OHDSI Tutorial:

Common Data Model, Vocabulary, Population-Level Estimation, Patient-Level Prediction

Seng Chan You

- 행사 내용 : OMOP-CDM 기반 ATLAS (OHDSI tool) 의료 빅데이터 분석 도구의 사용법을 배우고 이를 이용한 실습을 통해 의료연구 계획 및 분석 수행
- 일　　시 : 2019.08.10(토) 09:40 ~ 08. 11(일) 13:40
- 장　　소 : 수원 아주대학교 홍재관 6층 603호 강의실

세부 일정

8월 10일 (토)		
09:00~09:40	참가자 등록	현장등록 불가
09:40~10:00	인사말 및 행사 소개	박래웅 교수/ 이성원 연구강사 (아주대학교 의료정보학과)
10:00~13:00	OMOP-CDM 및 ATALS 교육	
13:00~14:00	점심식사	점심 제공
14:00~15:00	팀 구성 및 연구주제 선정	
15:00~20:00	팀 분석 실습	간편식 제공
8월 11일 (일)		
09:40~10:00	행사 브리핑	
10:00~12:30	팀 분석 실습 (계속)	간편식 제공
12:30~13:30	팀 발표	
13:30~13:40	수료증 배부 및 사진 촬영	
13:40	종료	

Federated E-Health Big Data for Evidence Renovation Network
분산형 바이오헬스 빅데이터 사업단

산업통상자원부

아주대학교의료원
Ajou University Medical Center

Goal of the DataThon

- *행복한 가정은 모두 비슷한 이유로 행복하지만 불행한 가정은 저마다의 이유로 불행하다 (Все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива по-своему)*

- 구동이 되는 패키지는 모두 비슷한 이유로 작동하지만, 에러가 나는 패키지는 **모두 저마다의 문제**가 있다.

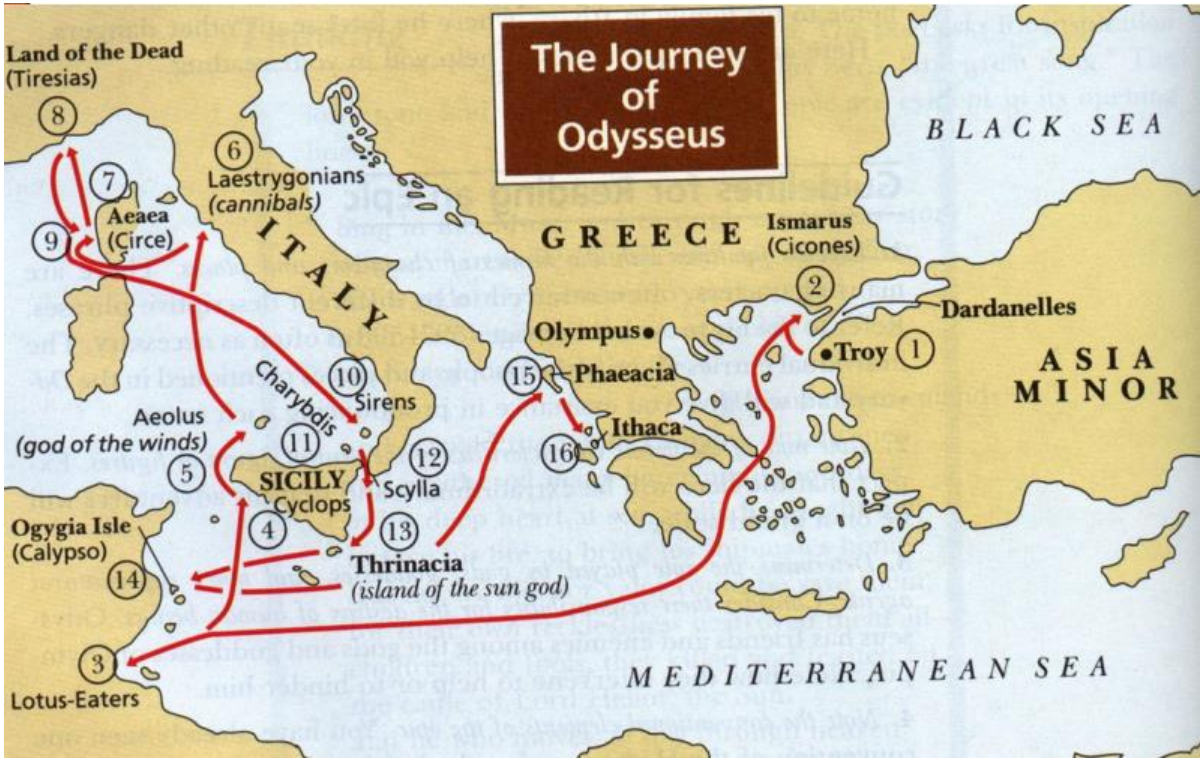
➔구동이 되는 PLE or PLP package를 완성하고, 결과를 발표, GitHub Upload

Three-day tutorial during an hour

- View video of full-day tutorial for **Common Data Model and Standardized Vocabularies**
  - https://www.ohdsi.org/past-events/2018-tutorials-omop-common-data-model-and-standardized-vocabularies/
- View video of full-day tutorial for **Population-Level Estimation**
  - https://www.ohdsi.org/past-events/population-level-estimation/
- View video of full-day tutorial for **Patient-Level Prediction**
  - https://www.ohdsi.org/past-events/patient-level-prediction/

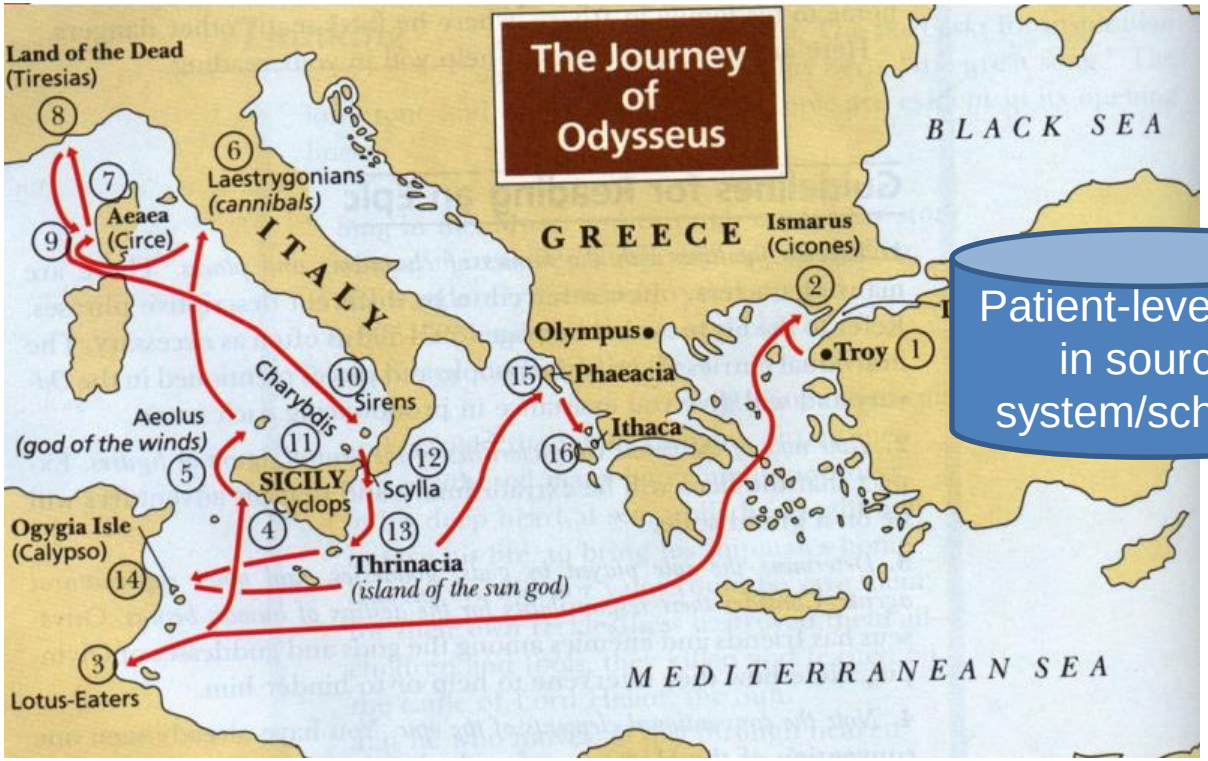
OHDSI (Observational Health Data Sciences and Informatics)

The odyssey to evidence generation



OHDSI (Observational Health Data Sciences and Informatics)

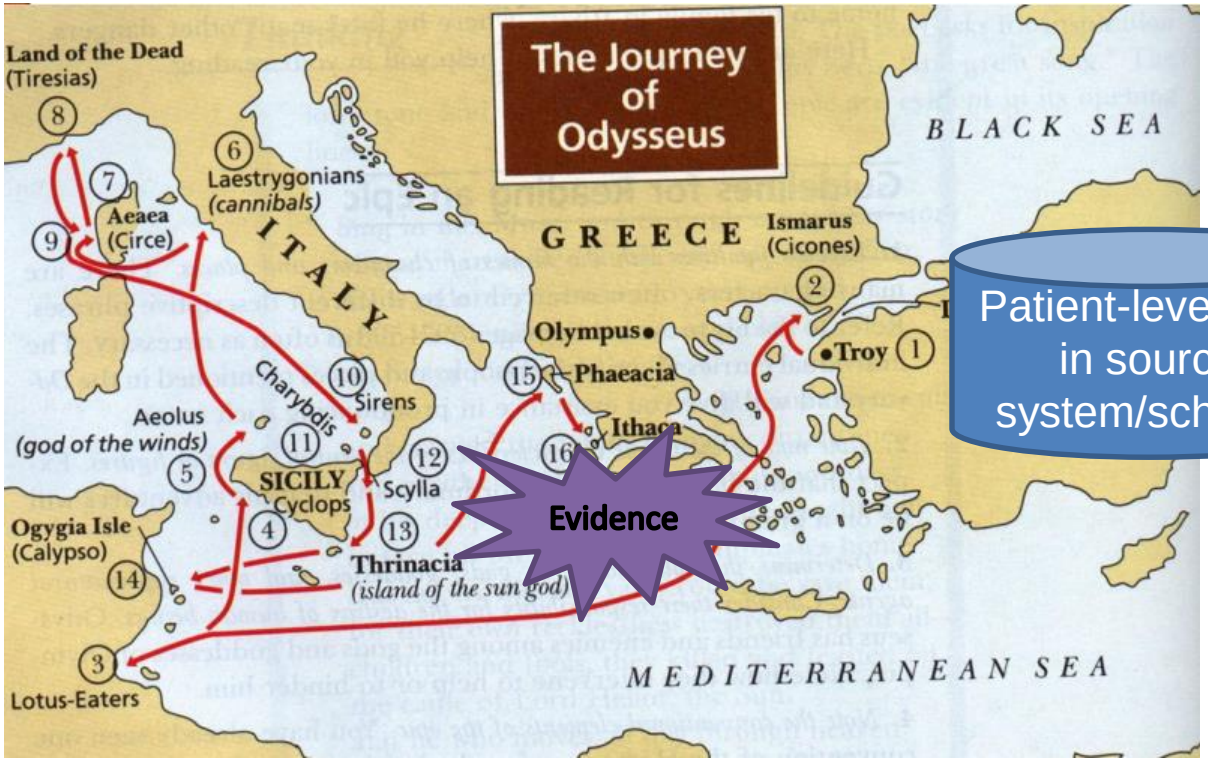
The odyssey to evidence generation



Patient-level data in source system/schema

OHDSI (Observational Health Data Sciences and Informatics)

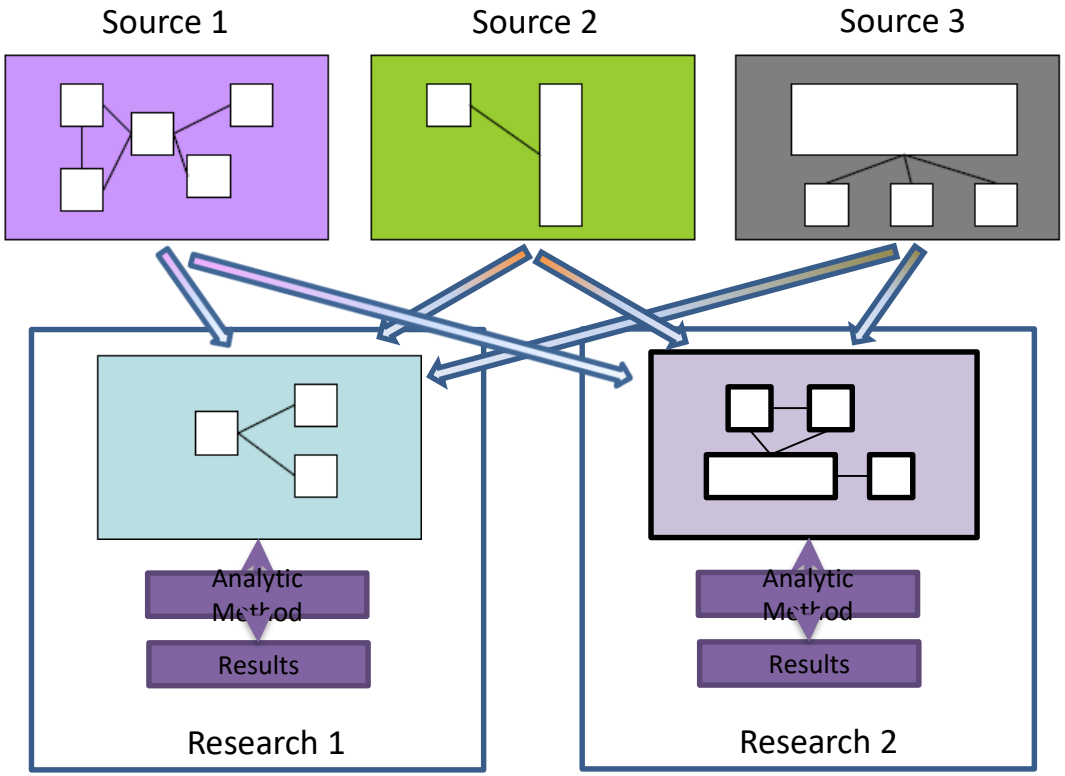
The odyssey to evidence generation



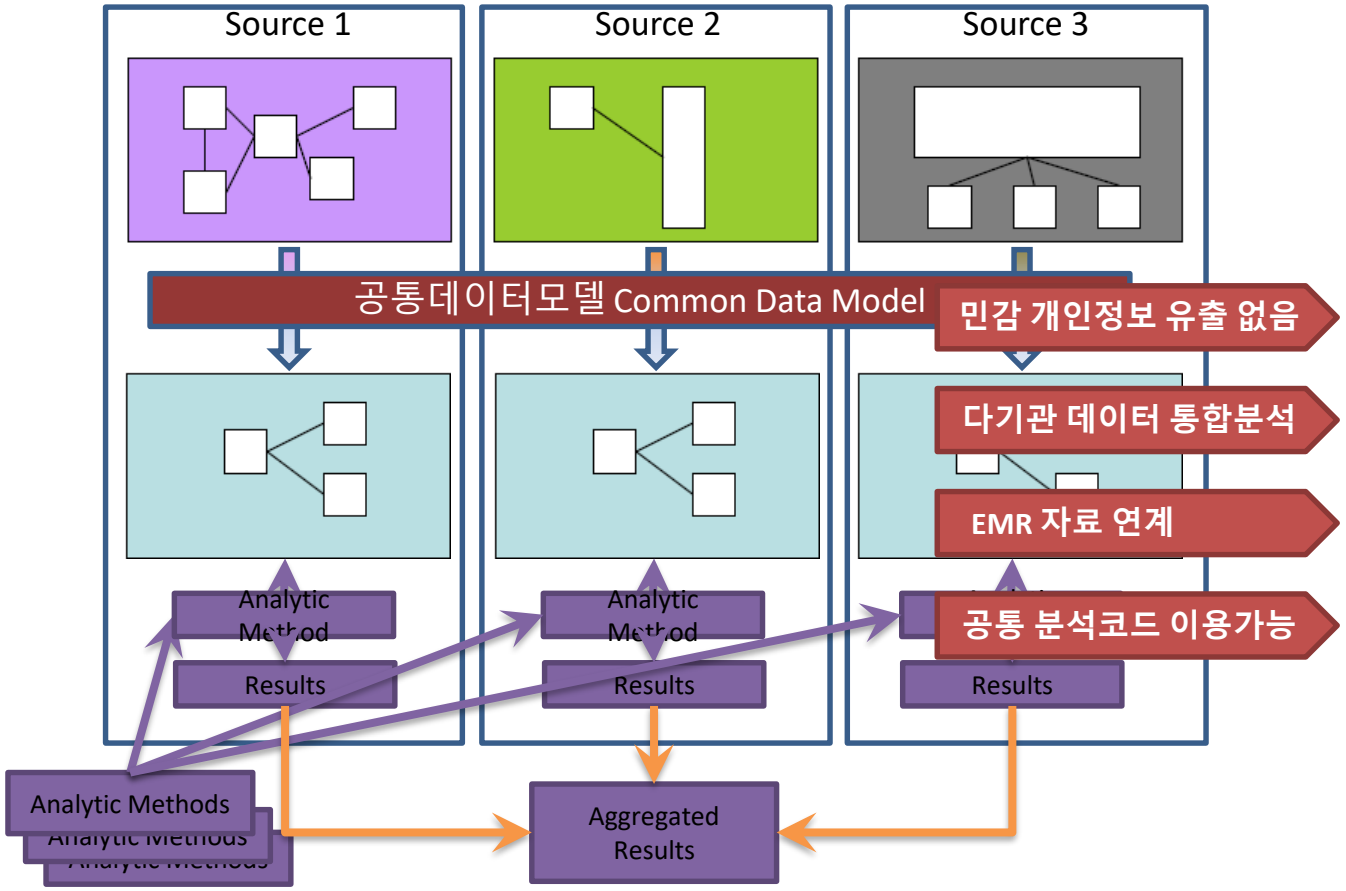
Why Common Data Model?

