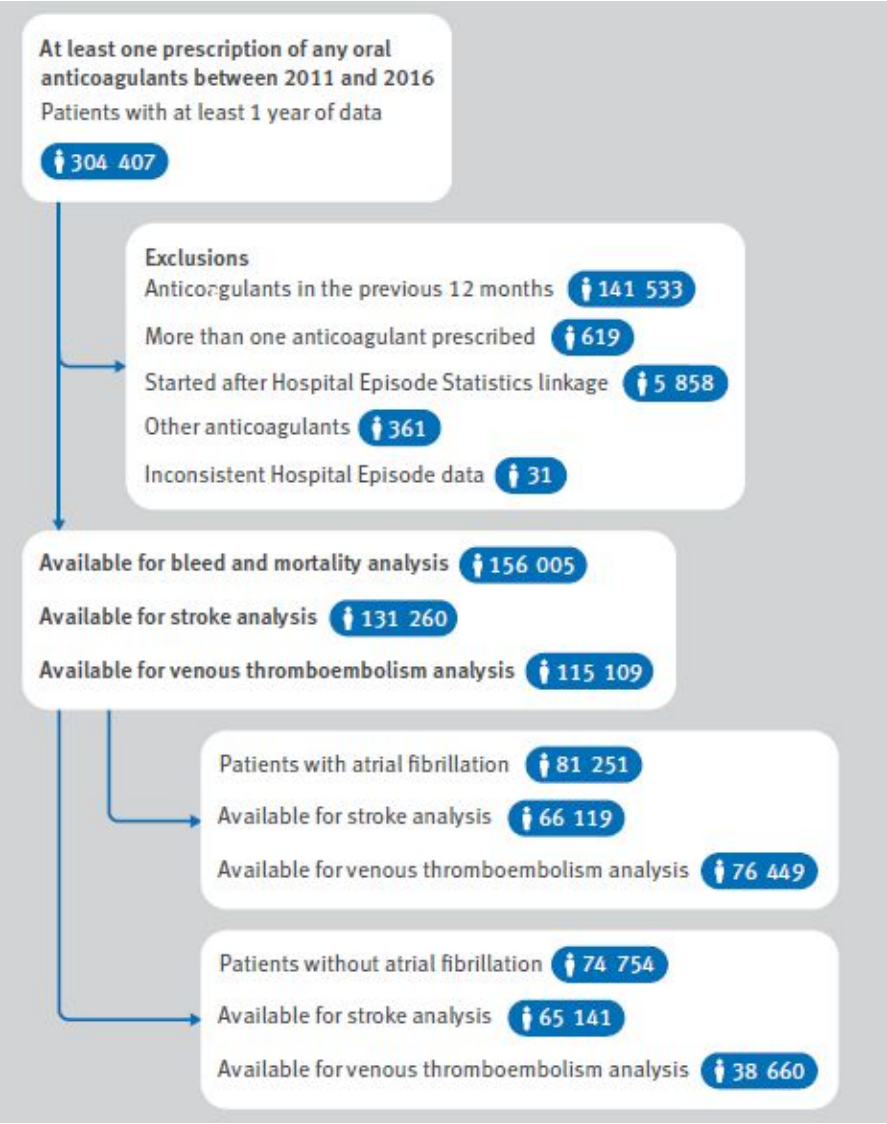
(AKA, read the paper in case anyone didn't do the homework ☺)

Get ready to start dissecting soon!



"Never bring a pet to school . . . especially if it's a frog!"

Dissecting the study



eTable 5 ***All patients****: Rates per 1000py of outcomes by drug and adjuste*

	Person-years	Adjusted hazard ratios (95%CI)[a]
Major bleeding		
Warfarin	111,823	Reference
Dabigatran	6,117	
Rivaroxaban	20,581	
Apixaban	10,744	
Intracranial bleed		
Warfarin	113,705	Reference
Dabigatran	6,218	
Rivaroxaban	20,823	
Apixaban	10,805	
Haematuria		
Warfarin	112,791	Reference
Dabigatran	6,173	
Rivaroxaban	20,730	
Apixaban	10,789	
Haemoptysis		
Warfarin	113,667	Reference
Dabigatran	6,204	
Rivaroxaban	20,820	
Apixaban	10,811	
All GI bleed		
Warfarin	113,044	Reference
Dabigatran	6,180	
Rivaroxaban	20,717	
Apixaban	10,774	
Upper GI bleed		
Warfarin	113,143	Reference
Dabigatran	6,181	
Rivaroxaban	20,728	
Apixaban	10,779	

Vinogradova Yana, Coupland Carol, Hill Trevor, Hippisley-Cox Julia. Risks and benefits of direct oral anticoagulants versus warfarin in a real world setting: cohort study in primary care BMJ 2018; 362 :k2505

Target/Comparator Cohort

Study design

We used a new-user design to capture all events occurring after starting treatment and to reduce the impact of confounding.[30] For a study period from January 2011 to the latest date of Hospital Episode Statistics linked data (October 2016 for QResearch and March 2016 for CPRD), patients prescribed the oral anticoagulants warfarin, dabigatran, rivaroxaban, and apixaban, and aged from 21 to 99 at study entry date, formed the cohort. The entry date was defined as the date of the first prescription of any of the anticoagulant drugs. To facilitate a direct comparison between new users of direct oral anticoagulants (DOACs) against new users of warfarin, and to reduce the impact of indication bias, patients were excluded if they had any anticoagulant prescription in the last 12 months before the entry date. To ensure the quality of data, patients were also excluded if they had either fewer than 12 months of records before entry or had no valid ~~Townsend score.~~

Patients were followed from their first prescription of an anticoagulant until they experienced an outcome of interest or were censored. Patients were censored when: they stopped or suspended treatment (at 30 days after the expected end date of any prescription, where the gap between the expected prescription end date and the start date of any subsequent prescription was more than 30 days); they switched treatment (at the day before the prescription start for a different anticoagulant); they left a practice (at the day of deregistration); they died; or the study period ended. For the analyses of dosages, we also censored patients when they changed to a different dose.

Cohorts in the paper?

OBJECTIVE

To investigate the associations between direct oral anticoagulants (DOACs) and risks of bleeding, ischaemic stroke, venous thromboembolism, and all cause mortality compared with warfarin.

MAIN OUTCOME MEASURES

Major bleeding leading to hospital admission or death. Specific sites of bleeding and all cause mortality were also studied.

Target cohort:
- Person
- cohort start date
- cohort end date

What is the target (T) cohort?
People on Warfarin

Comparator cohort:
- Person
- cohort start date
- cohort end date

What are the comparator cohort?
Directly Acting Oral Anticoagulants (DOACs)
(dabigatran, rivaroxaban, apixaban - each is a cohort)

Outcome cohort:
- Person
- cohort start date
- cohort end date

Ischemic stroke
Venous thromboembolism
Hematuria
Haemoptysis
Lower GI Bleed
Death

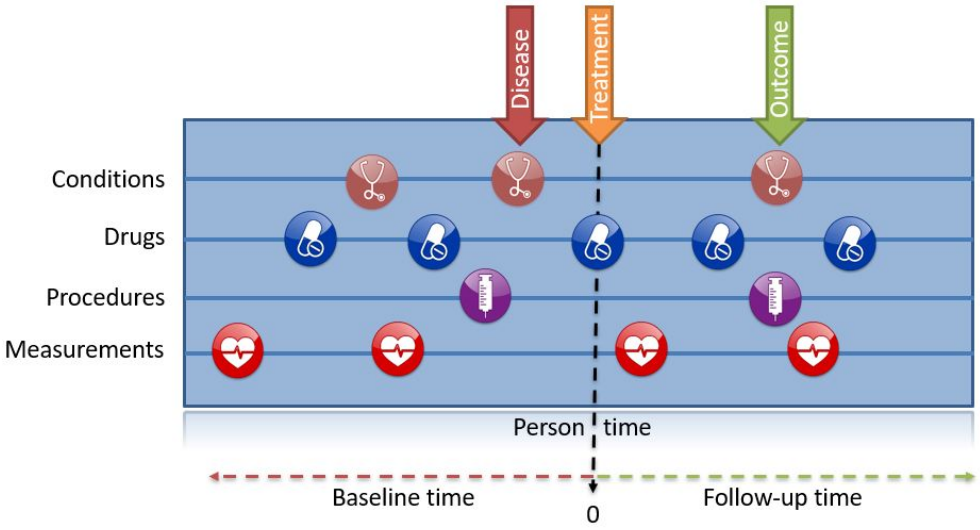
Vinogradova Yana, Coupland Carol, Hill Trevor, Hippisley-Cox Julia. Risks and benefits of direct oral anticoagulants versus warfarin in a real world setting: cohort study in primary care BMJ 2018; 362 :k2505