기존의 다기관 연구방법

연구 수행 때마다 데이터 모델을 맞추는 변환 작업을 수행해야 함



CDM in Distributed Research Network



OMOP 공통데이터모델

- OMOP (Observational Medical Outcomes Partnerships) CDM (common data model)
  - 2008년 관찰 데이터를 이용한 연구 기반 구축을 위해 시작된 공공-민간 협력 파트너십
  - OMOP에서 개발한 공통데이터모델은 계속 진화·발전하고 있음

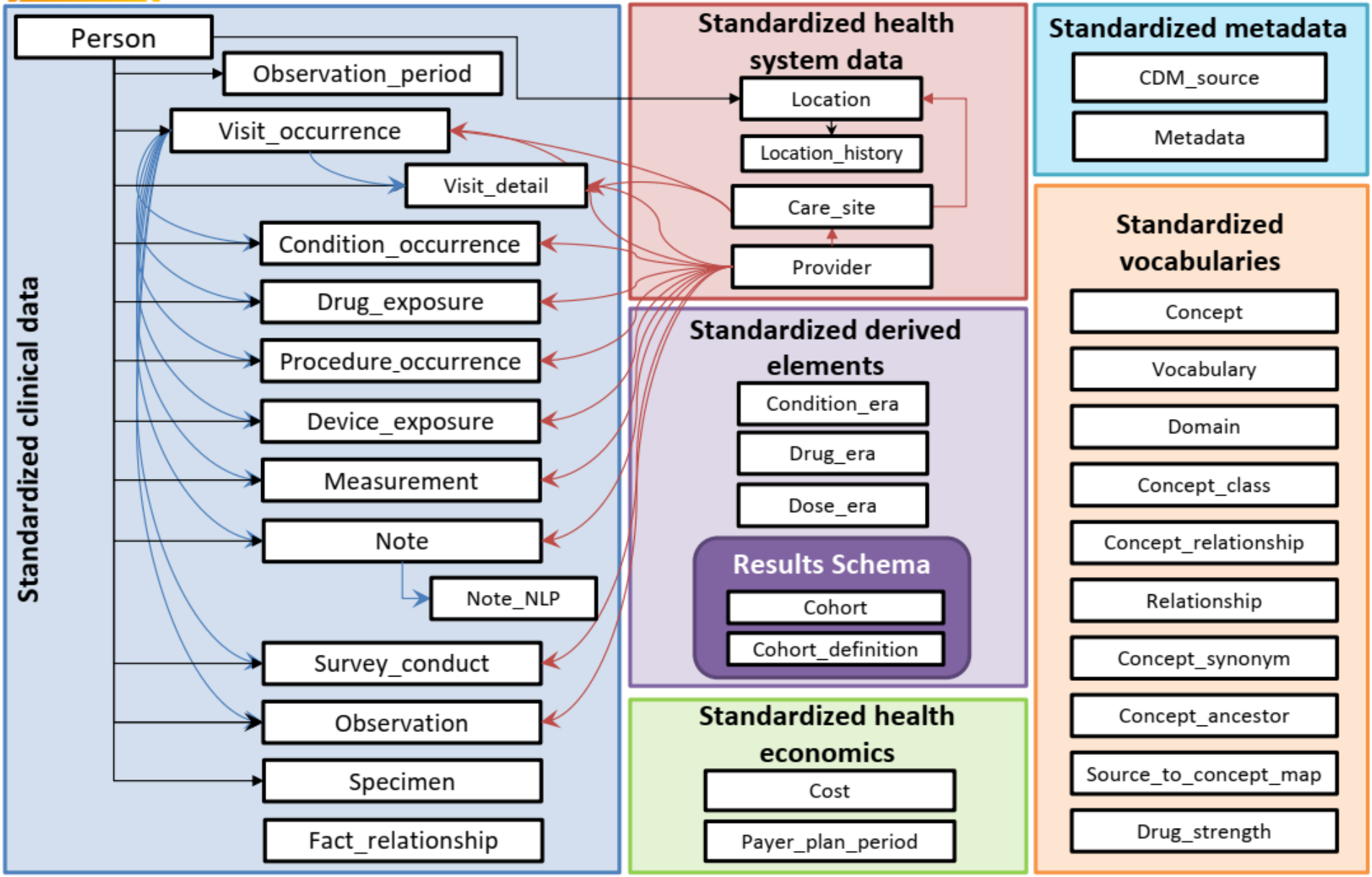
간결한 구조
- 24개의 변환 테이블로 구성

국제 표준 모델
- 전 세계적으로 동일한 구조와 의미
- 높은 재사용성
- 공동 연구 및 비교 연구 용이

표준 용어 이용
- SNOMED-CT, LOINC 등

오픈 정책 채택
- 다양한 오픈소스 툴 보유
- 많은 사용자 및 개발자
- 높은 창의성과 혁신성

CDM Version 6 Key Domains



OMOP CDM Principles

- Patient centric
- Vocabulary and Data Model are blended
- Domain-oriented concepts
- Accommodates data from various sources
- Preserves data provenance
- Extendable & Evolving
- Database Platform Independent

OMOP CDM Standard Domain Features

Feature	Description & Purpose	Field Name Convention	Example
Patient centric	Every domain table has **patient identifier**. Patient data can be retrieved independently from other domains.	**person_id**	person_id 123
Unique domain identifiers	Ever domain table has a unique primary key to identify domain entities.	<entity>**_id**	condition_occurrence_id 470985
Standard concept from a respective vocabulary domain	Integration with the Vocabulary. Foreign key into the Standard Vocabulary for **Standard Concept.**	<entity>_concept_id	condition_concept_id 313217 (SNOMED "Atrial Fibrillation")
Source value	Provenance. Verbatim information from the source data, **not to be used** by any standard analytics.	<entity>**_source_value**	condition_source_value 427.31 (ICD9CM "Atrial Fibrillation")
Source concept from a respective vocabulary domain	Provenance. Foreign key into Standard Vocabulary for **Source Concept**.	<entity>**_source_concept_id**	condition_source_concept_id 44821957 (ICD9CM "Atrial Fibrillation")
Source type	Provenance. Foreign key into Vocabulary for the **origin of the data.**	<entity>**_type_concept_id**	condition_type_concept_id 38000199 ("Inpatient header – primary")

OMOP Common Vocabulary Model

What it is

- **Standardized structure** to house existing vocabularies used in the public domain
- **Compiled standards** from disparate public and private sources and some OMOP-grown concepts

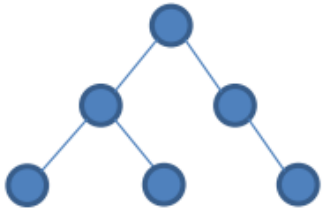
What it's not

- **Static dataset** – the vocabulary updates regularly to keep up with the continual evolution of the sources
- **Finished product** – vocabulary maintenance and improvement is ongoing activity that requires community participation and support

Structure of OMOP Vocabulary

All content: concepts in **`concept`**

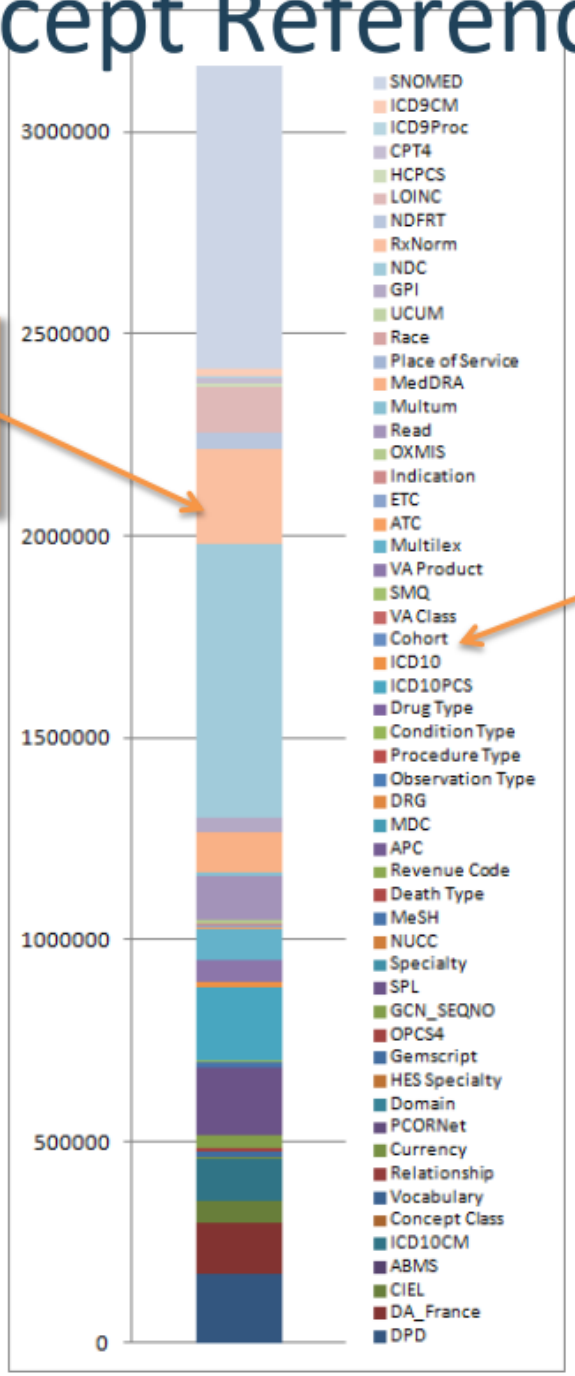
Direct relationships between concepts in **`concept_relationship`**



Multi-step hierarchical relationships pre-processed into **`concept_ancestor`**

Single Concept Reference Table

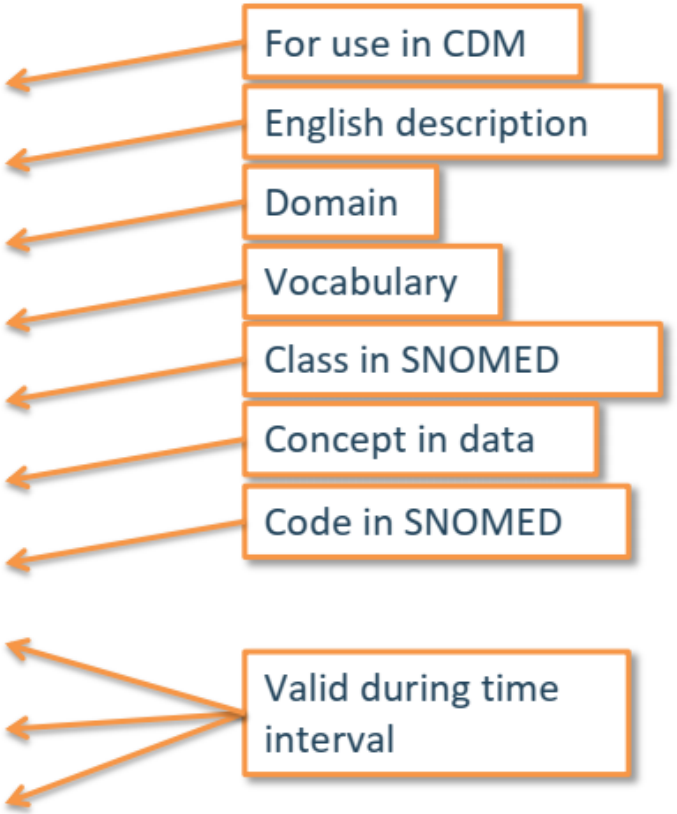
All vocabularies stacked up in one table



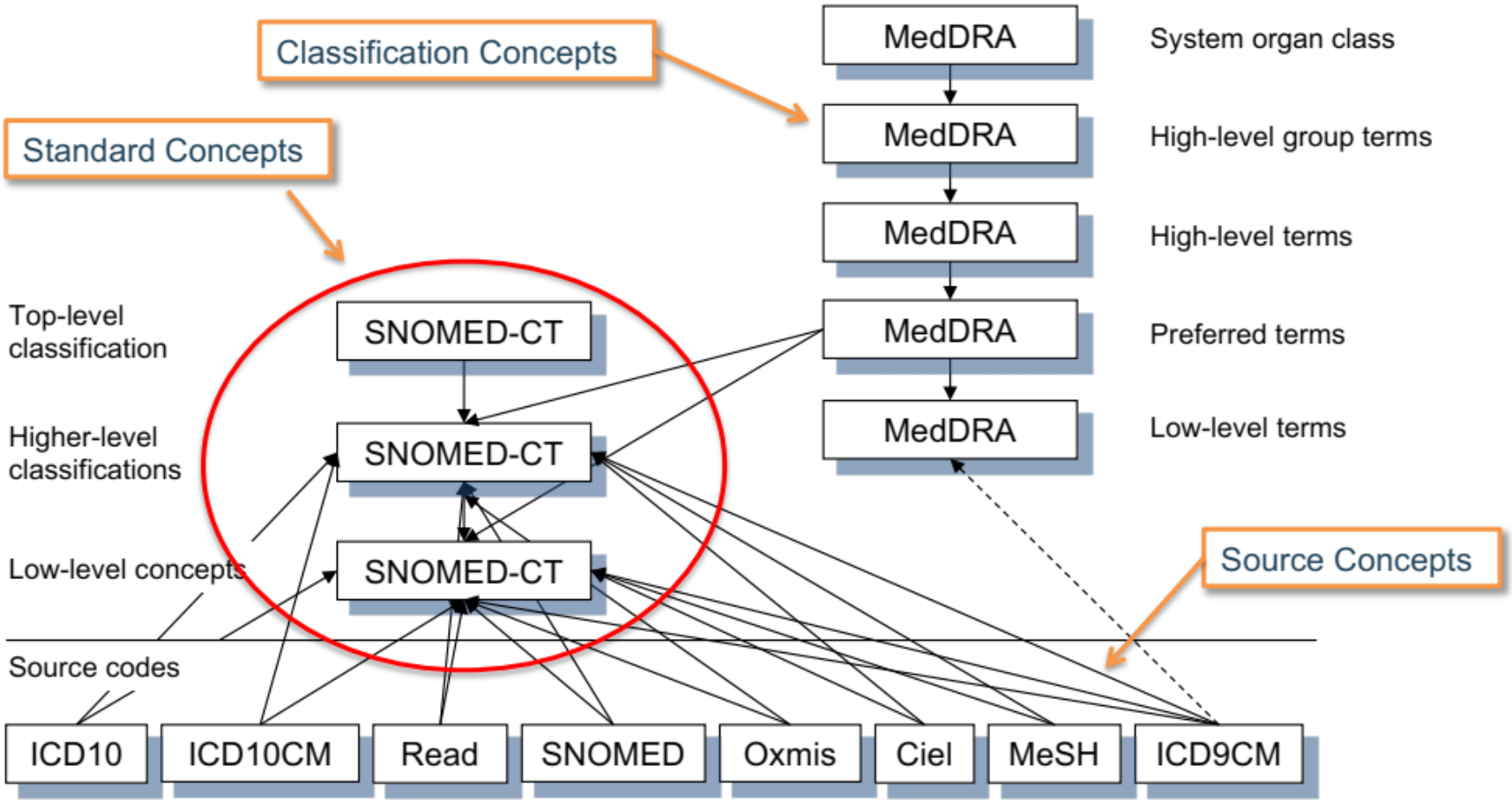
Vocabulary ID

What's in a Concept

CONCEPT_ID	313217
CONCEPT_NAME	Atrial fibrillation
DOMAIN_ID	Condition
VOCABULARY_ID	SNOMED
CONCEPT_CLASS_ID	Clinical Finding
STANDARD_CONCEPT	S
CONCEPT_CODE	49436004
VALID_START_DATE	01-Jan-1970
VALID_END_DATE	31-Dec-2099
INVALID_REASON	