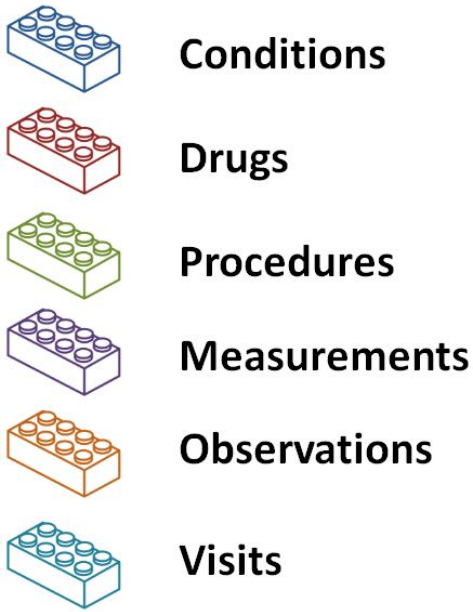
OHDSI fundamentals

Person level dataset

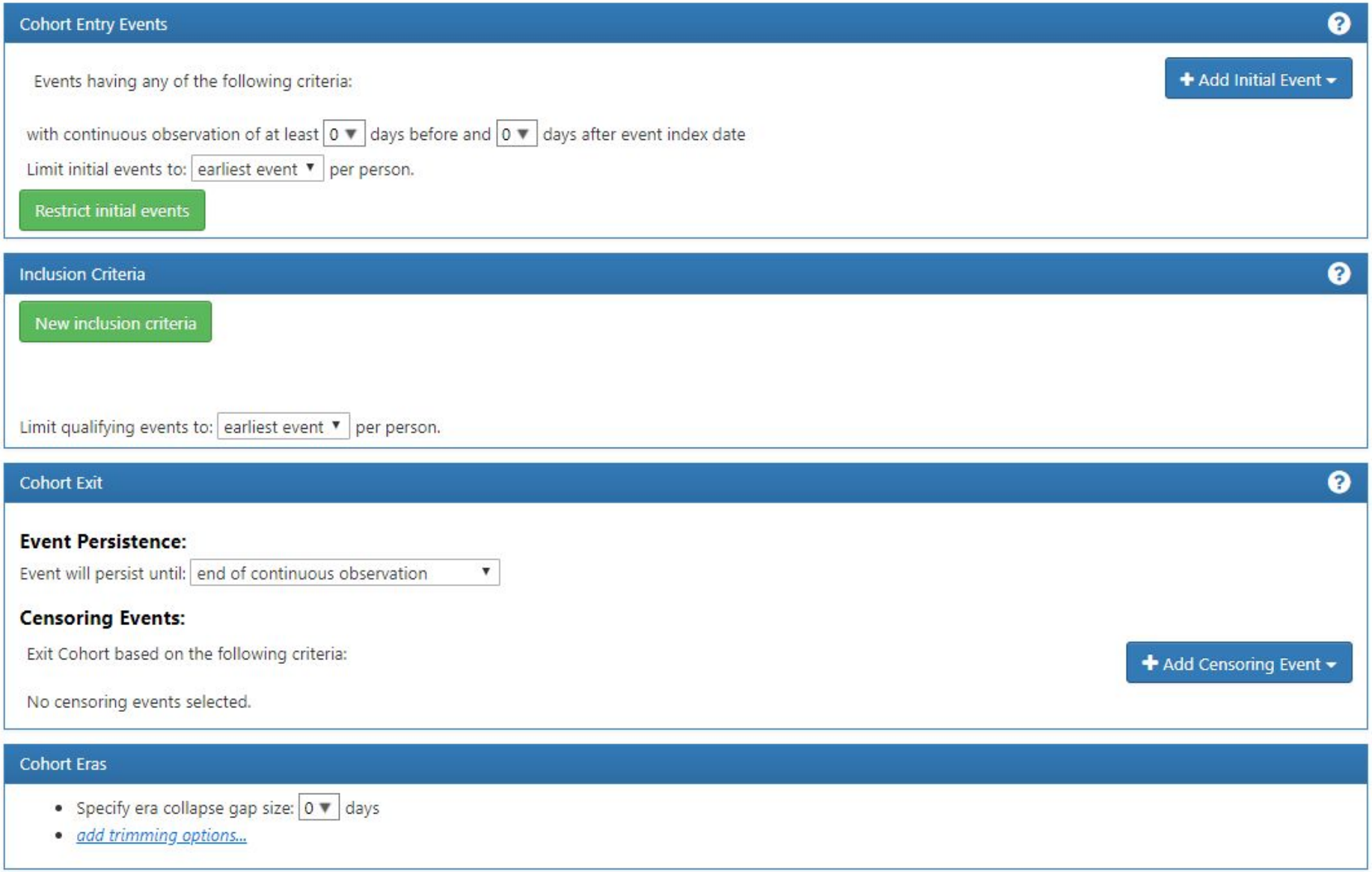
With time-stamped events

Events organized in domains





Tools to Standardize Cohort Building



Cohort (T/C) Entry Events

Cohort Entry Events

- What are the initial/index events that entered a person into the cohort?
- Observation before/after event?
- Earliest, latest, or all events?

Cohort (T/C) Entry Events

Cohort Entry Events

- What are the initial/index events that entered a person into the cohort?
  - *First prescription of dabigatran, rivaroxaban or apixaban (T); warfarin (C)*
- Observation before/after event?
  - *At least 12 mo of observation before index date*
- Earliest, latest, or all events?
  - *1st event (new user)*

Cohort (T/C) Inclusion Criteria

Inclusion Criteria

- Any additional inclusion criteria? (or exclusion criteria)

- Inclusion criteria vs additional criteria attributes of initial event?

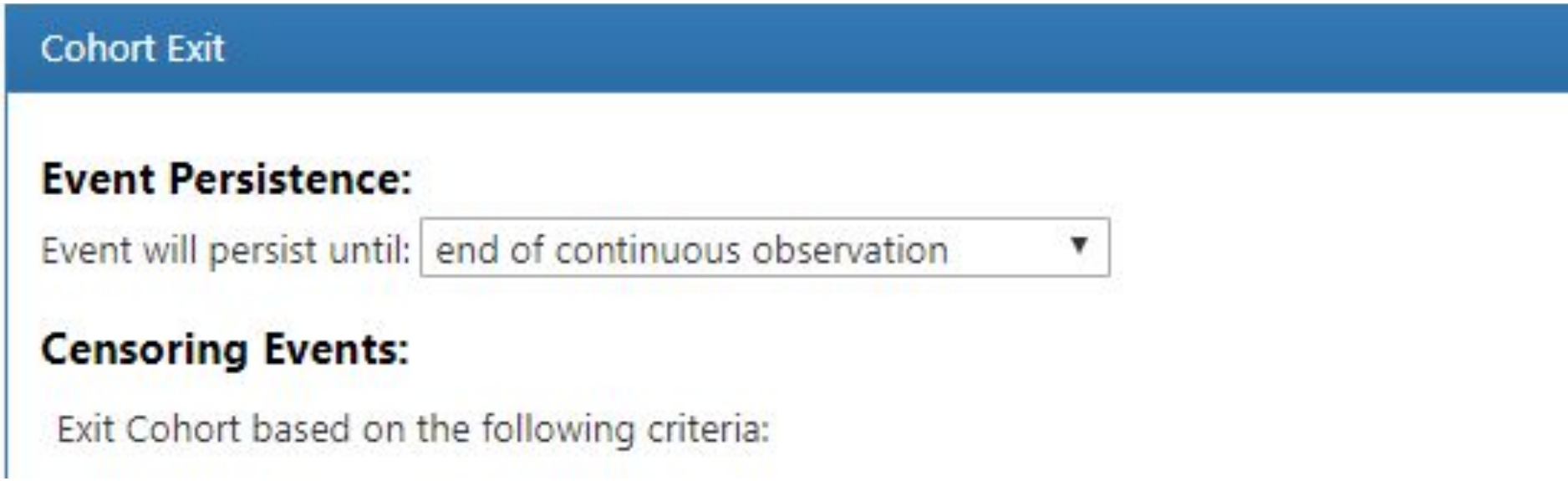
- Earliest, latest, or all events?

Cohort (T/C) Inclusion Criteria

Inclusion Criteria

- Any additional inclusion criteria? (or exclusion criteria)
  - Age 21-99
  - Valid Townsend score; can we represent this?
  - Excluded if any anticoagulation rx in past 12 mo
  - Require 0 occurrences in Atlas (is exclusion)
- Inclusion criteria vs additional criteria attributes of initial event?
  - Either?
- Earliest, latest, or all events?

Cohort (T/C) Cohort Exit



Cohort (T/C) Exit - Event Persistence

Cohort Exit

What is the cohort exit?

Fixed time or end of exposure?

Cohort Exit

Cohort Exit

Censoring Events: Fixed time or end of exposure?

If the subject had the outcome of interest?

Cohort Exit

Cohort Exit

Censoring Events: Fixed time or end of exposure?

- End of study period;
- End date: Oct 2016-QResearch or March 2016-CPRD

If the subject had the outcome of interest?

Rule Based Phenotyping

Christopher Knoll
Gowtham Rao

New User Cohort

Cohort Entry

- First exposure to drug of interest (use first event per person)

Inclusion Criteria

- Is there age restrictions? Gender?
- Are there any prior indications required?
- Are there any prior indications to exclude?

Cohort Exit

- How long should the remaining events persist cohort presence?
- Should we identify any observed data that would cancel cohort presence?

Cohort Eras

- Since we are only using first events, there will only be one era per person, so no need for ‘gap size’ value.

Outcome Cohort

Cohort Entry

- Allow multiple events per person
- Can the outcome be defined from looking at data across multiple domains?

Inclusion Criteria

- Is the entry event enough? Should we use additional data to improve positive predictive value?

Cohort Exit

- Outcome events are usually ‘moments’ in time, but can be repeatedly recorded in patient data. So, it is reasonable to specify a fixed duration from the event to cover the expected duration of the outcome.

Cohort Eras

- Alternatively, we can specify a collapse gap to group outcomes together if they are within days of each other into a single cohort era.

Building Concept Sets

Concept Sets Are:

- Lists of vocabulary concepts that are used to identify patient records.
- They can be constructed by using CONCEPT_ANCESTOR
- They can include concepts from different domains, but usually you should restrict a single concept to a single clinical idea for a specific CDM domain.
- They resemble 'code expressions' you may have seen in publications:
  - 401.x 402.x (Hypertension), 580-589 (Chronic Kidney Disease)

Concept Sets Are Not:

- Temporal associations between occurrence of one observation and another. You can't say '401.1 with 580.1' in a Concept Set. Those would be defined as 2 separate concept sets (Hypertension, Chronic Kidney Disease), and cohort criteria would be defined to look for those codes in the same visit/time period.

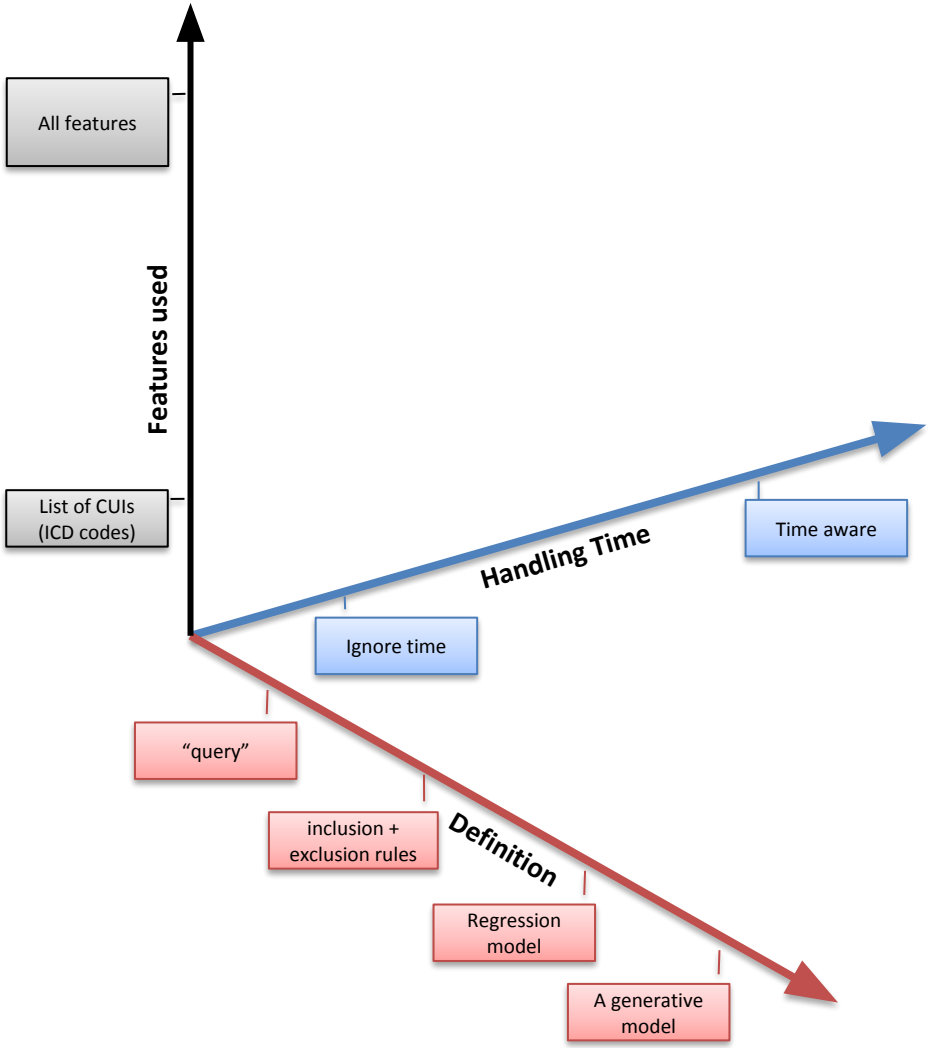
Lunch and Student Exercise

Probabilistic Phenotyping in Aphrodite

Juan M. Banda

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.