Condition Concepts



Finding the Right Concept #1

1. ..if I know the **ID**

```
SELECT * FROM concept WHERE concept_id = 313217
```

CONCEPT_ID	CONCEPT_ NAME	DOMAIN_ID	VOCABULARY_ID	CONCEPT_CLASS_ID	STANDARD_CONCEPT	CONCEPT_CODE	VALID_START_DATE	VALID_END_DATE	INVALID_REASON
313217	Atrial fibrillation	Condition	SNOMED	Clinical Finding	S	49436004	01-Jan-1970	31-Dec-2099	

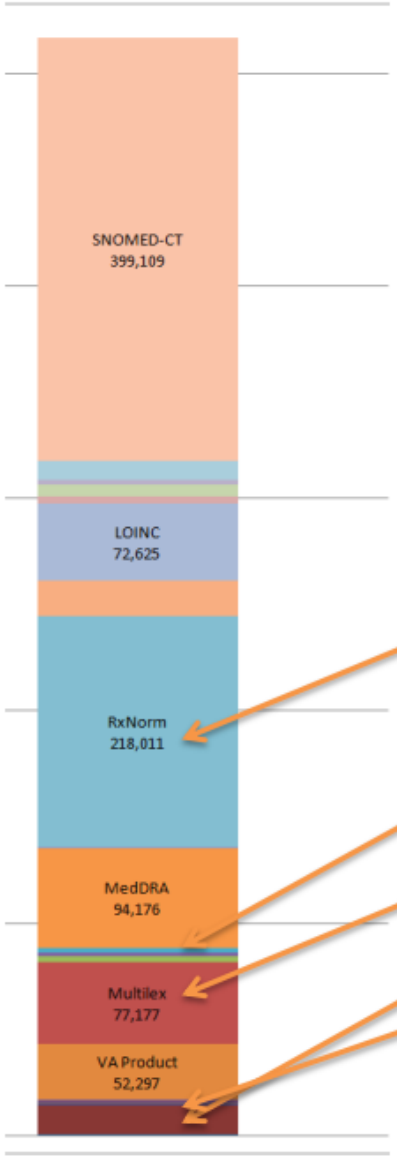
2. ..if I know the **code**

SNOMED code

```
SELECT * FROM concept WHERE concept_code = '49436004'
```

CONCEPT_ID	CONCEPT_ NAME	DOMAIN_ID	VOCABULARY_ID	CONCEPT_CLASS_ID	STANDARD_CONCEPT	CONCEPT_CODE	VALID_START_DATE	VALID_END_DATE	INVALID_REASON
313217	Atrial fibrillation	Condition	SNOMED	Clinical Finding	S	49436004	01-Jan-1970	31-Dec-2099	

Concept ID versus Concept Code




```
SELECT *
FROM concept
WHERE concept_code = '1001';
```

Same code

Concept_Name	Concept Class	Vocabulary_ID	Concept_Code
Antipyrine	Ingredient	RxNorm	1001
Acepromazine maleate	Ingredient	BDPM	1001
Serum	Specimen	CIEL	1001
methixene hydrochloride	Ingredient	Multilex	1001
Brompheniramine Maleate, 10 mg/mL injectable solution	Multum	Multum	1001
ABBOTT COLD SORE BALM 4%/0.06% W/	Drug Product	LPD_Australia	1001
Residential Treatment - Psychiatric	Revenue Code	Revenue Code	1001

Finding the Right Concept #2

3. ..if I know the **name**

```
SELECT * FROM concept WHERE concept_name = 'Atrial fibrillation';
```

CONCEPT _ID	CONCEPT_ NAME	DOMAIN _ID	VOCABULARY _ID	CONCEPT_ CLASS_ID	STANDARD_ CONCEPT	CONCEPT_ CODE
313217	Atrial fibrillation	Condition	SNOMED	Clinical Finding	S	49436004
44821957	Atrial fibrillation	Condition	ICD9CM	5-dig billing code		427.31
35204953	Atrial fibrillation	Condition	MedDRA	PT	C	10003658
45500085	Atrial fibrillation	Condition	Read	Read		G573000
45883018	Atrial fibrillation	Meas Value	LOINC	Answer	S	LA17084-7

Finding the Right Concept #3

1. if don't know any of this, but I know the code in another vocabulary

ICD-9 is not a Standard Concept

```
SELECT * FROM concept WHERE concept_code = '427.31';
```

CONCEPT _ID	CONCEPT_ NAME	DOMAIN _ID	VOCABULARY _ID	CONCEPT_ CLASS_ID	STANDARD_ CONCEPT	CONCEPT_ CODE
44821957	Atrial fibrillation	Condition	ICD9CM	5-dig billing code		427.31

```
SELECT * FROM concept_relationship WHERE concept_id_1 = 44821957;
```

Mapping to different vocabularies

Kind of relationship

_ID_1	CONCEPT ID_2	RELATIONSHIP _ID	VALID_START _DATE	VALID_END _DATE	INVALID _REASON
44821957	21001551	ICD9CM - FDB Ind	01-Oct-13	31-Dec-2099	
44821957	35204953	ICD9CM - MedDRA	01-Jan-70	31-Dec-2099	
44821957	44824248	Is a	01-Oct-14	31-Dec-2099	
44821957	44834731	Is a	01-Oct-14	31-Dec-2099	
44821957	313217	Maps to	01-Jan-70	31-Dec-2099	

Mapping = Translating

Step 1. Lookup the Source Concept

```
SELECT * FROM concept WHERE concept_code = '427.31';
```

CONCEPT _ID	CONCEPT_ NAME	DOMAIN _ID	VOCABULARY _ID	CONCEPT_ CLASS_ID	STANDARD_ CONCEPT	CONCEPT_ CODE
44821957	Atrial fibrillation	Condition	ICD9CM	5-dig billing code		427.31

Step 2. Translate to Standard

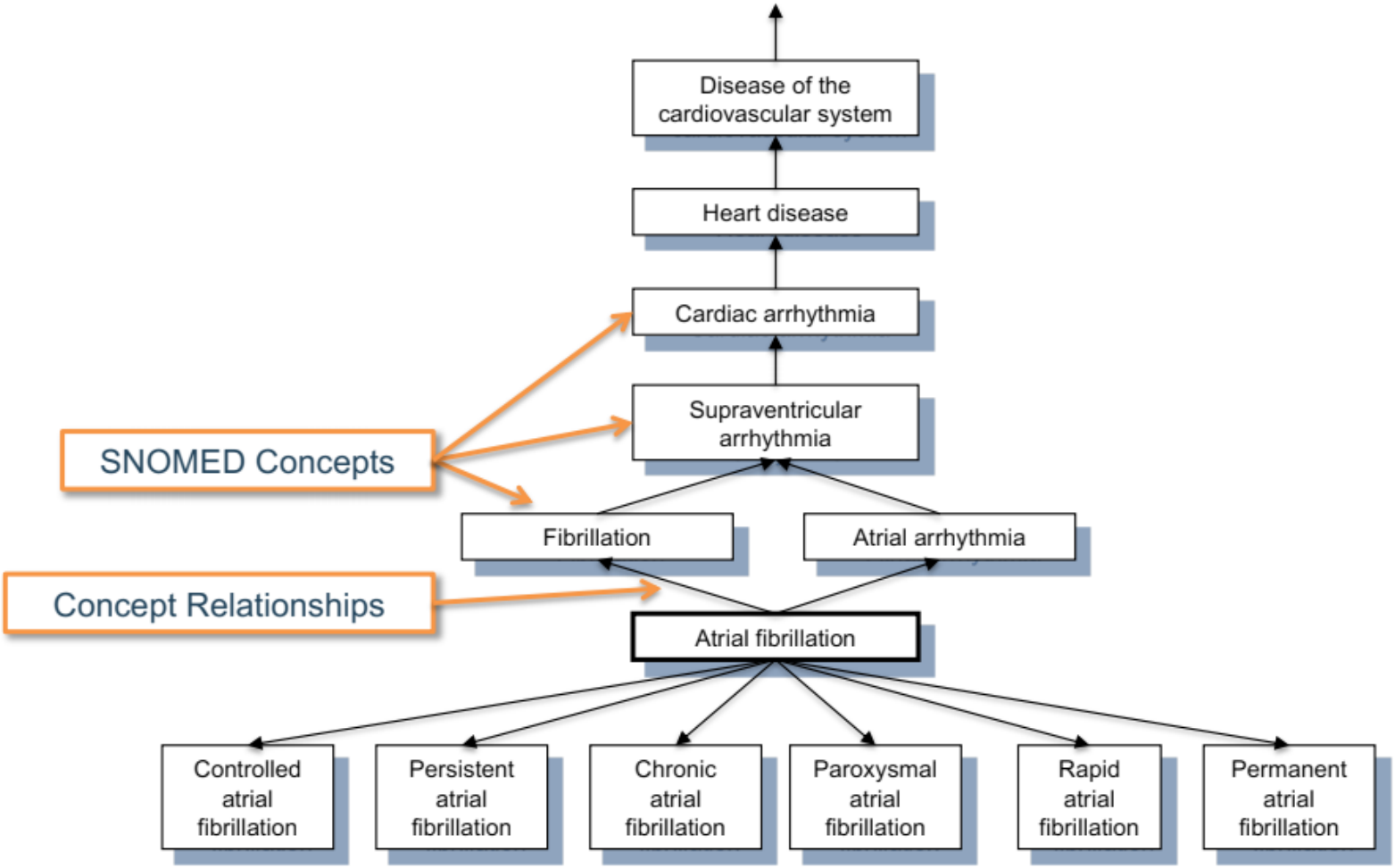
```
SELECT * FROM concept_relationship WHERE concept_id_1 = 44821957
    AND relationship_id = 'Maps to';
```

CONCEPT _ID_1	CONCEPT _ID_2	RELATIONSHIP _ID	VALID_START _DATE	VALID_END _DATE	INVALID _REASON
44821957	313217	Maps to	01-Jan-1970	31-Dec-2099	

Step 3. Check out the translated Concept

```
SELECT * FROM concept WHERE concept_id = 313217;
```

Reason #2: Disease Hierarchy



Exploring Relationships

```
SELECT *
FROM concept_relationship
WHERE concept_id_1 = 313217
```

CONCEPT ID_1	CONCEPT ID_2	RELATIONSHIP _ID
313217	4232697	Subsumes
313217	4181800	Focus of
313217	35204953	SNOMED - MedDRA eq
313217	4203375	Asso finding of
313217	4141360	Subsumes
313217	4119601	Subsumes
313217	4117112	Subsumes
313217	4232691	Subsumes
313217	4139517	Due to of
313217	4194288	Asso finding of
313217	44782442	Subsumes
313217	44783731	Focus of
313217	21003018	SNOMED - ind/CI
313217	40248987	SNOMED - ind/CI
313217	21001551	SNOMED - ind/CI
313217	21001540	SNOMED - ind/CI
313217	45576876	Mapped from
313217	44807374	Asso finding of
313217	21013834	SNOMED - ind/CI
313217	21001572	SNOMED - ind/CI
313217	21001606	SNOMED - ind/CI
313217	21003176	SNOMED - ind/CI
313217	4226399	Is a
313217	500001801	SNOMED - HOI
313217	500002401	SNOMED - HOI
313217	4119602	Subsumes
313217	40631039	Subsumes
313217	4108832	Subsumes
313217	21013671	SNOMED - ind/CI
313217	21013390	SNOMED - ind/CI
313217	313217	Maps to
313217	44821957	Mapped from
313217	2617597	Mapped from
313217	45500085	Mapped from
313217	313217	Mapped from
313217	45951191	Mapped from
313217	21013856	SNOMED - ind/CI
313217	21001575	SNOMED - ind/CI
313217	21001594	SNOMED - ind/CI

Related Concepts

Relationship ID

Exploring Relationships

```
SELECT cr.relationship_id, c.*
FROM concept_relationship cr
JOIN concept c ON cr.concept_id_2 = c.concept_id
WHERE cr.concept_id_1 = 313217
```