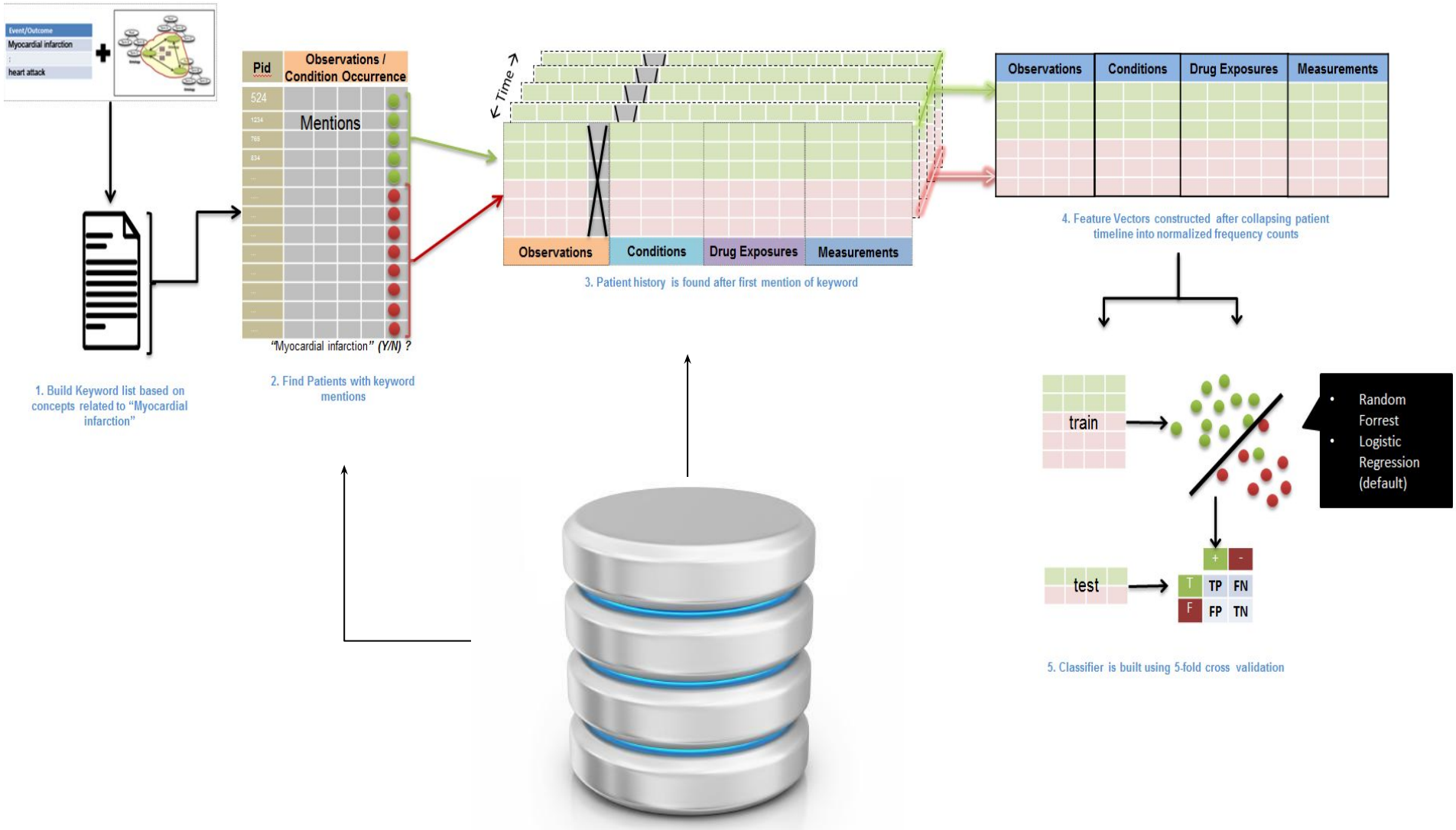
Learning using imperfectly labeled data

1. Create a heuristic , i.e. “labeling function”, that is high precision (can be low recall).
2. Use the imperfect labeling function to create large amounts of training data.
3. Train a classifier using the imperfectly labeled data (noisy data)
4. Evaluate performance.

Does it work?

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 %

APHRODITE Overview



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

2. Find Patients with keyword mentions

3. Patient history is found after first mention of keyword

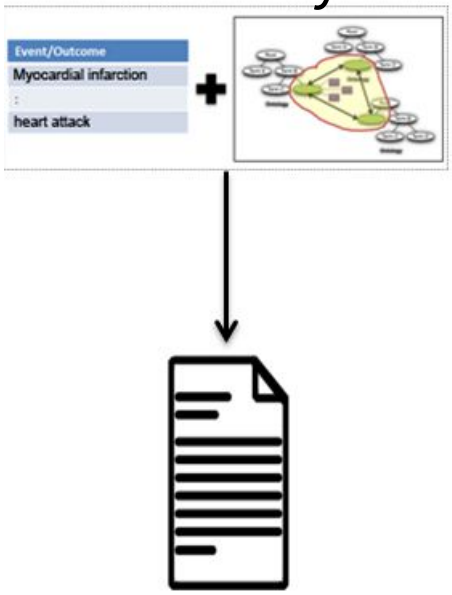
4. Feature Vectors constructed after collapsing patient timeline into normalized frequency counts

5. Classifier is built using 5-fold cross validation

OHDSI /OMOP CDMv5

Step 1 – Build Keyword List

Vocabulary v5



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

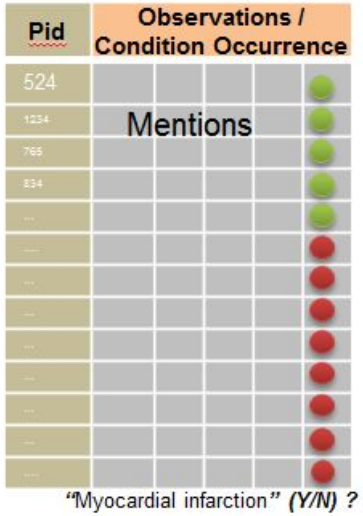
Generated keywords for "Myocardial Infarction"

```
4329847 Myocardial infarction   312327  Acute myocardial infarction
4329847 Myocardial infarction   314666  Old myocardial infarction
4329847 Myocardial infarction   319038  Postmyocardial infarction syndrome
4329847 Myocardial infarction   319039  Acute posterior myocardial infarction
4329847 Myocardial infarction   438447  Acute myocardial infarction of inferolateral wall
4329847 Myocardial infarction   439693  True posterior myocardial infarction
4329847 Myocardial infarction   4011131 Acute infarction of papillary muscle
4329847 Myocardial infarction   4030582 Postoperative myocardial infarction
4329847 Myocardial infarction   4096808 Post-infarction panhypopituitarism
4329847 Myocardial infarction   4108217 Subsequent myocardial infarction
4329847 Myocardial infarction   4108220 Rupture of papillary muscle as current complication following acute myocardial infarction
4329847 Myocardial infarction   4108668 Acute anteroapical infarction
4329847 Myocardial infarction   4108669 Acute myocardial infarction of atrium
4329847 Myocardial infarction   4108677 Subsequent myocardial infarction of anterior wall
4329847 Myocardial infarction   4110951 Anterior myocardial infarction NOS
4329847 Myocardial infarction   4110952 Posterior myocardial infarction NOS
4329847 Myocardial infarction   4110953 Lateral myocardial infarction NOS
4329847 Myocardial infarction   4110954 Inferior myocardial infarction NOS
4329847 Myocardial infarction   4111389 Other specified anterior myocardial infarction
```

Function call:

```
wordLists <- buildKeywordList(conn, aphrodite_concept_name, cdmSchema, dbms)
```

Step 2 – Find Cases and Controls

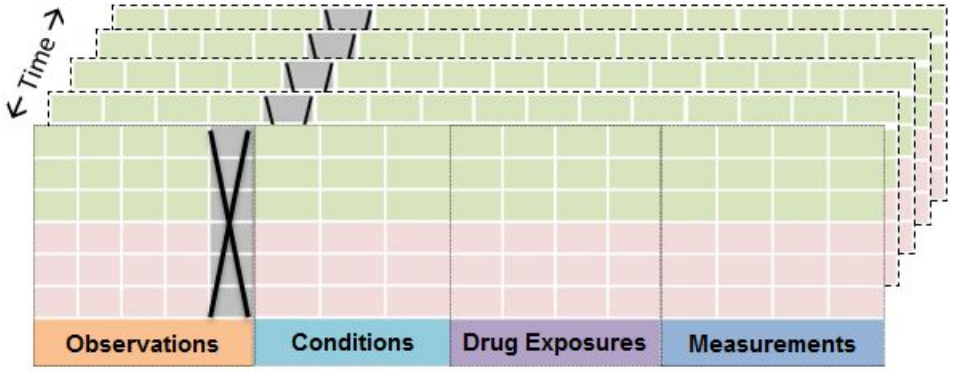


CDM v5

Function call:

```
casesANDcontrolspatient_ids_df<- getdPatientCohort(conn, dbms,
as.character(keywordList_FF$V3),as.character(ignoreList_FF$V3), cdmSchema,nCases,nControls)
```

Step 3 – Extract Patient Data

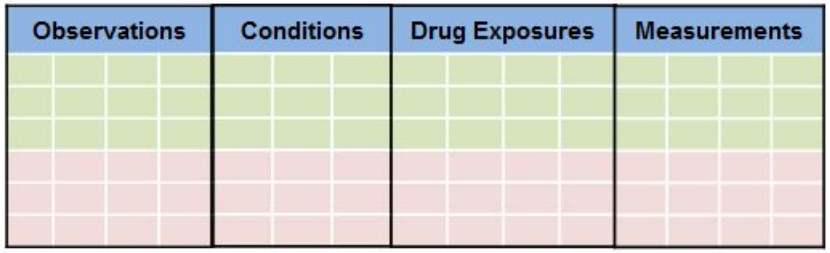


CDM v5

Function calls:

```
dataFcases <- getPatientData(conn, dbms, cases, as.character(ignoreList_FF$V3), flag, cdmSchema)
dataFcontrols <- getPatientData(conn, dbms, controls, as.character(ignoreList_FF$V3), flag, cdmSchema)
```

Step 4 – Create Feature Vector

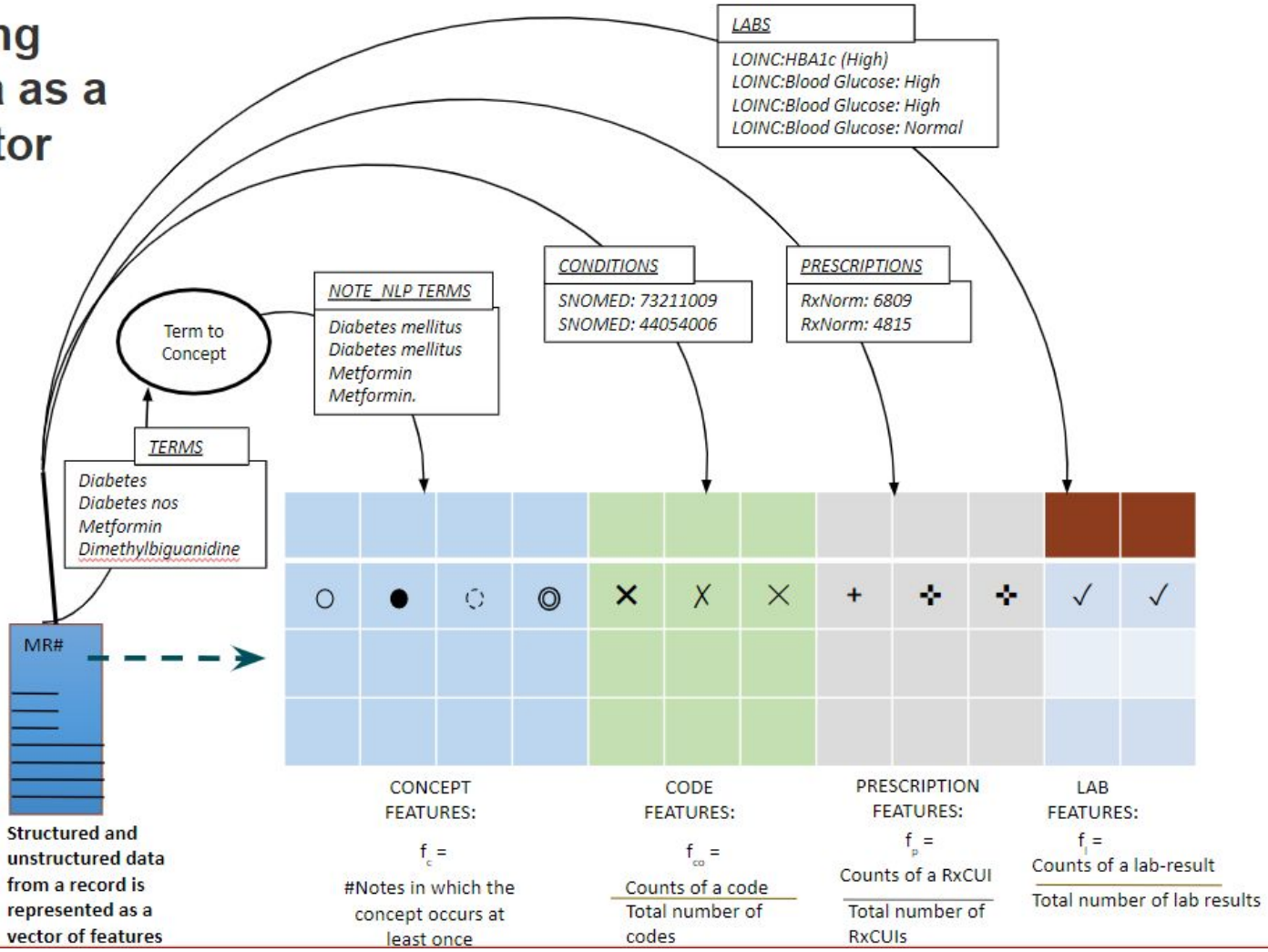


4. Feature Vectors constructed after collapsing patient timeline into normalized frequency counts

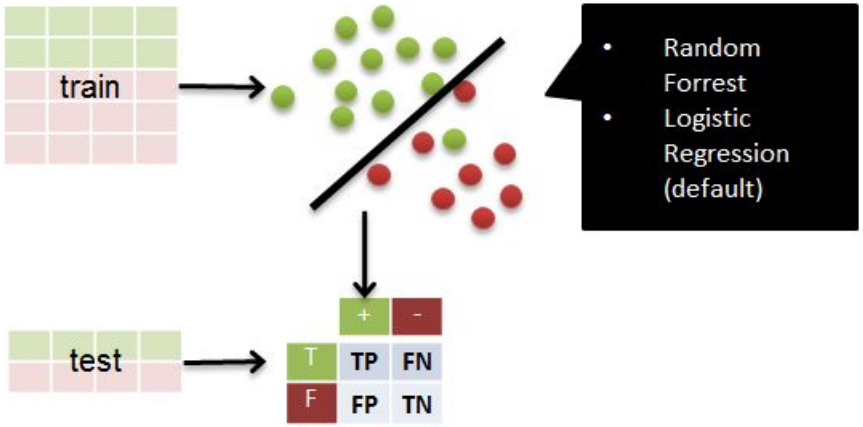
Function calls:

```
fv_all<-buildFeatureVector(flag, dataFcases,dataFcontrols)
fv_full_data <- combineFeatureVectors(flag, data.frame(cases), controls, fv_all, outcomeName)
```

Representing patient data as a feature vector



Step 5 – Build Model



5. Classifier is built using 5-fold cross validation

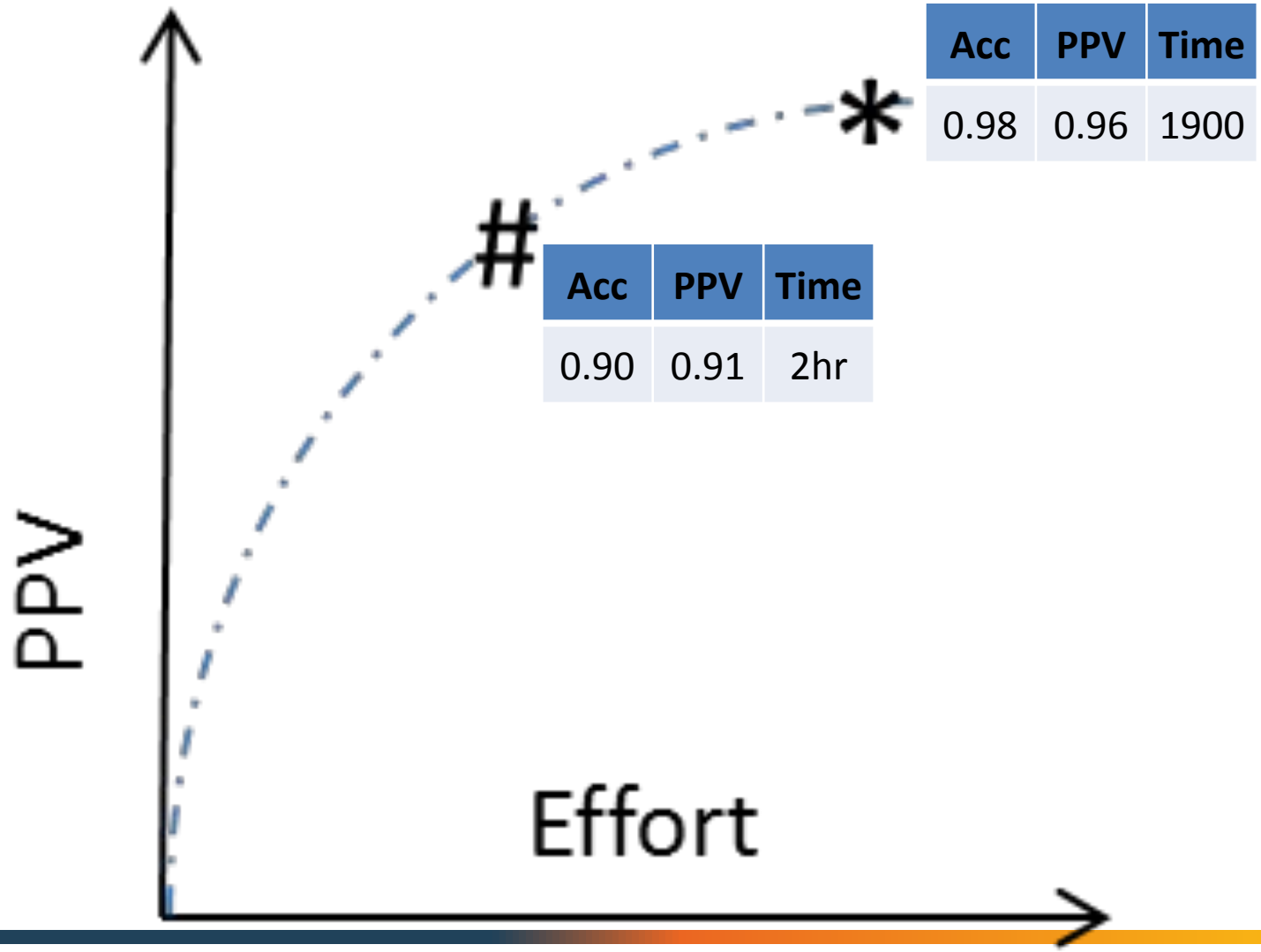
Function calls:

```
model_predictors <- buildModel(flag, fv_full_data, outcomeName, folder)
                    model<-model_predictors$model
predictorsNames<-model_predictors$predictorsNames
```