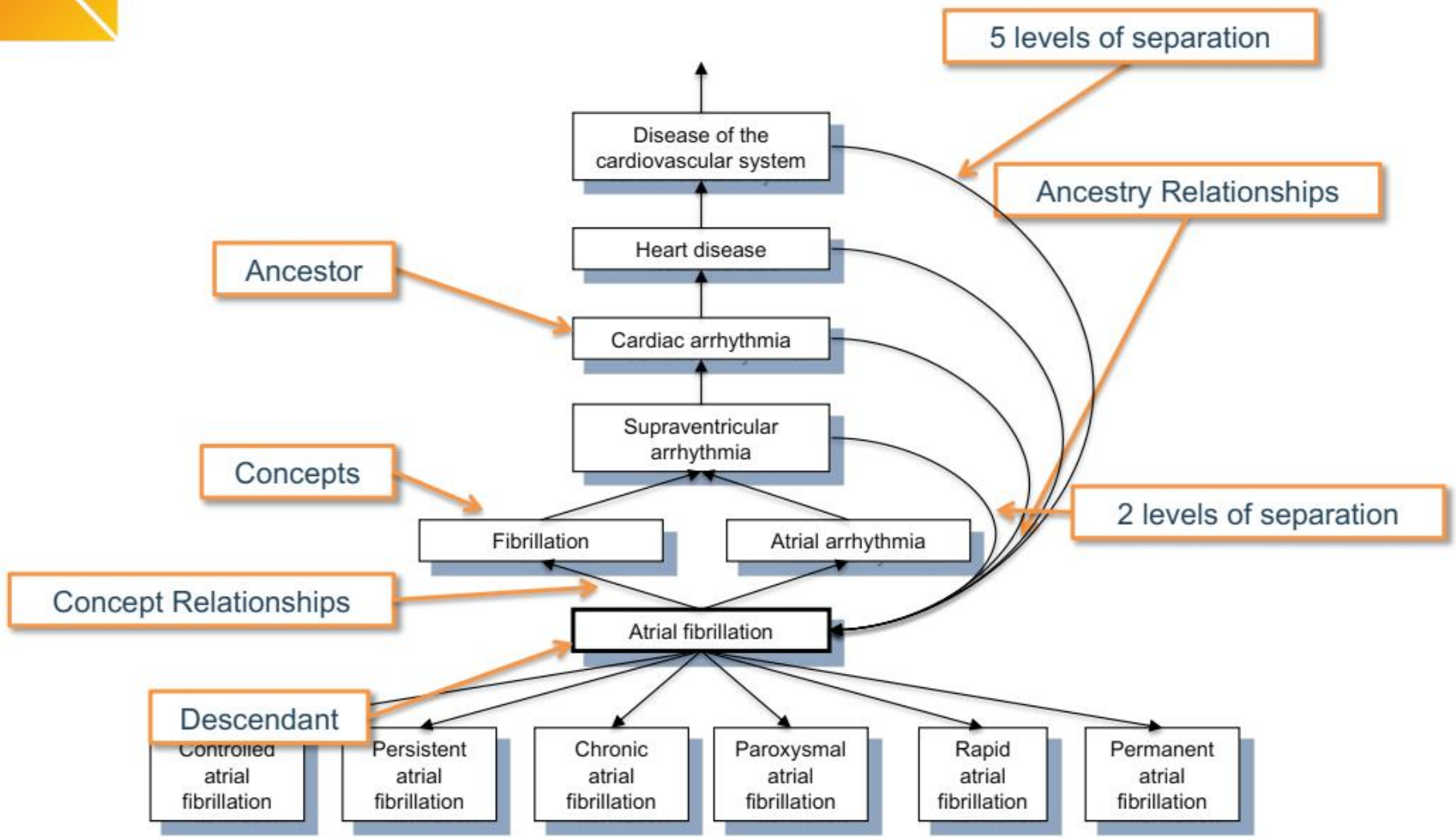
Find out related concept

Ancestor concepts

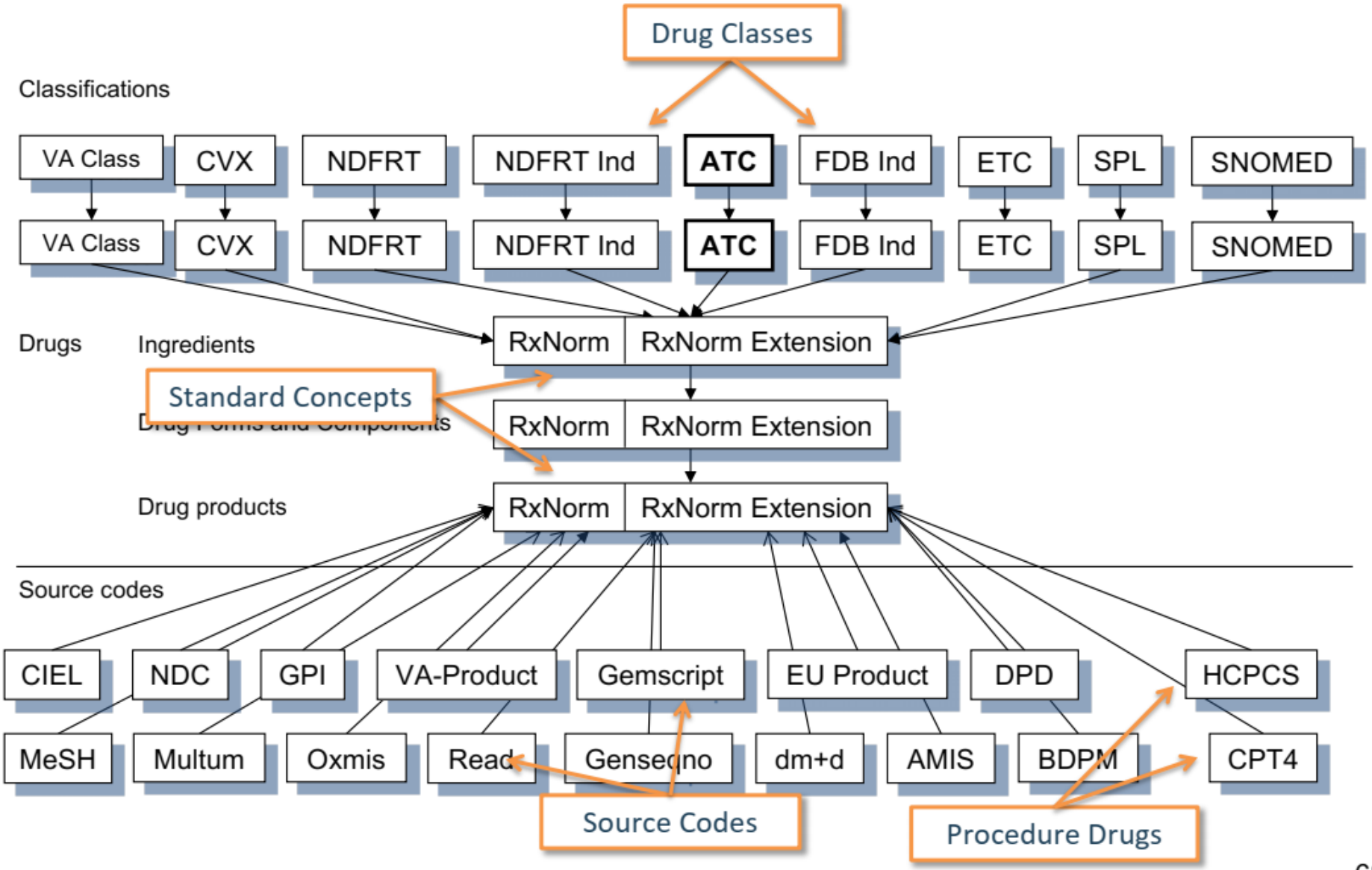
relationship_id	concept_id	concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code	valid_start_date	valid_end_date	invalid_reason
Asso finding of	4194288		Observation	SNOMED	Context-dependent	S	312442005	1/1/1970 0:00	12/31/2099 0:00	NULL
Asso finding of	4203375		Observation	SNOMED	Context-dependent	S	433276002	1/31/2009 0:00	12/31/2099 0:00	NULL
Asso finding of	42689685	Atrial [illegible] not [illegible]ected	Observation	SNOMED	Context-dependent	S	1.06706E+15	4/1/2017 0:00	12/31/2099 0:00	NULL
Asso finding of	44807374	Atrial fibrillation excluded	Observation	SNOMED	Context-dependent	S	8.16401E+14	4/1/2014 0:00	12/31/2099 0:00	NULL
Concept poss_eq from	40323929	Fibrillation - atrial	Condition	SNOMED	Clinical Finding	NULL	155364009	1/1/1970 0:00	3/11/2016 0:00	U
Concept poss_eq from	40345197	Fibrillation - atrial	Condition	SNOMED	Clinical Finding	NULL	266306001	1/1/1970 0:00	3/11/2016 0:00	U
Due to of	4139517	Transient cerebral ischemia due to atrial fibrillation	Condition	SNOMED	Clinical Finding	S	426814001	1/1/1970 0:00	12/31/2099 0:00	NULL
Focus of	4209991	Insertion of pacemaker for control of atrial fibrillation	Procedure	SNOMED	Procedure	S	449863006	1/31/2012 0:00	12/31/2099 0:00	NULL
Has finding site	4242112	Atrial structure	Spec Anatomic Site	SNOMED	Body Structure	S	59652004	1/1/1970 0:00	12/31/2099 0:00	NULL
Is a	4226399	Fibrillation	Condition	SNOMED	Clinical Finding	S	40593004	1/1/1970 0:00	12/31/2099 0:00	NULL
Is a	4068155	Atrial arrhythmia	Condition	SNOMED	Clinical Finding	S	17366009	1/1/1970 0:00	12/31/2099 0:00	NULL
Mapped from	40323929	Fibrillation - atrial	Condition	SNOMED	Clinical Finding	NULL	155364009	1/1/1970 0:00	3/11/2016 0:00	U
Mapped from	2617597	Patient with heart failure and atrial fibrillation documented to be on warfarin therapy	Observation	HCPCS	HCPCS	NULL	G8183	1/1/1970 0:00	11/11/2014 0:00	D
Mapped from	45576876	Unspecified atrial fibrillation	Condition	ICD10CM	5-char billing code	NULL	I48.91	12/30/2006 0:00	12/31/2099 0:00	NULL
Mapped from	45500085	Atrial fibrillation	Condition	Read	Read	NULL	G573000	1/1/1970 0:00	12/31/2099 0:00	NULL
Mapped from	45611600	Atrial Fibrillation	Condition	MeSH	Main Heading	NULL	D001281	1/1/1970 0:00	12/31/2099 0:00	NULL
Mapped from	40345197	Fibrillation - atrial	Condition	SNOMED	Clinical Finding	NULL	266306001	1/1/1970 0:00	3/11/2016 0:00	U
Mapped from	45951191	Atrial Fibrillation	Condition	CIEL	Diagnosis	NULL	148203	11/3/2007 0:00	12/31/2099 0:00	NULL
Mapped from	313217	Atrial fibrillation	Condition	SNOMED	Clinical Finding	S	49436004	1/1/1970 0:00	12/31/2099 0:00	NULL
Mapped from	44821957	Atrial fibrillation	Condition	ICD9CM	5-dig billing code	NULL	427.31	1/1/1970 0:00	12/31/2099 0:00	NULL
Maps to	313217	Atrial fibrillation	Condition	SNOMED	Clinical Finding	S	49436004	1/1/1970 0:00	12/31/2099 0:00	NULL
SNOMED - HOI	500002401	OMOP Atrial Fibrillation 1	Condition	Cohort	Cohort	C	500002401	1/1/1970 0:00	12/31/2099 0:00	NULL
SNOMED - HOI	500001801	OMOP Qt Prolongation/Torsade De Pointes 1	Condition	Cohort	Cohort	C	500001801	1/1/1970 0:00	12/31/2099 0:00	NULL
SNOMED - ind/CI	21005673	Prevention of Thromboembolism in Chronic Atrial Fibrillation	Drug	Indication	Indication	C	5673	1/1/1970 0:00	12/31/2099 0:00	NULL
SNOMED - ind/CI	21003176	Tachyarrhythmia	Drug	Indication	Indication	C	3176	1/1/1970 0:00	12/31/2099 0:00	NULL
SNOMED - ind/CI	21001542	Supraventricular Tachycardia	Drug	Indication	Indication	C	1542	1/1/1970 0:00	12/31/2099 0:00	NULL
SNOMED - ind/CI	21001594	Disease of Cardiovascular System	Drug	Indication	Indication	C	1594	1/1/1970 0:00	12/31/2099 0:00	NULL
SNOMED - MedDRA eq	35204953	Atrial fibrillation	Condition	MedDRA	PT	C	10003658	1/1/1970 0:00	12/31/2099 0:00	NULL
Subsumes	4117112	Controlled atrial fibrillation	Condition	SNOMED	Clinical Finding	S	300996004	1/1/1970 0:00	12/31/2099 0:00	NULL
Subsumes	4119601	Lone atrial fibrillation	Condition	SNOMED	Clinical Finding	S	233910005	1/1/1970 0:00	12/31/2099 0:00	NULL
Subsumes	4232697	Persistent atrial fibrillation	Condition	SNOMED	Clinical Finding	S	440059007	1/31/2009 0:00	12/31/2099 0:00	NULL
Subsumes	4141360	Chronic atrial fibrillation	Condition	SNOMED	Clinical Finding	S	426749004	1/1/1970 0:00	12/31/2099 0:00	NULL
Subsumes	44782442	Atrial fibrillation with rapid ventricular response	Condition	SNOMED	Clinical Finding	S	1.20041E+14	1/31/2014 0:00	12/31/2099 0:00	NULL
Subsumes	4199501	Rapid atrial fibrillation	Condition	SNOMED	Clinical Finding	S	314208002	1/1/1970 0:00	12/31/2099 0:00	NULL
Subsumes	4119602	Non-rheumatic atrial fibrillation	Condition	SNOMED	Clinical Finding	S	233911009	1/1/1970 0:00	12/31/2099 0:00	NULL

Descendant concepts

Ancestry Relationships: Higher-Level Relationships



Drug Hierarchy



Person

PERSON

field	required	type
person_id	Yes	INTEGER
gender_concept_id	Yes	INTEGER
year_of_birth	Yes	INTEGER
month_of_birth	No	INTEGER
day_of_birth	No	INTEGER
birth_datetime	No	DATETIME
death_datetime	No	DATETIME
race_concept_id	Yes	INTEGER
ethnicity_concept_id	Yes	INTEGER
location_id	No	INTEGER
provider_id	No	INTEGER
care_site_id	No	INTEGER
person_source_value	No	VARCHAR(50)
gender_source_value	No	VARCHAR(50)
gender_source_concept_id	Yes	INTEGER
race_source_value	No	VARCHAR(50)
race_source_concept_id	Yes	INTEGER
ethnicity_source_value	No	VARCHAR(50)
ethnicity_source_concept_id	Yes	INTEGER

- 환자 기본정보
- Person 당 한 개 record
- 생년월일 중 년도(year_of_birth)는 필수
- death_datetime: v6.0에서 적용

OBSERVATION PERIOD

OBSERVATION_PERIOD

field	required	type
observation_period_id	Yes	INTEGER
person_id	Yes	INTEGER
observation_period_start_date	Yes	DATE
observation_period_end_date	Yes	DATE
period_type_concept_id	Yes	INTEGER

- DB상 환자 관찰 기간 (데이터 보유 기간)
- EMR 데이터의 경우, person 당 1 개 record 생성 (첫 기록 ~ 마지막 기록)
- Observation period 내에 있는 데이터만 분석에 이용

VISIT OCCURRENCE

VISIT_OCCURRENCE

field	required	type
visit_occurrence_id	Yes	INTEGER
person_id	Yes	INTEGER
visit_concept_id	Yes	INTEGER
visit_start_date	No	DATE
visit_start_datetime	Yes	DATETIME
visit_end_date	No	DATE
visit_end_datetime	Yes	DATETIME
visit_type_concept_id	Yes	INTEGER
provider_id	No	INTEGER
care_site_id	No	INTEGER
visit_source_value	No	VARCHAR(50)
visit_source_concept_id	Yes	INTEGER
admitting_source_concept_id	Yes	INTEGER
admitting_source_value	No	VARCHAR(50)
discharge_to_concept_id	Yes	INTEGER
discharge_to_source_value	No	VARCHAR(50)
preceding_visit_occurrence_id	No	INTEGER

- 외래/입원/응급 등 수진정보
- 방문 유형(visit_concept_id)
  - 9201: Inpatient Visit(입원)
  - 9202 : Outpatient Visit(외래)
  - 9203: Emergency Room Visit(응급)
  - ....

CONDITION OCCURRENCE

CONDITION_OCCURRENCE

field	required	type
condition_occurrence_id	Yes	BIGINT
person_id	Yes	BIGINT
condition_concept_id	Yes	INTEGER
condition_start_date	No	DATE
condition_start_datetime	Yes	DATETIME
condition_end_date	No	DATE
condition_end_datetime	No	DATETIME
condition_type_concept_id	Yes	INTEGER
condition_status_concept_id	Yes	INTEGER
stop_reason	No	VARCHAR(20)
provider_id	No	INTEGER
visit_occurrence_id	No	INTEGER
visit_detail_id	No	INTEGER
condition_source_value	No	VARCHAR(50)
condition_source_concept_id	Yes	INTEGER
condition_status_source_value	No	VARCHAR(50)

- 진단(주진단/부진단), 주호소 등
- 사망진단은 v6.0에서 적용
- 표준용어: SNOMED CT, ICD-O

DRUG EXPOSURE

DRUG_EXPOSURE

field	required	type
drug_exposure_id	Yes	BIGINT
person_id	Yes	BIGINT
drug_concept_id	Yes	INTEGER
drug_exposure_start_date	No	DATE
drug_exposure_start_datetime	Yes	DATETIME
drug_exposure_end_date	No	DATE
drug_exposure_end_datetime	No	DATETIME
verbatim_end_date	No	DATE
drug_type_concept_id	Yes	INTEGER
stop_reason	No	VARCHAR(20)
refills	No	INTEGER
quantity	No	FLOAT
days_supply	No	INTEGER
sig	No	VARCHAR(MAX)
route_concept_id	Yes	INTEGER
lot_number	No	VARCHAR(50)
provider_id	No	INTEGER
visit_occurrence_id	No	INTEGER
visit_detail_id	No	INTEGER
drug_source_value	No	VARCHAR(50)
drug_source_concept_id	Yes	INTEGER
route_source_value	No	VARCHAR(50)
dose_unit_source_value	No	VARCHAR(50)

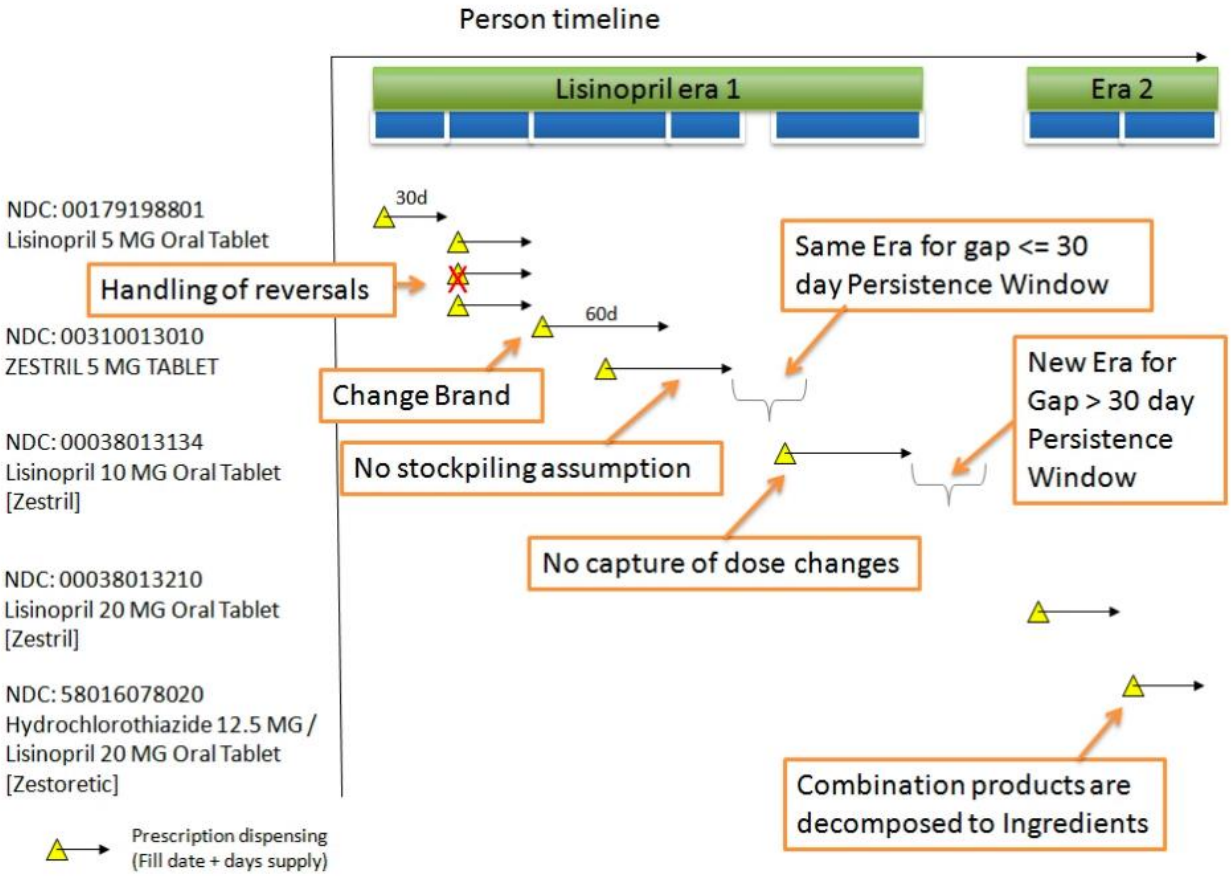
- 약물 처방 및 투약 기록
- 약품명, 처방일, 총수량, 투여경로

  (경구, 주사, 외용)
- 표준용어: RxNorm, RxNorm

  Extension

DRUG ERA

DRUG_ERA



분당서울대 유수영 교수님 Slide

DRUG ERA

DRUG_ERA

field	required	type
drug_era_id	Yes	INTEGER
person_id	Yes	INTEGER
drug_concept_id	Yes	INTEGER
drug_era_start_datetime	Yes	DATE
drug_era_end_datetime	Yes	DATE
drug_exposure_count	No	INTEGER
gap_days	No	INTEGER

- Ingredient level의 연속적인 약물 노출 기간 추론 데이터
- DRUG_EXPOSURE 테이블로부터 표준화된 로직을 적용하여 자동 계산(예, 30일 persistence window 허용 등)

PROCEDURE OCCURRENCE

PROCEDURE_OCCURRENCE

field	required	type
procedure_occurrence_id	Yes	INTEGER
person_id	Yes	INTEGER
procedure_concept_id	Yes	INTEGER
procedure_date	No	DATE
procedure_datetime	Yes	DATETIME
procedure_type_concept_id	Yes	INTEGER
modifier_concept_id	Yes	INTEGER
quantity	No	INTEGER
provider_id	No	INTEGER
visit_occurrence_id	No	INTEGER
visit_detail_id	No	INTEGER
procedure_source_value	No	VARCHAR(50)
procedure_source_concept_id	Yes	INTEGER
modifier_source_value	No	VARCHAR(50)

- 수술, 처치, 검사처방 등
- 표준용어: SNOMED CT
- 표준 용어 매핑 이슈 존재

MEASUREMENT

MEASUREMENT

field	required	type
measurement_id	Yes	INTEGER
person_id	Yes	INTEGER
measurement_concept_id	Yes	INTEGER
measurement_date	No	DATE
measurement_datetime	Yes	DATETIME
measurement_time	No	VARCHAR(10)
measurement_type_concept_id	Yes	INTEGER
operator_concept_id	No	INTEGER
value_as_number	No	FLOAT
value_as_concept_id	No	INTEGER
unit_concept_id	No	INTEGER
range_low	No	FLOAT
range_high	No	FLOAT
provider_id	No	INTEGER
visit_occurrence_id	No	INTEGER
visit_detail_id	No	INTEGER
measurement_source_value	No	VARCHAR(50)
measurement_source_concept_id	Yes	INTEGER
unit_source_value	No	VARCHAR(50)
value_source_value	No	VARCHAR(50)

- 수치형/범주형의 구조화된 검사결과, 임상관찰기록, 병리보고서 결과 등
- 표준용어: LOINC, SNOMED CT
- 표준 용어 매핑 이슈 존재

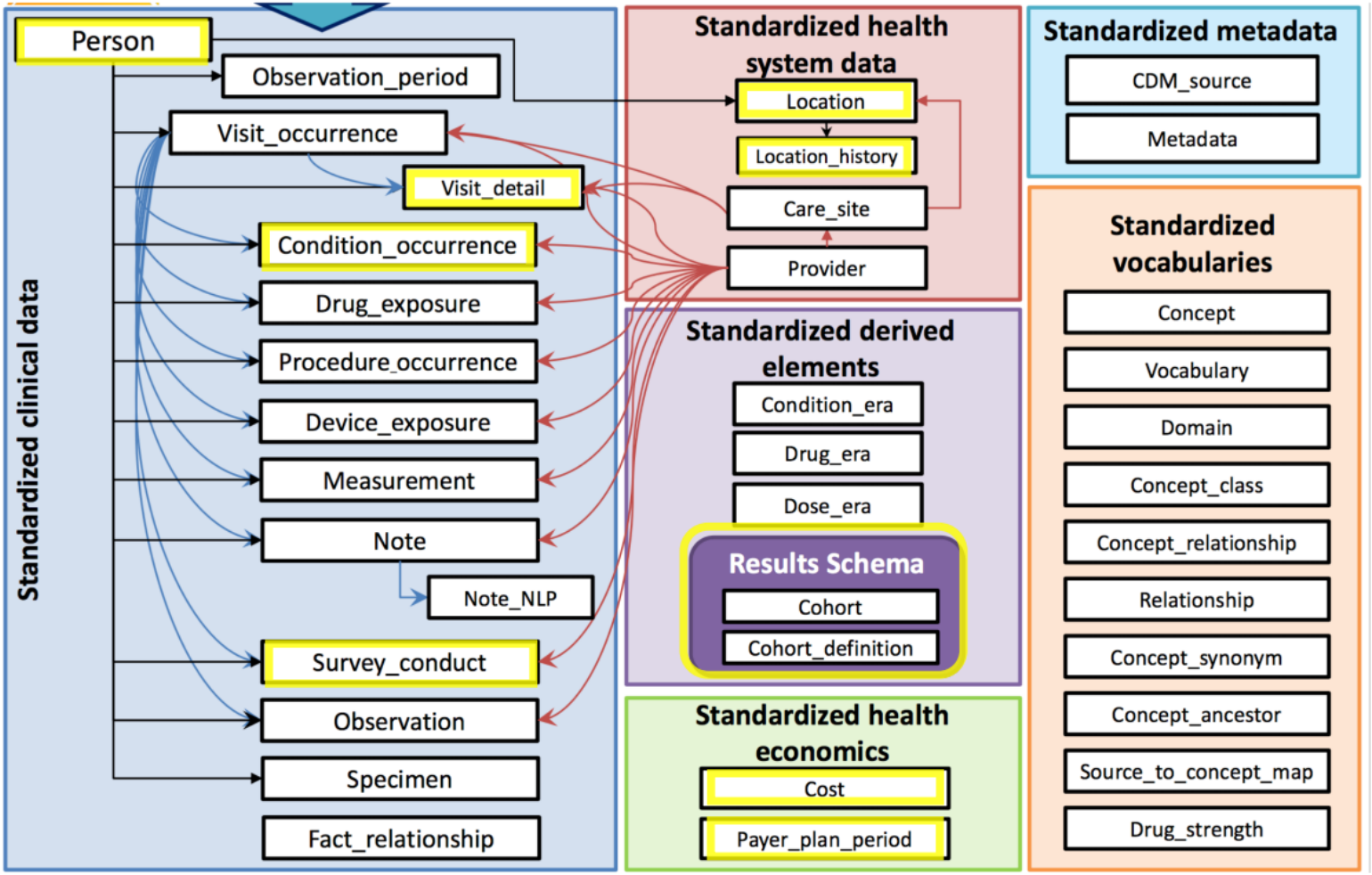
OBSERVATION

OBSERVATION

field	required	type
observation_id	Yes	INTEGER
person_id	Yes	INTEGER
observation_concept_id	Yes	INTEGER
observation_date	No	DATE
observation_datetime	Yes	DATETIME
observation_type_concept_id	Yes	INTEGER
value_as_number	No	FLOAT
value_as_string	No	VARCHAR(60)
value_as_concept_id	No	INTEGER
qualifier_concept_id	No	INTEGER
unit_concept_id	No	INTEGER
provider_id	No	INTEGER
visit_occurrence_id	No	INTEGER
visit_detail_id	No	INTEGER
observation_source_value	No	VARCHAR(50)
observation_source_concept_id	Yes	INTEGER
unit_source_value	No	VARCHAR(50)
qualifier_source_value	No	VARCHAR(50)
observation_event_id	No	INTEGER
obs_event_field_concept_id	Yes	INTEGER
value_as_datetime	No	INTEGER

- 가족력/과거력, 흡연, 음주, 문진 결과, 평가지 등 데이터
- 기타 다른 테이블에 저장할 수 없는 임상 데이터
- 표준용어: SNOMED CT, LOINC

OMOP Version 6 Key Domains



분당서울대 유수영 교수님 Slide

Explore the OMOP Vocabulary

http://athena.ohdsi.org/

What evidence does OHDSI seek to generate 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 hypoglycemia?
  - **Comparative effectiveness**: Does metformin cause hypoglycemia more than glyburide?
- **Patient-level prediction**
  - **Precision medicine**: Given everything you know about me and my medical history, if I start taking glyburide, what is the chance that I am going to have hypoglycemia during the first 30 days?
  - **Disease interception**: Given everything you know about me, what is the chance I will develop diabetes?

How to generate evidence by using cohorts? (target / comparator / outcome)

- **Population-level estimation**
  - **Safety surveillance**: Does metformin cause hypoglycemia?
    - ➔Target cohort: Patients with diabetes
    - ➔Outcome cohort: Patients developing hypoglycemia
  - **Comparative effectiveness**: Does metformin cause hypoglycemia more than glyburide?
    - ➔ Target cohort: DM Patients using metformin
    - ➔ Comparator cohort: DM Patient using glyburide
    - ➔Outcome cohort: Patients developing hypoglycemia

Most Published Research findings are False

Table 4. PPV of Research Findings for Various Combinations of Power ($1 - \beta$), Ratio of True to Not-True Relationships *(R)*, and Bias *(u)*

$1 - \beta$	*R*	*u*	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

Ioannidis, et al., "Why Most Published Research Findings Are False." *PLoS Medicine*, 2005

p-Hacking

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- ☐ Presidents
- ☐ Governors
- ☒ Senators
- ☒ Representatives

How do you want to measure economic performance?

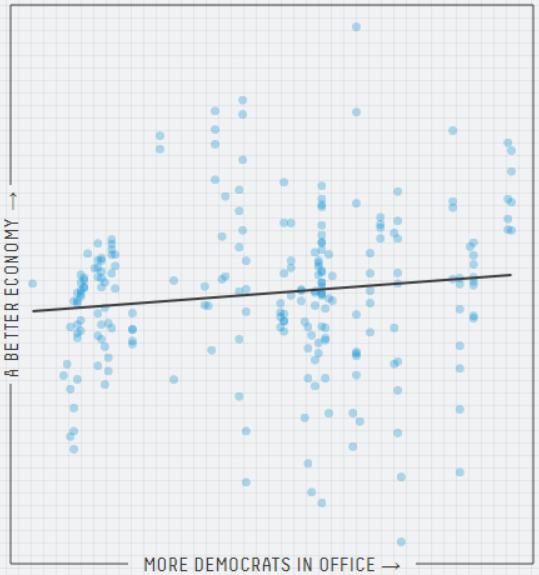
- ☒ Employment
- ☒ Inflation
- ☒ GDP
- ☐ Stock prices

Other options

- ☐ Factor in power
  Weight more powerful positions more heavily
- ☒ Exclude recessions
  Don't include economic recessions

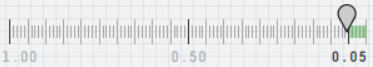
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Almost

Your **0.06** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Threats to reproducible science

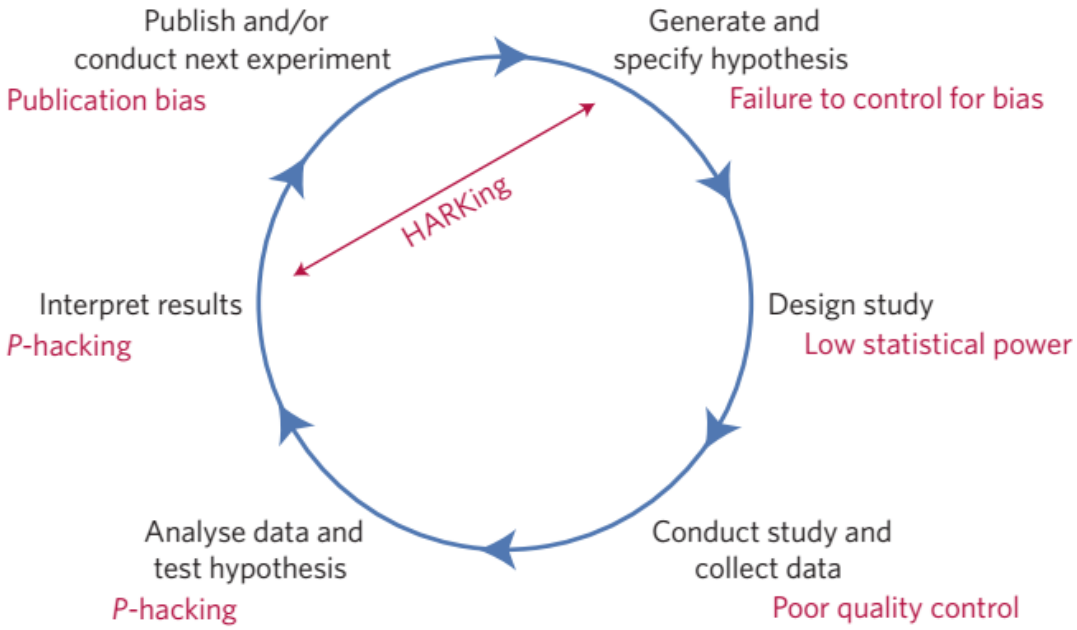
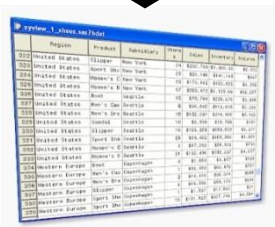


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[1], poor study design, low statistical power[2], analytical flexibility[3], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Munafò et al., *Nature Human Behaviour*, 2017

What do epi studies currently look like?