Step 6 – Get Evaluation Results

```
          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'
```

Effort precision trade off

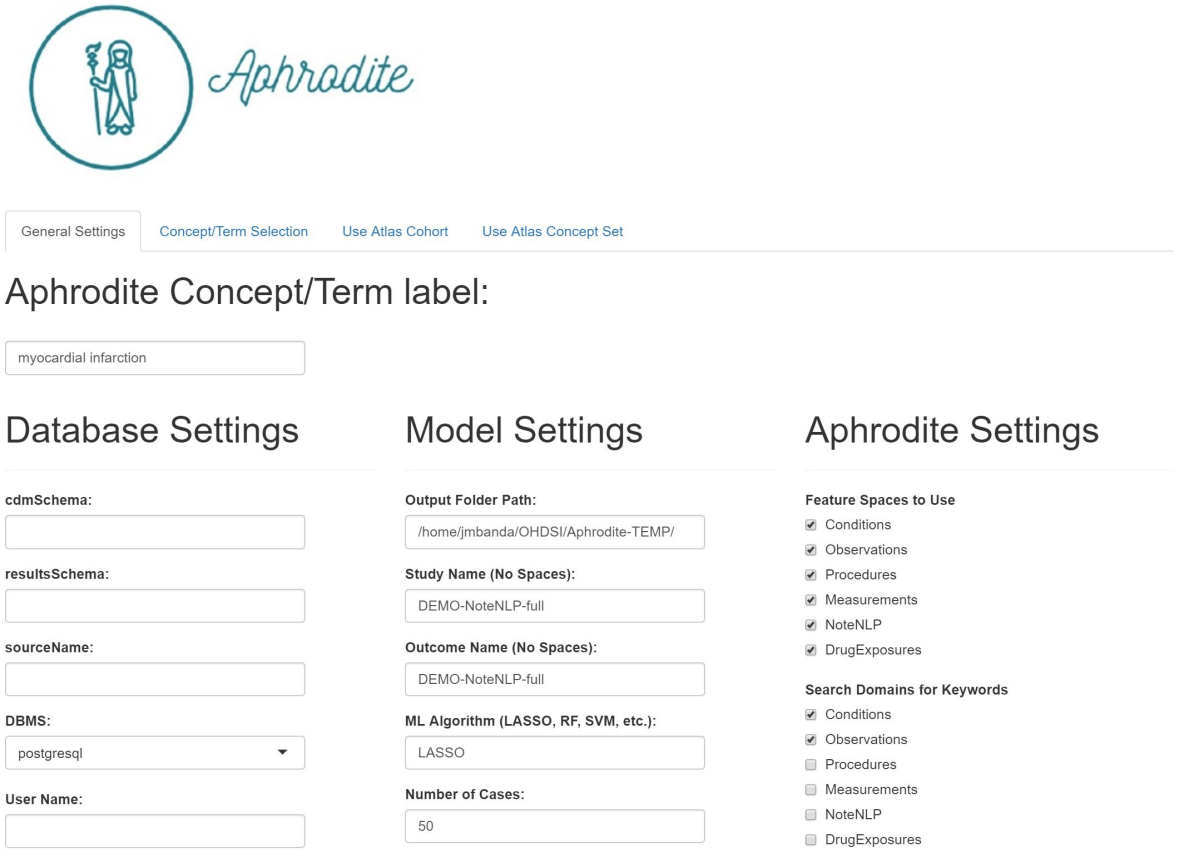


Acc	PPV	Time
0.90	0.91	2hr

Acc	PPV	Time
0.98	0.96	1900

Live Demo

Set all configuration parameters on the Shiny app:



Pick labeling heuristic terms or concepts

General Settings | Concept/Term Selection | Use Atlas Cohort | Use Atlas Concept Set

Include keyword Selection | Ignore keyword Selection

Aphrodite Concept/Term label Selected:

myocardial infarction

Show 10 entries

Search:

	Label concept_id	concept_name	Related concept_id	concept_name
1	4329847	Myocardial infarction	312327	Acute myocardial infarction
2	4329847	Myocardial infarction	314666	Old myocardial infarction
3	4329847	Myocardial infarction	319038	Postmyocardial infarction syndrome
4	4329847	Myocardial infarction	319039	Acute posterior myocardial infarction
5	4329847	Myocardial infarction	434376	Acute myocardial infarction of anterior wall
6	4329847	Myocardial infarction	436706	Acute myocardial infarction of lateral wall
7	4329847	Myocardial infarction	438170	Acute myocardial infarction of inferior wall
8	4329847	Myocardial infarction	438438	Acute myocardial infarction of anterolateral wall
9	4329847	Myocardial infarction	438447	Acute myocardial infarction of inferolateral wall

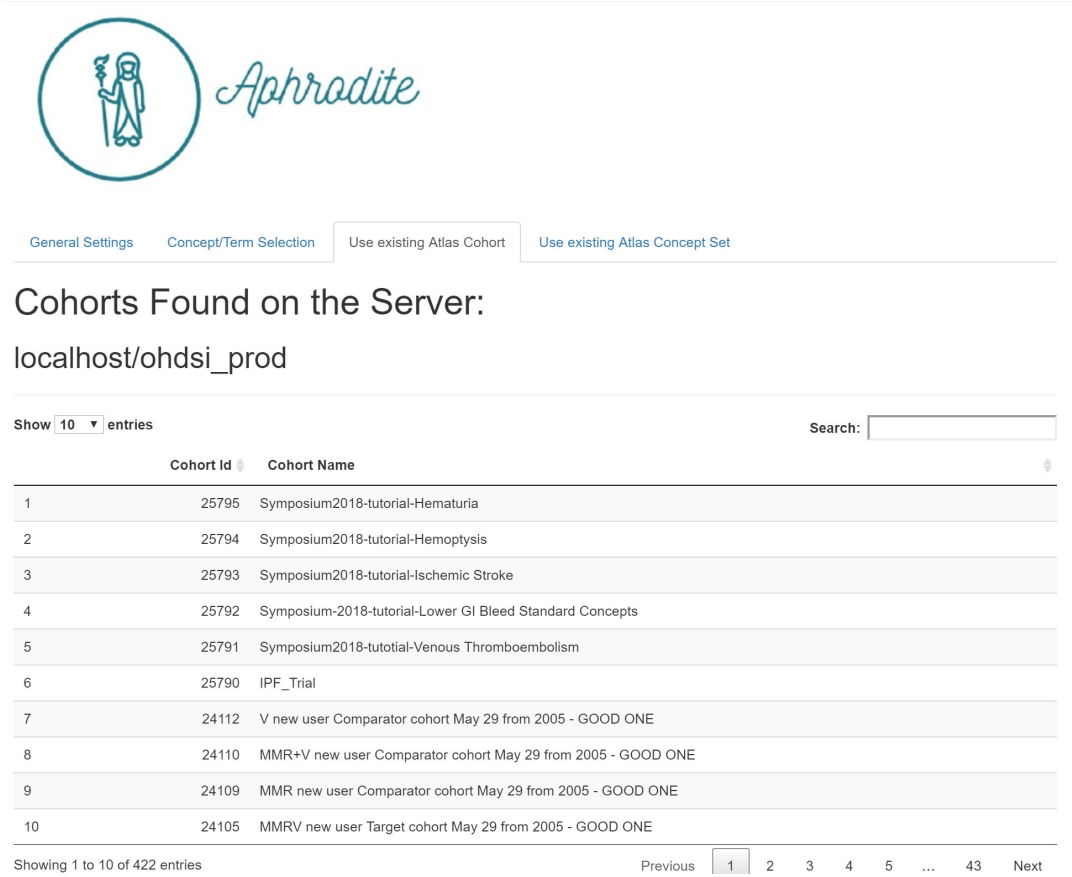
Execution code will be generated by Shiny app

```
# Copyright 2015 Observational Health Data Sciences and Informatics
#
# This file is part of:
#
#   APHRODITE
#
# Licensed under the Apache License, Version 2.0 (the "License");
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#
# @author Stanford University Center for Biomedical Informatics - Shah Lab
# @author Juan M. Banda

# Install necessary packages if needed, remove comments
# install.packages("devtools")
# install_github("ohdsi/Aphrodite")

#library(Aphrodite)
source('/home/jmbanda/OHDSI/Aphrodite/R/functions.R')
library("SqlRender", lib.loc="/usr/local/lib/R/site-library")
library("plyr",lib.loc="/usr/local/lib/R/site-library")
library("caret",lib.loc="/usr/local/lib/R/site-library")
library("pROC",lib.loc="/usr/local/lib/R/site-library")
library("data.table",lib.loc="/usr/local/lib/R/site-library")
library("DatabaseConnector",lib.loc="/usr/local/lib/R/site-library")
```

…And copy and paste/Run in your R session… done!

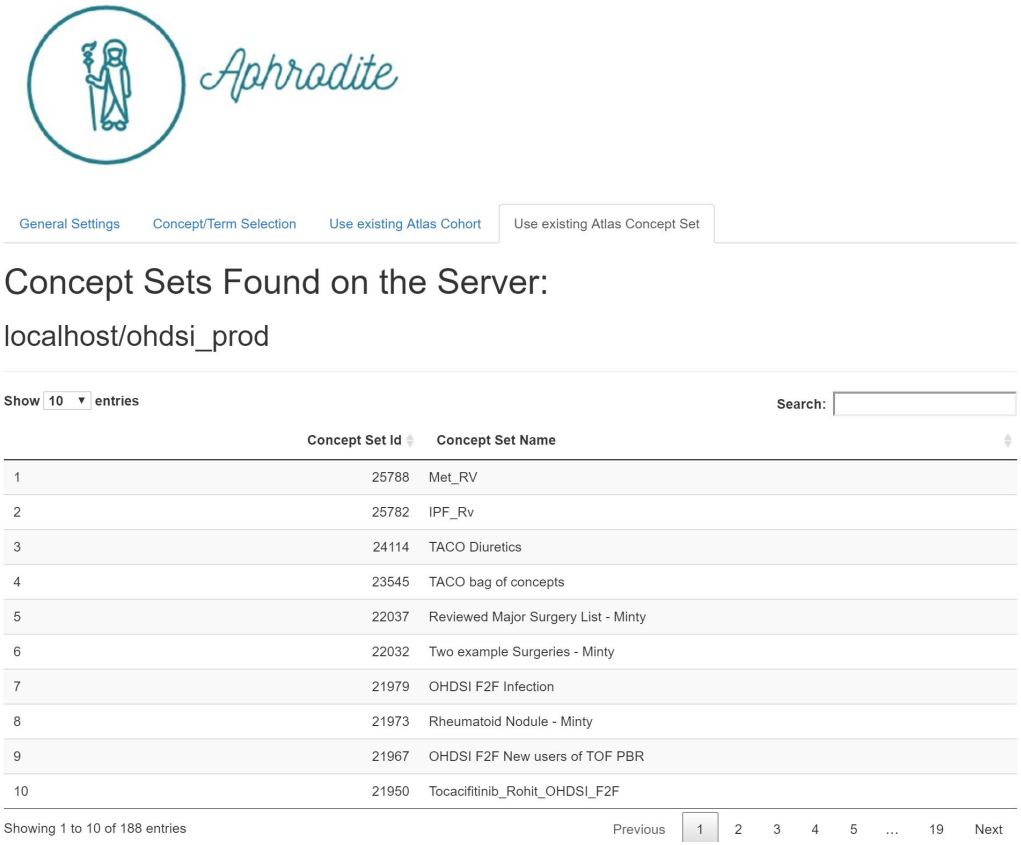
New functionality:

Build APHRODITE models using Atlas cohorts



New functionality:

Build APHRODITE models using Atlas concept sets



Patients found on Atlas cohorts

Outcome Cohort	Stanford Patients
Venous Thromboembolism	11,916
Lower GI Bleed	6,799
Ischemic Stroke	23,029
Hemoptysis	4,633
Hematuria	31,793

APHRODITE models

	Aphrodite Models (75% training - 25% unseen testing)								
Outcome Cohort	**ROC**	**PPV**	**Sensitivity**	**Specificity**	**F1 score**	**TP**	**TN**	**FP**	**FN**
Venous Thromboembolism	0.9012	0.9661	0.8462	0.9605	0.9022	2,878	2,456	101	523
Lower GI Bleed	0.8143	0.9547	0.9118	0.9525	0.9328	1,623	1,543	77	157
Ischemic Stroke	0.8398	0.8913	0.9403	0.8967	0.9151	5,132	5,432	626	326
Hemoptysis	0.8432	0.8740	0.9388	0.8822	0.9053	1,013	1,093	146	66
Hematuria	0.9231	0.9045	0.8133	0.8924	0.8565	7,190	6,298	759	1,651

Can we find new patients for the cohorts?

Yes!, however, trying to make predictions on 1 million patients.... not ideal… on our infrastructure

On a smaller set 100,000 non-overlapping patients

Venous Thromboembolism: 717 patients with 95% probability of being in the cohort!
Lower GI Bleed: 313 patients with 95% probability
Ischemic Stroke: 987 patients with 95% probability
Hemoptysis: 242 patients with 95% probability
Hematuria: 1,312 patients with 95% probability

Are those right?? ..... according to the models... ~9 out of 10 SHOULD be

Unsolved questions

- Do we share learned models, or do we share the modeling building workflow?
- How do we share the model or the workflow?
- How do we know if those patients are correct?

Need more information?

How-to Video Demo:

https://www.youtube.com/watch?v=u4PrWhbQkzY

Code Repository:

https://github.com/OHDSI/Aphrodite

Questions?

Try Aphrodite:

- https://github.com/OHDSI/Aphrodite

…. Thanks!

@drjmbanda

Phenotype Evaluation

Joel Swerdel

Evaluating the Performance of Phenotype Algorithms

Joel Swerdel
Epidemiology Analytics
Janssen R&D

Agenda

- Why do we need a evaluate phenotype performance?
- Development of a tool for the evaluation
- Results from the evaluation

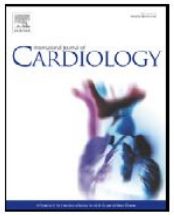
Validating Algorithms

Many research studies have attempted to validate algorithms

Contents lists available at ScienceDirect

International Journal of Cardiology

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



Review

Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations

Bruna Rubbo [a,*], Natalie K. Fitzpatrick [a], Spiros Denaxas [a], Marina Daskalopoulou [b], Ning Yu [a], Riyaz S. Patel [a,c], UK Biobank Follow-up and Outcomes Working Group, Harry Hemingway [a]

- Examined 33 studies
- Found significant heterogeneity in algorithms used, validation methods, and results

Validating an Algorithm

		Truth	
		Positive	**Negative**
Test	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Test – Comes from the algorithm/cohort definition

Truth – Some form of "gold standard" reference

Ex.: True Positive (TP) – Test and Truth agree Positive

For a complete validation of the algorithm we need:

1) Sensitivity: TP / (TP + FN)
2) Specificity: TN / (TN + FP)
3) Positive Predictive Value: TP / (TP + FP)

Evaluating Performance of Algorithm - Examples

Abstract

Purpose—To validate an algorithm based upon International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) codes for acute myocardial infarction (AMI) documented within the Mini-Sentinel Distributed Database (MSDD).

Methods—**Using an ICD-9-CM-based algorithm (hospitalized patients with 410.x0 or 410.x1 in primary position),** we identified a random sample of potential cases of AMI in 2009 from 4 Data Partners participating in the Mini-Sentinel Program. **Cardiologist reviewers used information abstracted from hospital records to assess the likelihood of an AMI diagnosis based on criteria from the joint European Society of Cardiology and American College of Cardiology Global Task Force.** Positive predictive values (PPVs) of the ICD-9-based algorithm were calculated.

Results—Of the 153 potential cases of AMI identified, hospital records for 143 (93%) were retrieved and abstracted. **Overall, the PPV was 86.0% (95% confidence interval; 79.2%, 91.2%).** PPVs ranged from 76.3% to 94.3% across the 4 Data Partners.

Conclusions—The overall PPV of potential AMI cases, as identified using an ICD-9-CM-based algorithm, may be acceptable for safety surveillance; however, PPVs do vary across Data Partners. This validation effort provides a contemporary estimate of the reliability of this algorithm for use in future surveillance efforts conducted using the FDA's MSDD.

Evaluating Performance of Algorithm - Examples

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2009; **18**: 1064–1071
Published online 28 August 2009 in Wiley InterScience (www.interscience.wiley.com) **DOI**: 10.1002/pds.1821

SUMMARY
Purpose Studies of non-steroidal anti-inflammatory drugs (NSAIDs) and cardiovascular events using administrative data require identification of incident acute myocardial infarctions (AMIs) and information on whether confounders differ by NSAID status.
Methods **We identified patients with a first AMI hospitalization from Tennessee Medicaid files as those with primary ICD-9 discharge diagnosis 410.x and hospitalization stay of>2 calendar days.** Eligible persons were non-institutionalized, aged 50–84 years between 1999–[illegible] potential first AMI, a systematic sample (n¼350) was selected for review. **Using defined criteria, we classified events using chest pain history, EKG, and cardiac enzymes, and calculated the positive predictive value (PPV) for definite or probable AMI.**
[illegible] and no AMI, respectively. **PPV for any definite or probable AMI was 92.8% (95% CI 89.6–95.2); for an AMI without an event in the past year 91.7% (95% CI 88.3–94.2), and for an incident AMI was 72.7% (95% CI 67.7–77.2).** Age-adjusted prevalence of current smoking (46.4% vs. 39.1%, p¼0.35) and aspirin use (36.9% vs. 35.9%, p¼0.90) was similar among NSAID users and non-users
Conclusions ICD-9 code 410.x had high predictive value for identifying AMI. Among those with AMI, smoking and aspirin use was similar in NSAID exposure groups, suggesting these factors will not confound the relationship between NSAIDs and cardiovascular outcomes.

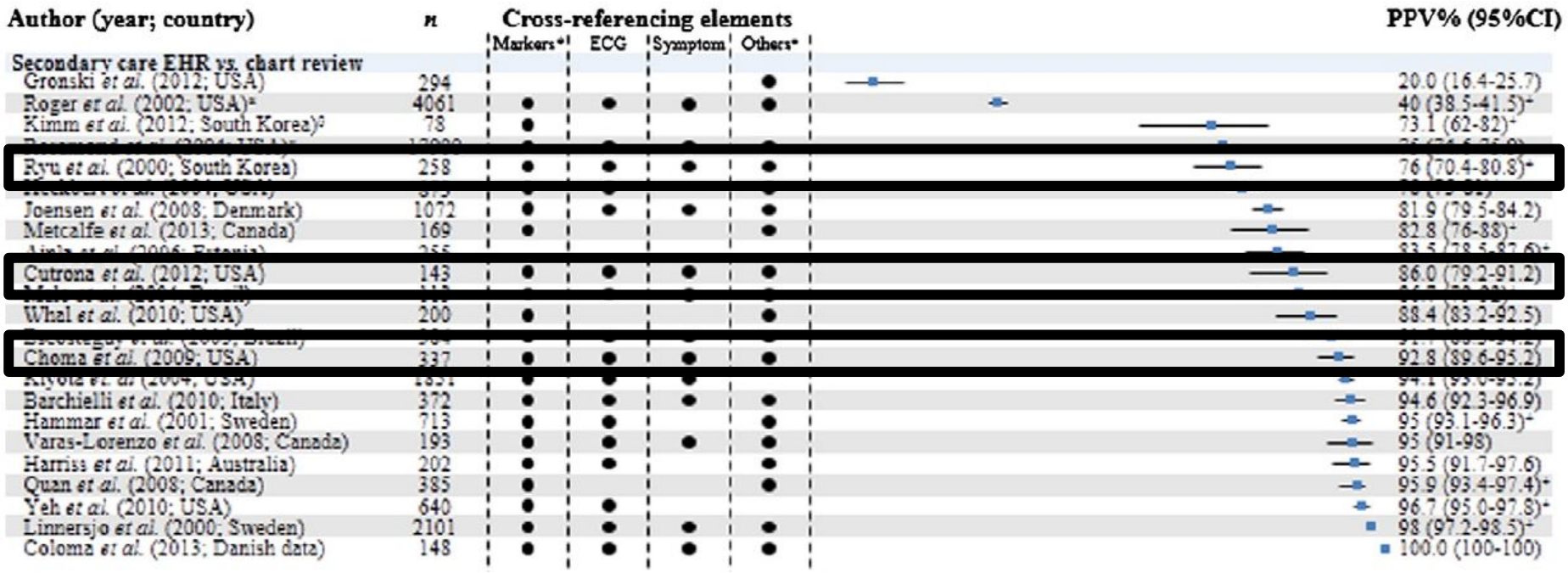
Evaluating Performance of Algorithm - Examples

Yonsei Medical Journal
Vol. 41, No. 5, pp. 570–576, 2000

Abstract

We attempted to assess the accuracy of the International Classification of Diseases (ICD) codes for myocardial infarction (MI) in medical insurance claims, and to investigate the reasons for any inaccuracy. This study was designed as a preliminary study to establish a surveillance system for cardiovascular diseases in Korea. A sample of 258 male patients who were diagnosed with MI from 1993 to 1997 was selected from the Korea Medical Insurance Corporation cohort (KMIC cohort: 183,461 people). The registered medical record administrators were trained in the survey technique, and gathered data by investigating the medical records of the study subjects from March 1999 to May 1999. **The definition of MI for this study included symptoms pursuant to the diagnostic criteria of chest pain, electrocardiogram (ECG) findings, cardiac enzyme and results of coronary angiography or nuclear scan.** We asked the record administrators for the reasons of incorrectness for cases where the final diagnosis was 'not MI'. **The accuracy rate of the ICD codes for MI in medical insurance claims was 76.0% (196 cases) of the study sample,** and 3.9% (ten cases) of the medical records were not available due to hospital closures, non-computerization or missing information. Nineteen cases (7.4%) were classified as insufficient due to insufficient records of chest pain, ECG findings, or cardiac enzymes. The major reason of inaccuracy in the disease code for MI in medical insurance claims was 'to meet the review criteria of medical insurance benefits (45.5%)'. The department responsible for the inaccuracy was the department of inspection for medical insurance benefit of the hospitals.