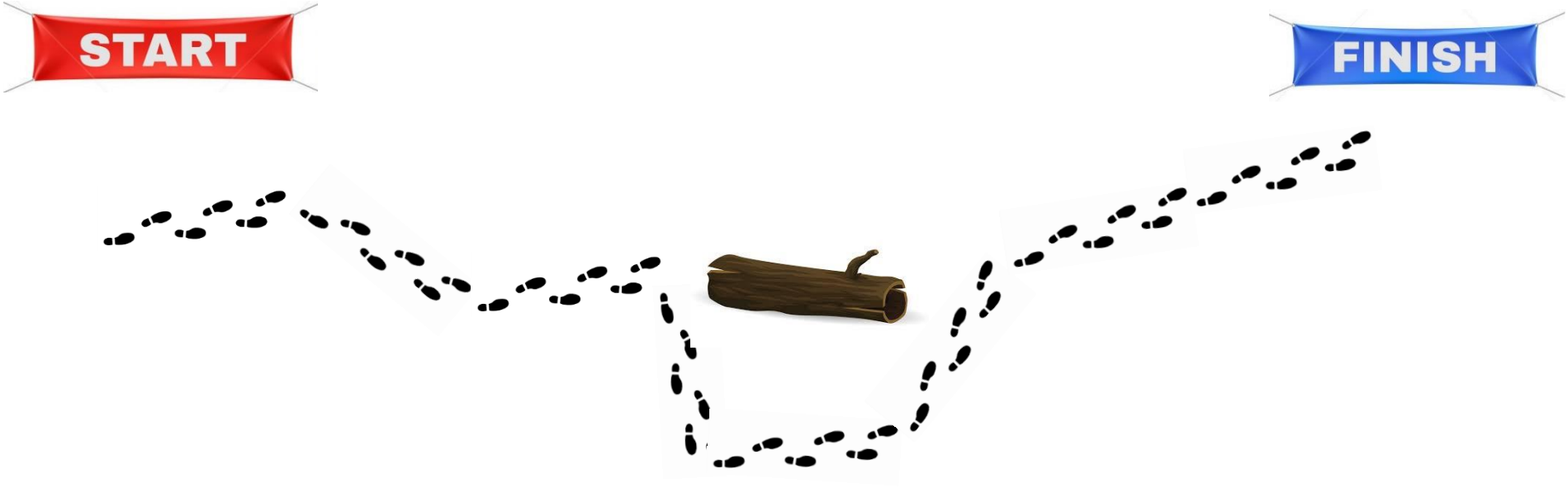


Data set

```
#> Model type: cox
#> Stratified: FALSE
#> Use covariates: FALSE
#> Status: OK
#>
#>            Estimate lower .95 upper .95     logRr seLogRr
#> treatment  0.944681  0.811174  1.096721 -0.056908  0.0769
```

Paper

A journey from data set to paper



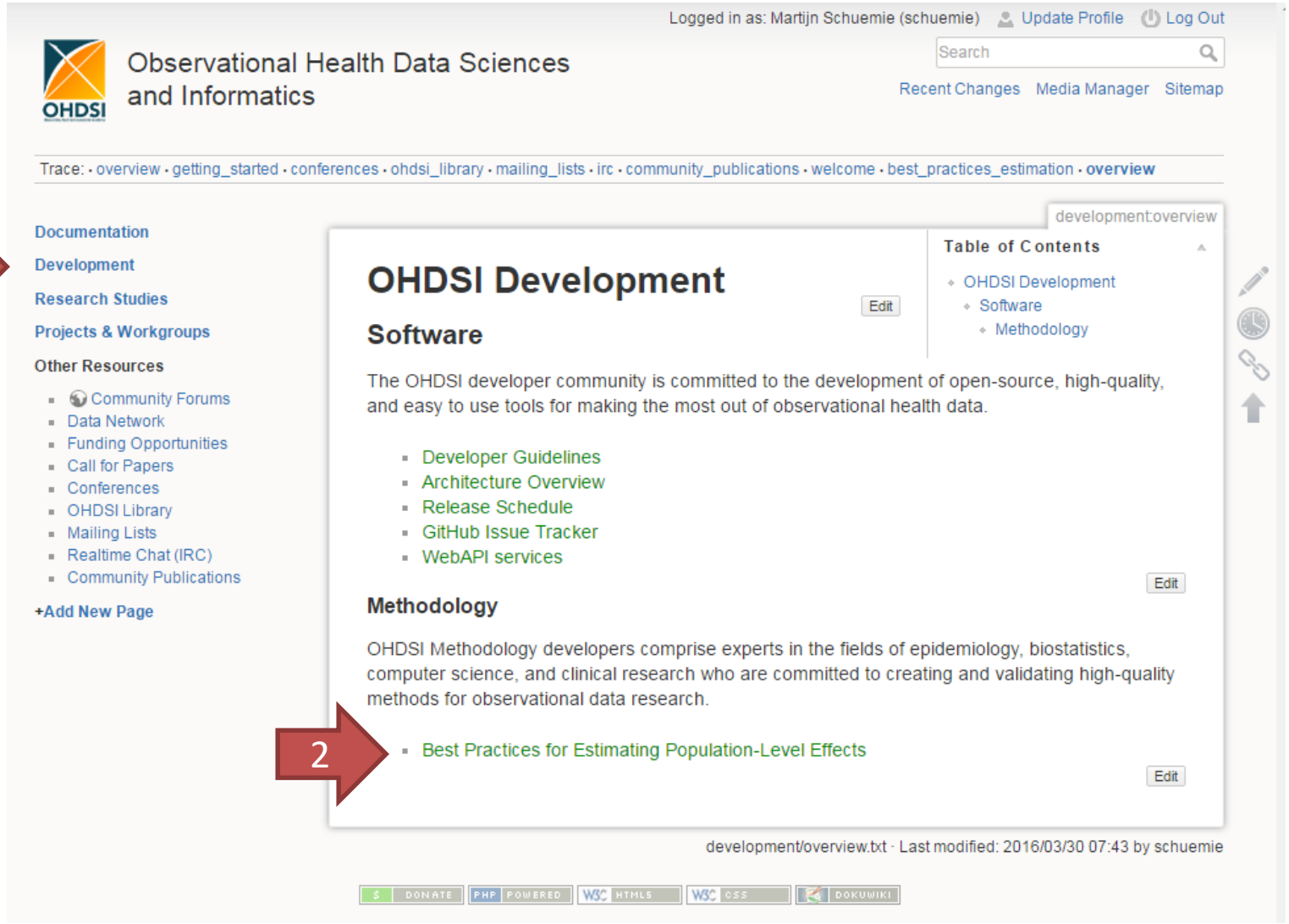
Most epidemiologists view a study as a journey from data set to paper.

- The protocol might be your map
- You will come across obstacles that you will have to overcome
- Several steps will require manual intervention
- In the end, it will be impossible to retrace your exact steps

Current epi studies are non-reproducible

- How do we know what happened?
- How do we know if it was done correctly?
- How do we know how well it worked?
- How could we be more efficient?
- How can we deal with more complex studies?
- How can multiple people work together on the same analysis?
- How could other reproduce this study on a different database?

OHDSI best practices



General principles

- **Prespecify** what you're going to estimate and how: this will avoid hidden multiple testing (publication bias, p-value hacking). Run your analysis only once.

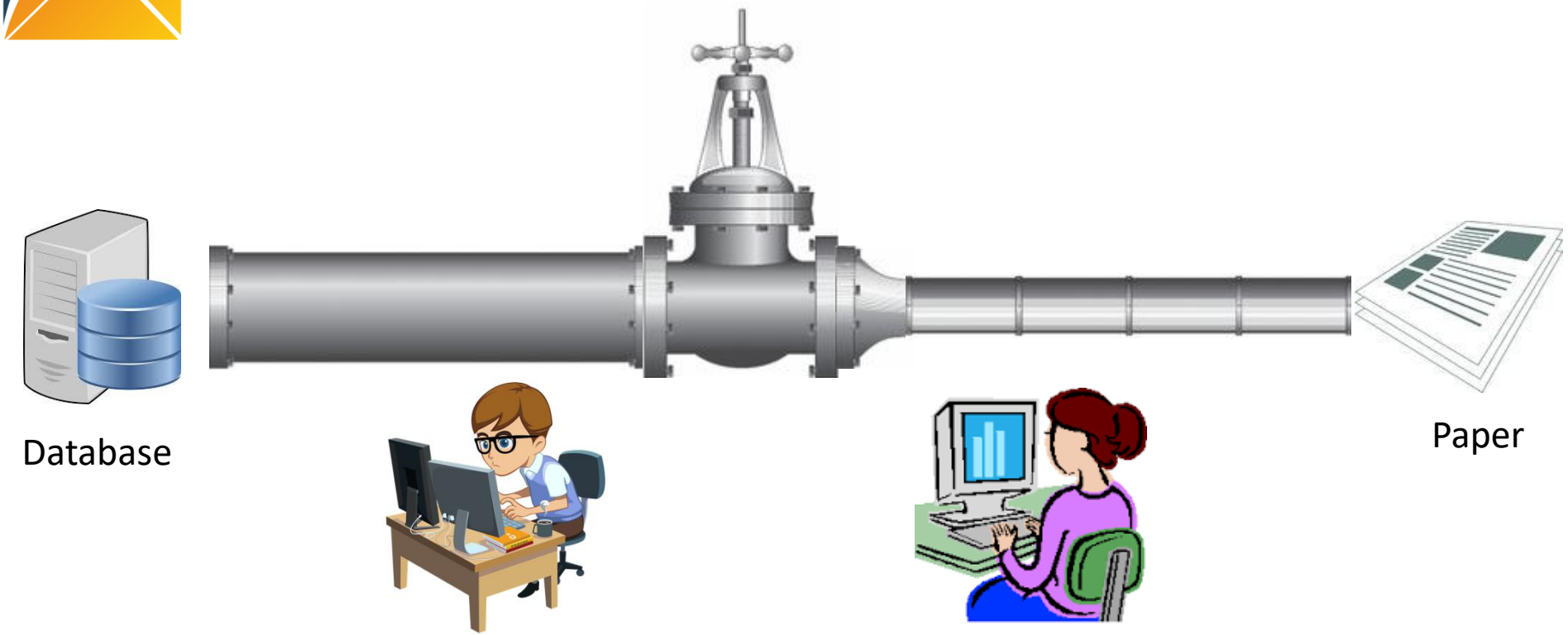
- **Validation of your analysis**: you should have evidence that your analysis does what you say it does (showing that statistics that are produced have nominal operating characteristics (e.g. p-value calibration), showing that specific important assumptions are met (e.g. covariate balance), using unit tests to validate pieces of code, etc.)

- **Transparency**: others should be able to reproduce your study in every detail using the information you provide.

Best practices (generic)

- **Write a full protocol**, and make it public prior to running the study
  - Research question + hypotheses to be tested
  - Which method(s), data, cohort definitions. What is the primary analyses and what are sensitivity analyses?
  - Quality control
  - Amendments and Updates
- **Validate all code** used to produce estimates. The purpose of validation is to ensure the code is doing what we require it to do. Possible options are:
  - Unit testing
  - Simulation
  - Double coding
  - Code review
- Include **negative controls** (exposure-outcome pairs where we believe there is no effect)
- Produce **calibrated p-values**
- Make all analysis code available as **open source** so others can easily replicate your study

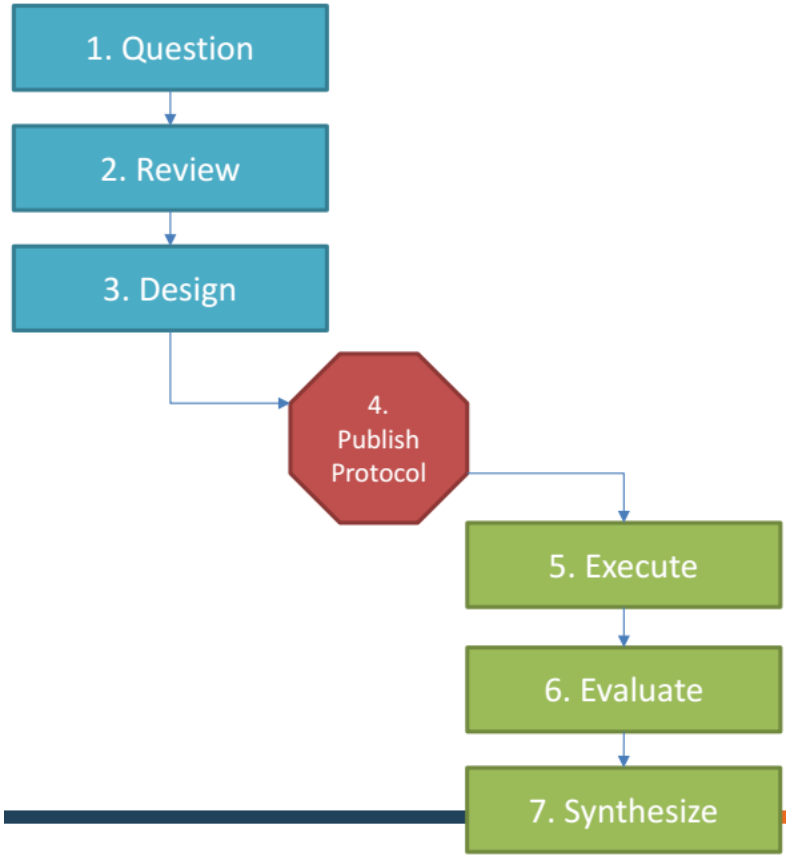
What should OHDSI studies look like?



A study should be like a pipeline

- A fully automated process from database to paper
- ‘Performing a study’ = building the pipeline

A standardized process for evidence generation and dissemination



How OHDSI is trying to help:

OHDSI community

Open-source knowledgebase (LAERTES)

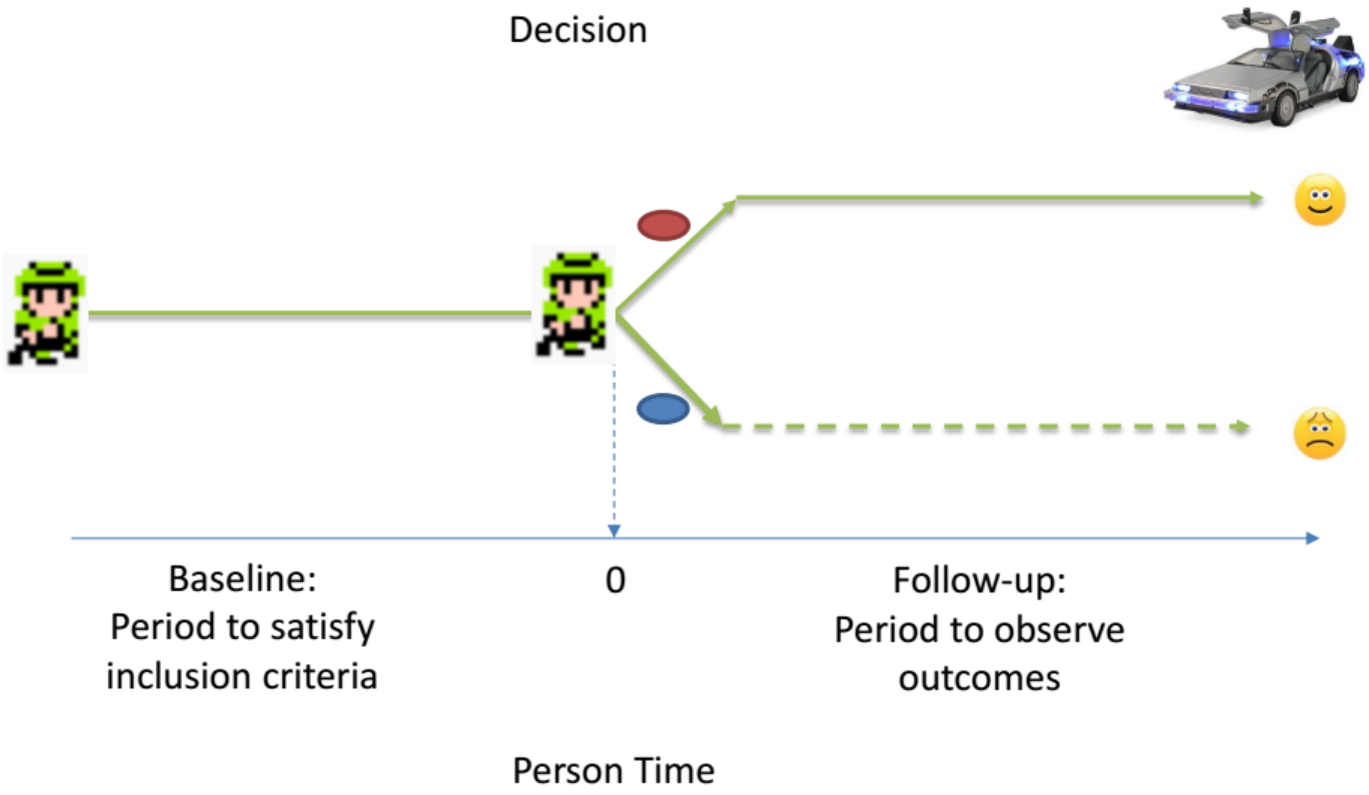
Open-source front-end web applications (ATLAS)

Open-source back-end statistical packages (R Methods Library)

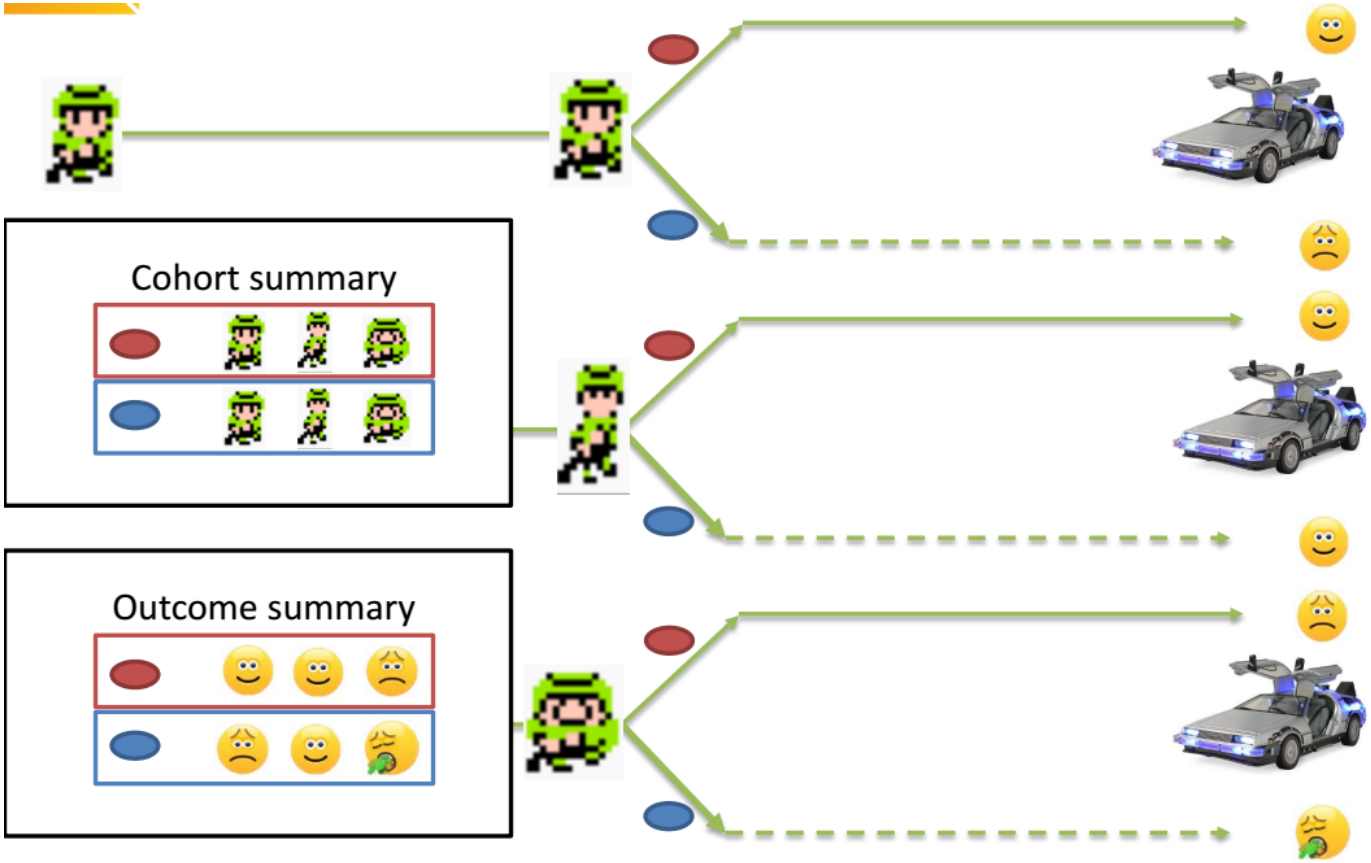
OHDSI network studies

Counterfactual reasoning for one person

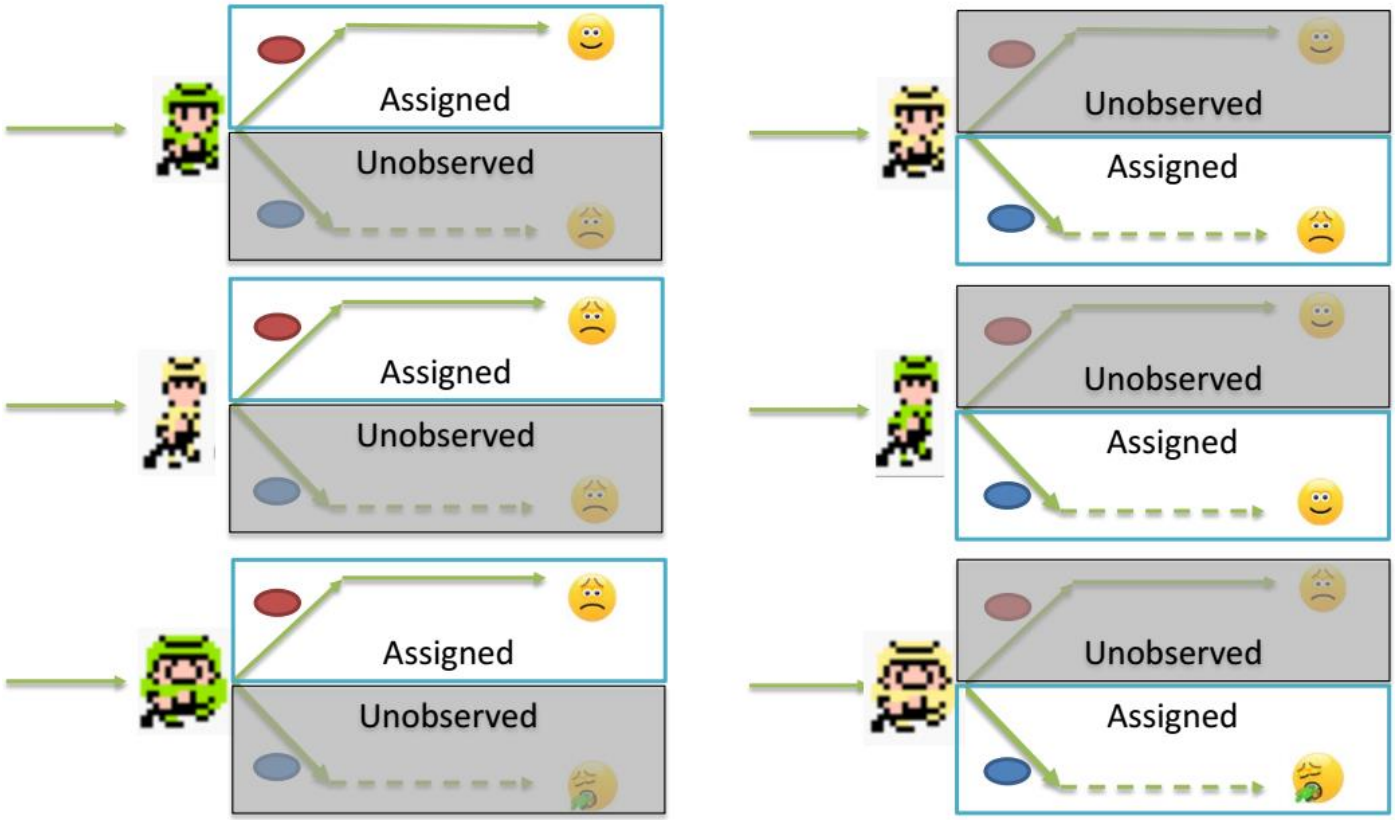
Counterfactual reasoning for one person



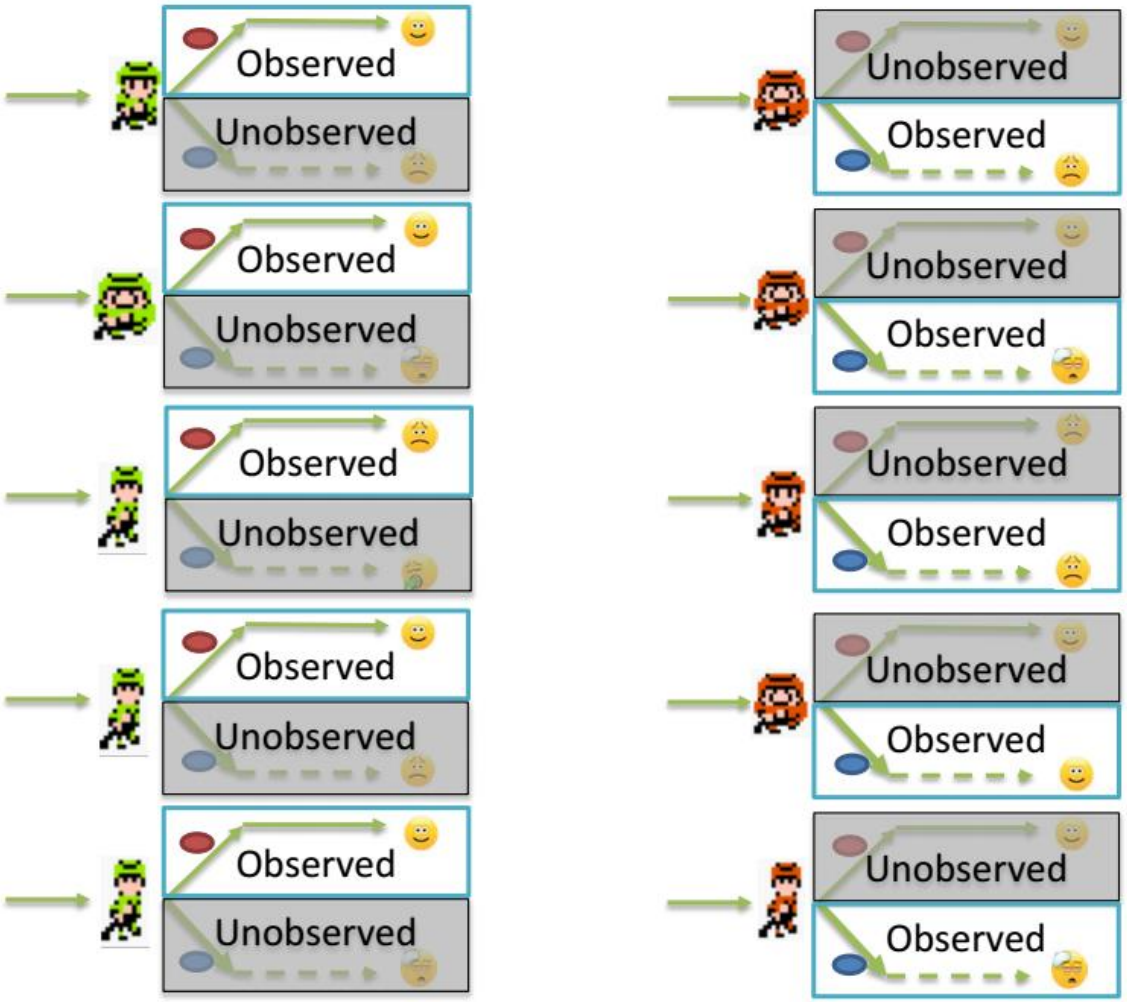
Counterfactual reasoning for a population



Randomized treatment assignment to approximate counterfactual outcomes



An observational comparative cohort design to approximate counterfactual outcomes



Propensity score

- $e(x) = Pr(Z=1|x)$
  - Z is treatment assignment
  - x is a set of all covariates at the time of treatment assignment
- Propensity score = probability of belonging to the target cohort vs. the comparator cohort, given the baseline covariates
- Propensity score can be used as a 'balancing score': if the two cohorts have similar propensity score distribution, then the distribution of covariates should be similar (need to perform diagnostic to check)