Evaluating Performance of Algorithm



Evaluating Performance of Algorithm

- Conclusion – for MI → no "gold standard" algorithm available
- Process is very costly and time consuming
- Small sample sizes → wide variation in estimates with wide confidence intervals

- In 33 studies "validating" algorithms, all reported PPV but:
  - Only 11 reported sensitivity
  - Only 5 reported specificity
  - **Is this really validation?**

The Value of Positive Predictive Value

- PPV is almost always reported in validation studies – easiest to assess
- Sensitivity and Specificity much less frequently reported
  - High cost and time to evaluate
- BUT – sensitivity and specificity are the actual characteristics of the test
  - PPV is a function of sensitivity, specificity and **prevalence** of Heath Outcome of Interest (HOI)

PPV Example – 1 Test, 2 Populations

Test Characteristics:
Sensitivity = 75%
Specificity = 99.9%

Population = 10,000

Prevalence = 1%		Truth	
		Positive	Negative
Test	Positive	75	10
	Negative	25	9890
	Total	100	9900

PPV =
75 / (75 + 10) =
88%

Prevalence = 5%		Truth	
		Positive	Negative
Test	Positive	375	10
	Negative	125	9490
		500	9500

PPV =
375 / (375 + 10) =
97%

PPV Example – 1 Population, 2 Tests

PPV = 90%

Population = 10,000

Prevalence = 5%		Truth	
		Positive	Negative
Test #1	Positive	90	10
	Negative	410	9490
	Total	500	9500

PPV = 90/(90+10) = 90%

Sens = 90/500 = 18%

Spec = 9490/9500 = 99.9%

Prevalence = 5%		Truth	
		Positive	Negative
Test #2	Positive	360	40
	Negative	140	9460
		500	9500

PPV = 360/(360+40) = 90%

Sens = 360/500 = 72%

Spec = 9460/9500 = 99.6%

Living with Algorithms

- Algorithms are used in most research with observational data
- Many ways to define an algorithm for any health outcome
- Each definition will have its own performance characteristics
  - Need to validate the algorithm to understand these characteristics
- Validation of an algorithm to be used in an observational dataset through chart review is likely not possible
  - Costly
  - Time consuming
  - Data is usually not available

Validating Algorithms in Observational Data

		Truth	
		Positive	**Negative**
Test	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Test – Comes from the algorithm/cohort definition

Truth – Some form of "gold standard" reference

Possible alternative for finding the "Truth"

Diagnostic Predictive Models

Prediction models used to estimate the probability of having a particular disease or outcome.

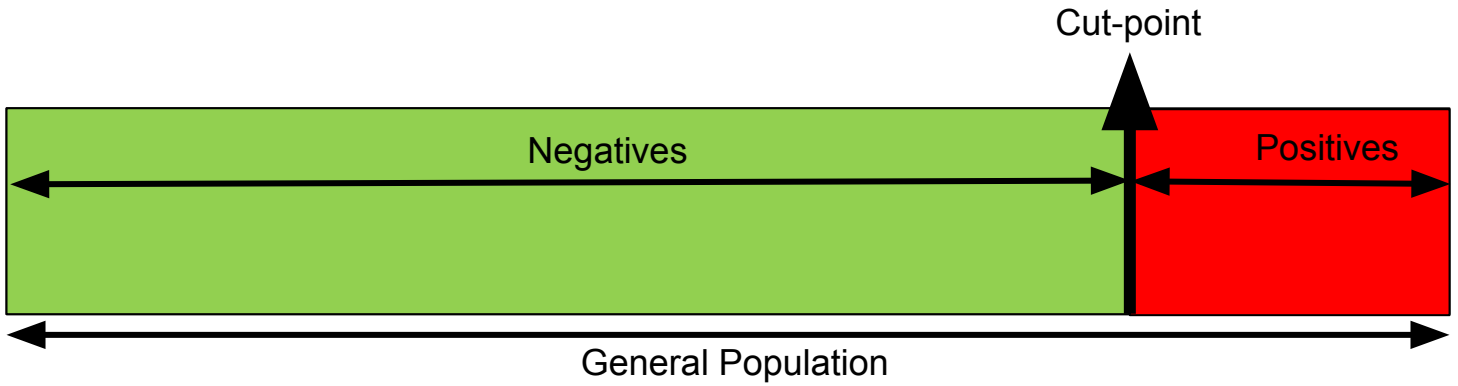
Finding the Truth – using Diagnostic Predictive Models

Step 1: Find a Gold standard of subjects for the HOI

Step 2: Develop the predictive model

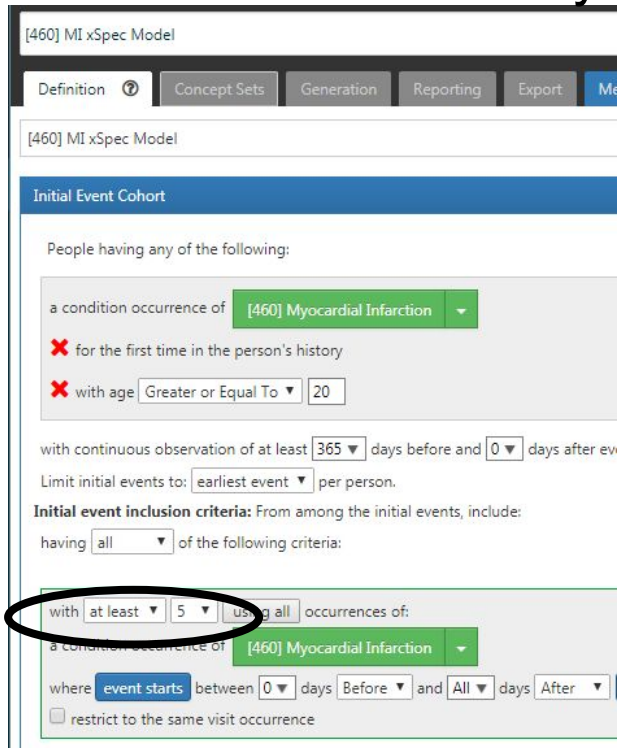
Step 3: Apply the model to a general population

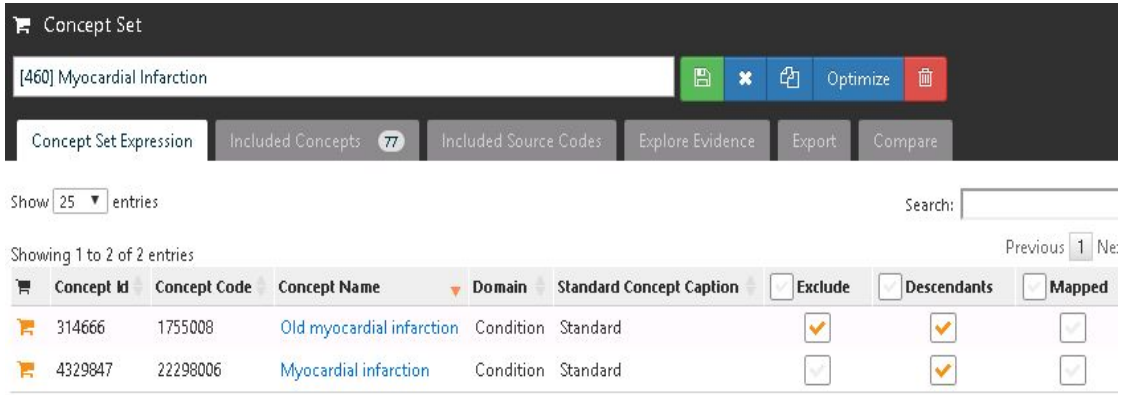
Step 4: Determine a cut-point from the model:
At what level of predicted probability is positive?



Finding a Gold Standard

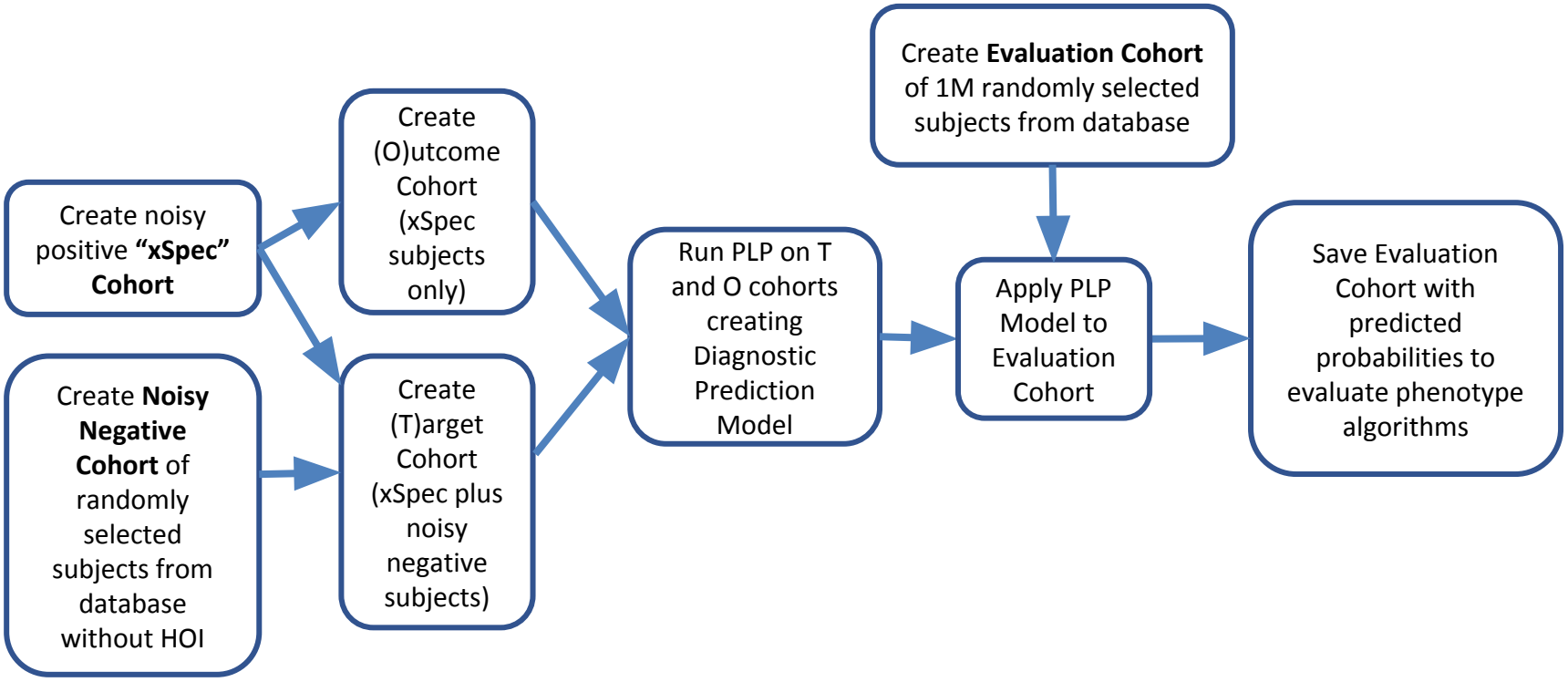
- It turns out that having a very good set of positives is good enough – a "noisy" model
- We use an "extremely specific" (xSpec) cohort





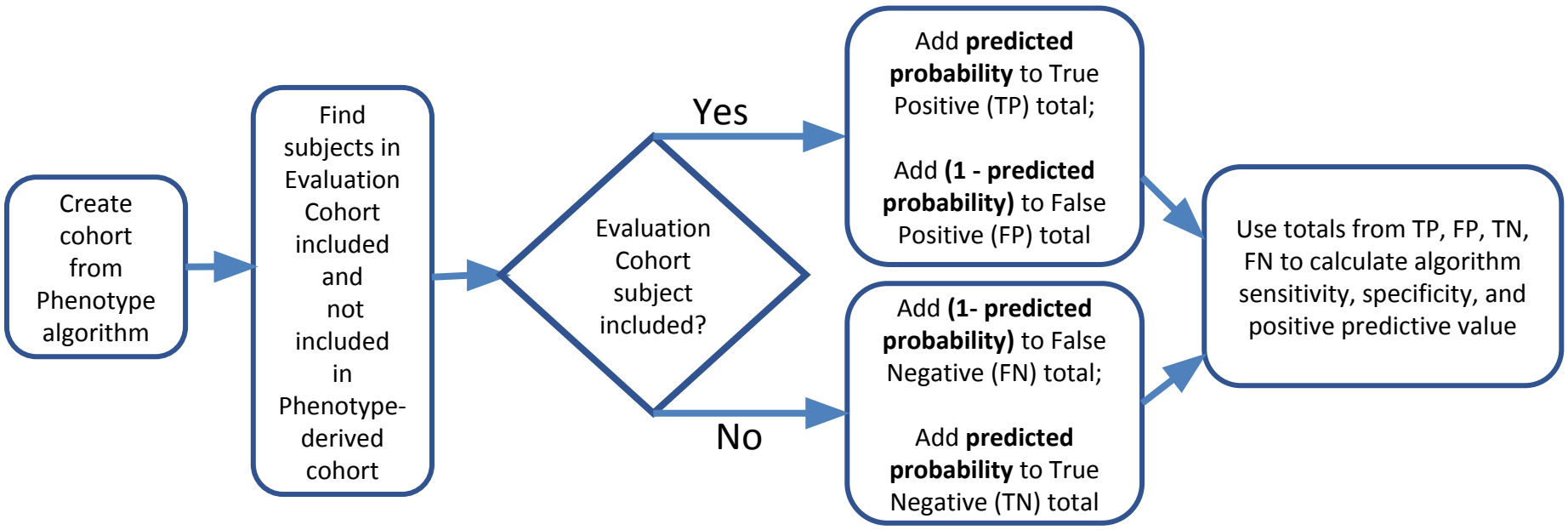
Running the Model

Step 1: Develop Evaluation Cohort from Diagnostic Predictive Model



Running the Model

Step 2: Evaluate Phenotype Algorithms



Example: Subject 1
Predicted Probability of HOI = 0.9 and **included** in Phenotype algorithm cohort:
0.9 added to TP
("90% correct case")
1 - 0.9 = 0.1 added to FP
("10% incorrect case")

Example: Subject 2
Predicted Probability of HOI = 0.1 and **not included** in Phenotype algorithm cohort:
1 - 0.1 = 0.9 added to TN
("90% correct non-case")
0.1 added to FN
("10% incorrect non-case")

Calculations:
Sensitivity = Sum TP / (Sum TP + Sum FN)
Specificity = Sum TN / (Sum TN + Sum FP)
PPV = Sum TP / (Sum TP + Sum FP)

Testing the Phenotypes

Typical Phenotypes for Health Outcomes of Interest (HOI):

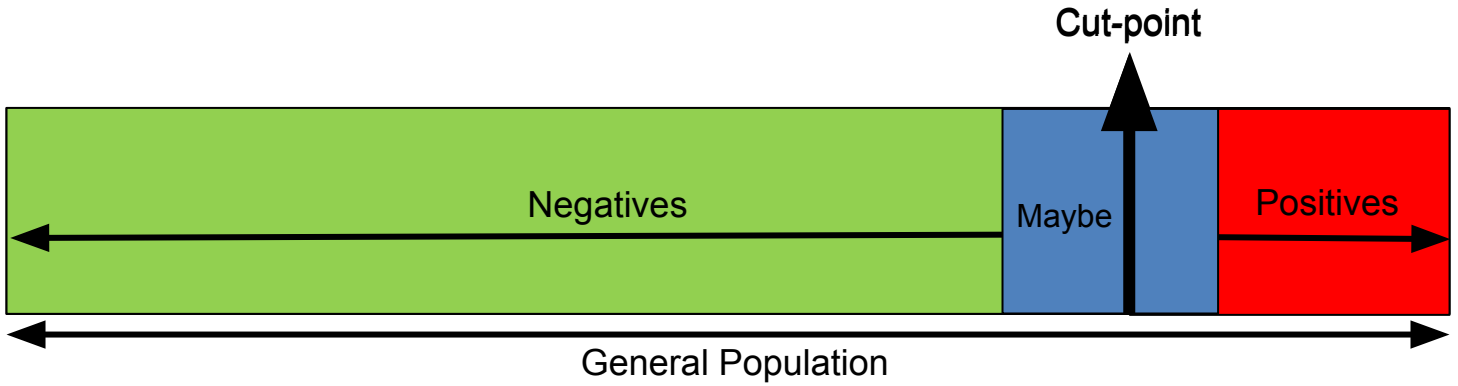
- 1 X HOI - e.g., Myocardial Infarction - SNOMED concept ID 22298006
- 2 X HOI - e.g., second MI diagnosis within 5 days of first MI diagnosis
- 1 X HOI, In-patient
- 1 X HOI, In-patient in first position

Comparing Results from Multiple Datasets

Phenotype Algorithm	CDM	Myocardial Infarction			Ischemic Stroke			Diabetes Mellitus			Atrial Fibrillation		
		Sens	PPV	Spec	Sens	PPV	Spec	Sens	PPV	Spec	Sens	PPV	Spec
>=1 x HOI	CCAE	99.9	83.5	99.9	99.9	56.8	99.6	99.9	75.6	96.8	99.9	70.0	99.6
	Optum	99.9	86.9	99.7	99.8	54.1	98.9	99.9	86.3	97.5	99.9	81.7	99.3
	MDCD	99.9	83.1	99.6	99.9	73.2	99.0	99.9	83.6	96.3	99.9	69.3	98.5
	MDCR	99.9	90.2	99.3	99.9	62.5	96.6	99.9	83.1	93.2	99.9	80.2	96.3
>=2 x HOI	CCAE	61.9	91.6	99.9	61.2	73.6	99.9	82.8	79.8	97.9	75.4	75.7	99.8
	Optum	53.3	90.7	99.9	61.1	66.7	99.6	81.8	87.9	98.2	77.7	83.4	99.5
	MDCD	54.3	88.6	99.9	61.8	81.5	99.6	83.6	86.1	97.5	74.8	74.9	99.2
	MDCR	53.2	92.1	99.7	61.4	74.9	98.8	84.9	84.7	94.9	80.2	83.0	97.5
>=1 x HOI - In-Patient	CCAE	78.7	89.6	99.9	66.2	73.3	99.9	32.3	87.4	99.5	47.9	90.6	99.9
	Optum	73.8	90.2	99.9	72.6	76.9	99.7	39.7	90.2	99.3	59.2	92.8	99.9
	MDCD	74.8	87.4	99.8	58.0	80.2	99.6	64.5	87.5	98.3	64.4	79.9	99.5
	MDCR	74.1	91.5	99.6	70.4	77.4	98.8	43.8	89.8	98.3	64.7	90.5	99.0
>=1 x HOI, IP - 1st Position	CCAE	71.9	90.2	99.9	62.3	76.6	99.9	7.4	90.9	99.9	37.7	95.1	99.9
	Optum	60.8	90.6	99.9	66.3	81.5	99.8	7.6	93.1	99.9	37.3	97.0	99.9
	MDCD	58.6	90.3	99.9	49.1	82.1	99.7	15.9	88.8	99.6	34.8	91.7	99.9
	MDCR	63.5	91.7	99.6	65.7	79.5	99.0	13.6	88.8	99.4	48.7	96.0	99.7

Limitations

- Sparse data for subjects
- Databases vary with overall level of detail



- Cutrona – 10% of patients with insufficient evidence
- Ryo – 7.5% of patients with insufficient evidence

Conclusion

- Having metrics for phenotype performance increases confidence in the use of observational data in research.
- Potential to use results of phenotype evaluation to correct/adjust our estimates.
- Using diagnostic predictive models to assess algorithm performance appears promising.

Questions

Student Workgroup

Advanced Atlas

Christopher Knoll

Cohort Applications

Christopher Knoll

Final Review