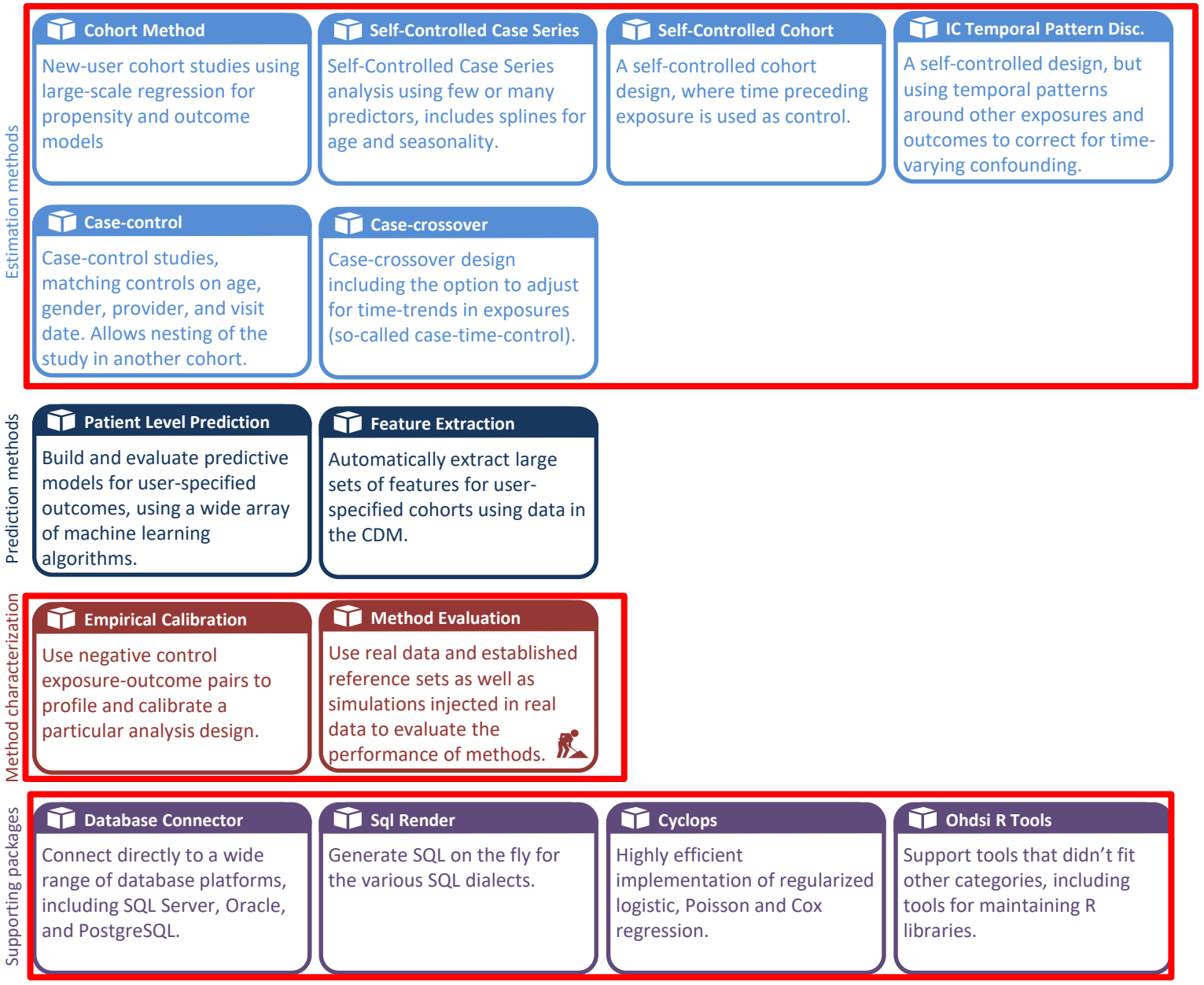
Rubin Biometrika 1983

"Five reasons to use propensity score in pharmacoepidemiology"

- Theoretical advantages
  - Confounding by indication is the primary threat to validity, PS focuses directly on indications for use and non-use of drug under study
- Value of propensity scores for matching or trimming the population
  - Eliminate 'uncomparable' controls without assumptions of linear relationship between PS and outcome
- Improved estimation with few outcomes
  - PS allows matching on one scalar value rather than needing degrees of freedom for all covariates
- Propensity score by treatment interactions
  - PS enables exploration of patient-level heterogeneity in response
- Propensity score calibration to correct for measurement error

Glynn et al, BCPT 2006

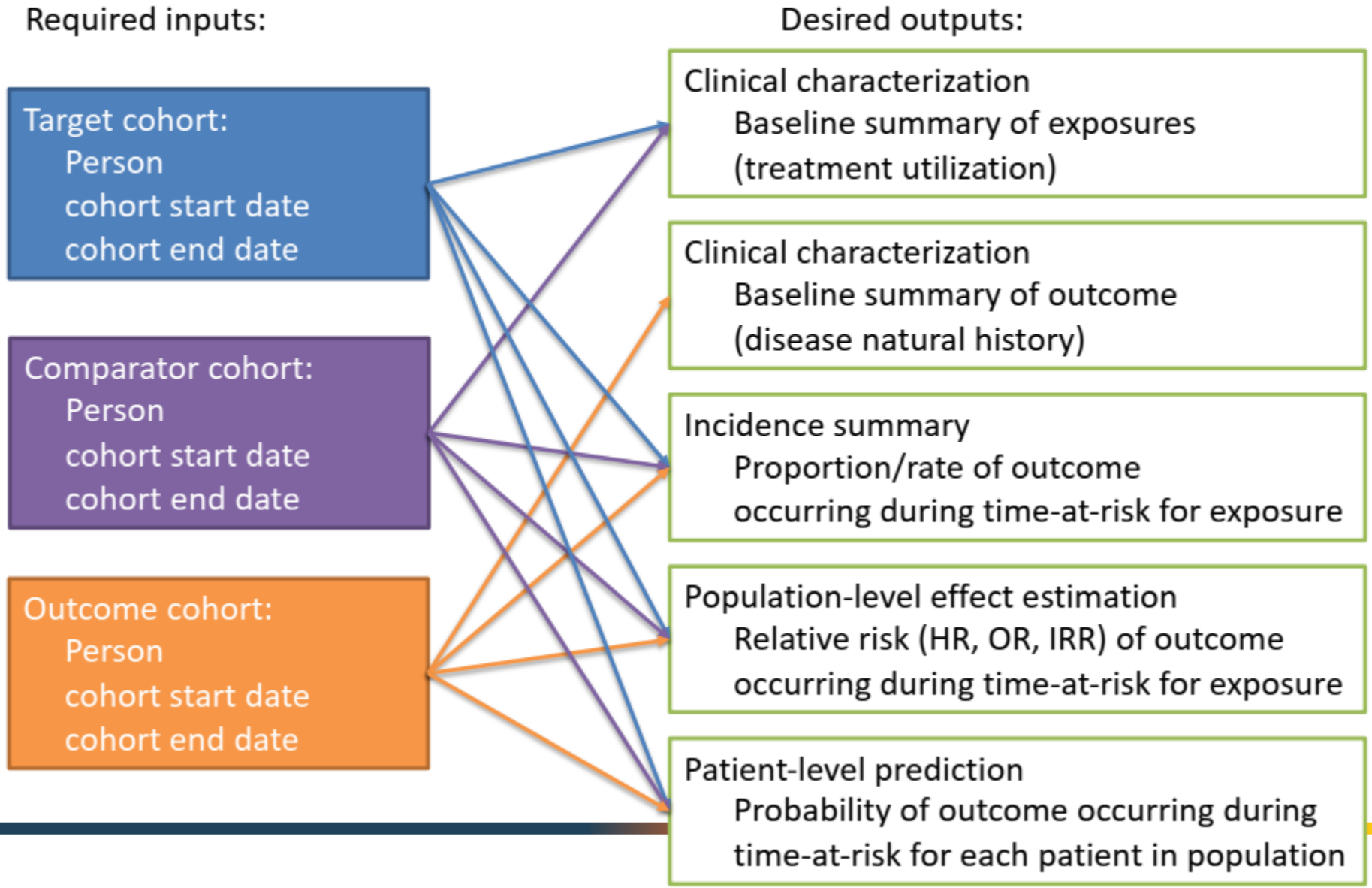
OHDSI Methods Library for **Medical researchers**



Design an observational study like a randomized trial

Input parameter	Design choice
Target cohort (T)	Metformin User
Comparator cohort (C)	Sulfonylurea User
Outcome cohort (O)	Hypoglycemia; CV outcome
Time-at-risk	On-treatment; ITT
Model specification	Cox regression

The common building block of all observational analysis: cohorts



OHDSI's definition of 'cohort'

Cohort = a set of persons who satisfy one or more inclusion criteria 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

Exercise: Define target and outcome cohort for Lactic acidosis in metformin users

- What **initial event**(?) define cohort entry?
- What **inclusion criteria** are applied to the initial events?
- What defines a person's **cohort exit**?

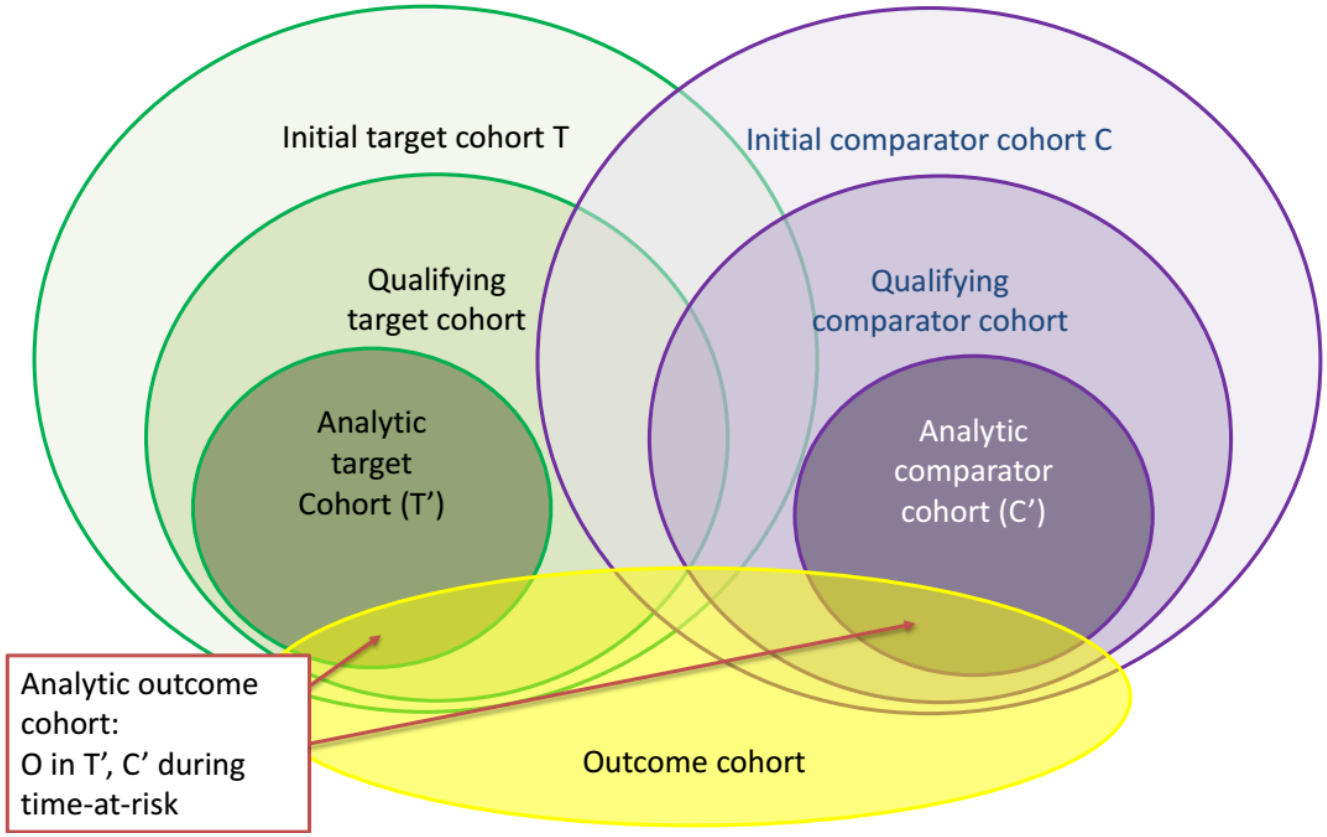
- OHDSI Cohort must have:

Start Date & **End Date**

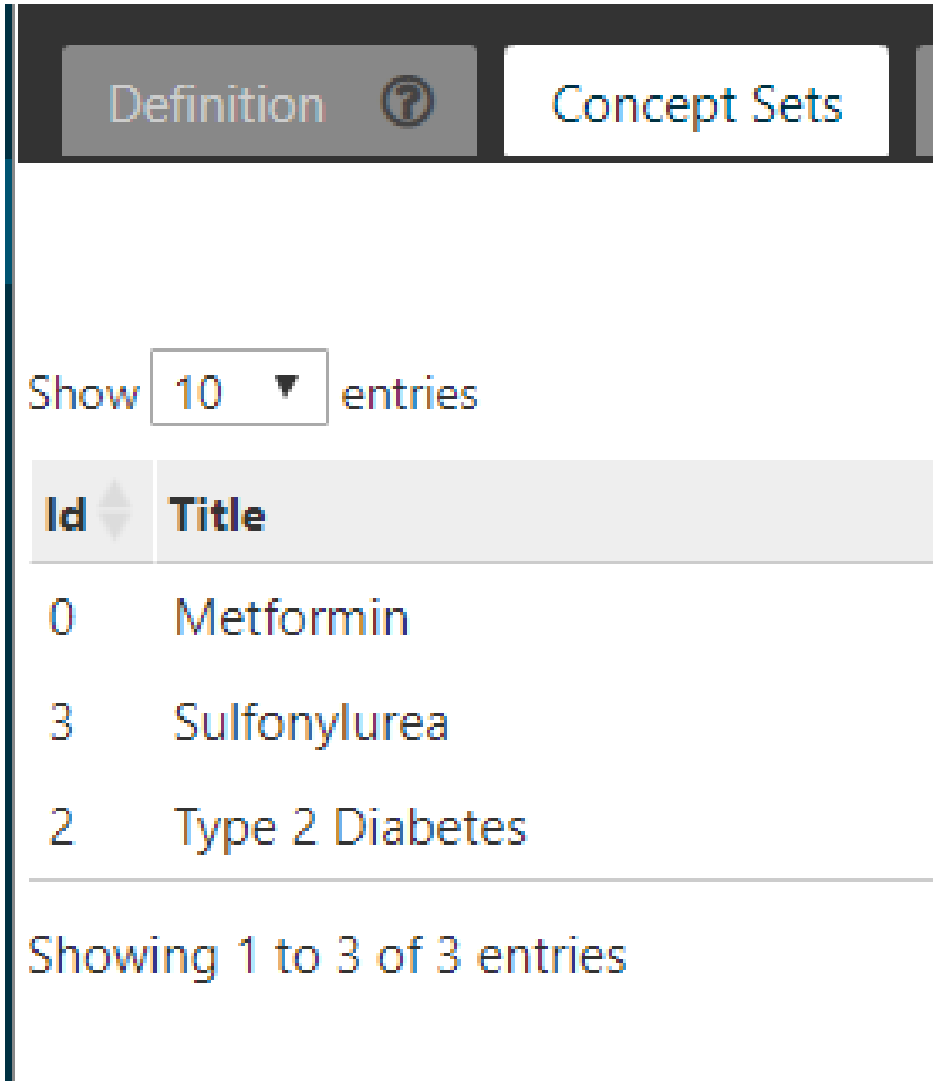
Demo: Defining cohort using ATLAS

- Follow along at: http://ohdsi.org/web/ATLAS

Define cohort: A database is full of cohorts, some of which may represent valid comparisons

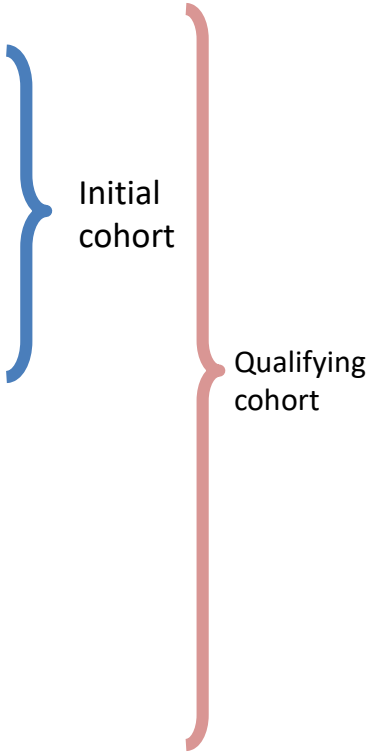


Before start: Making Concept Sets

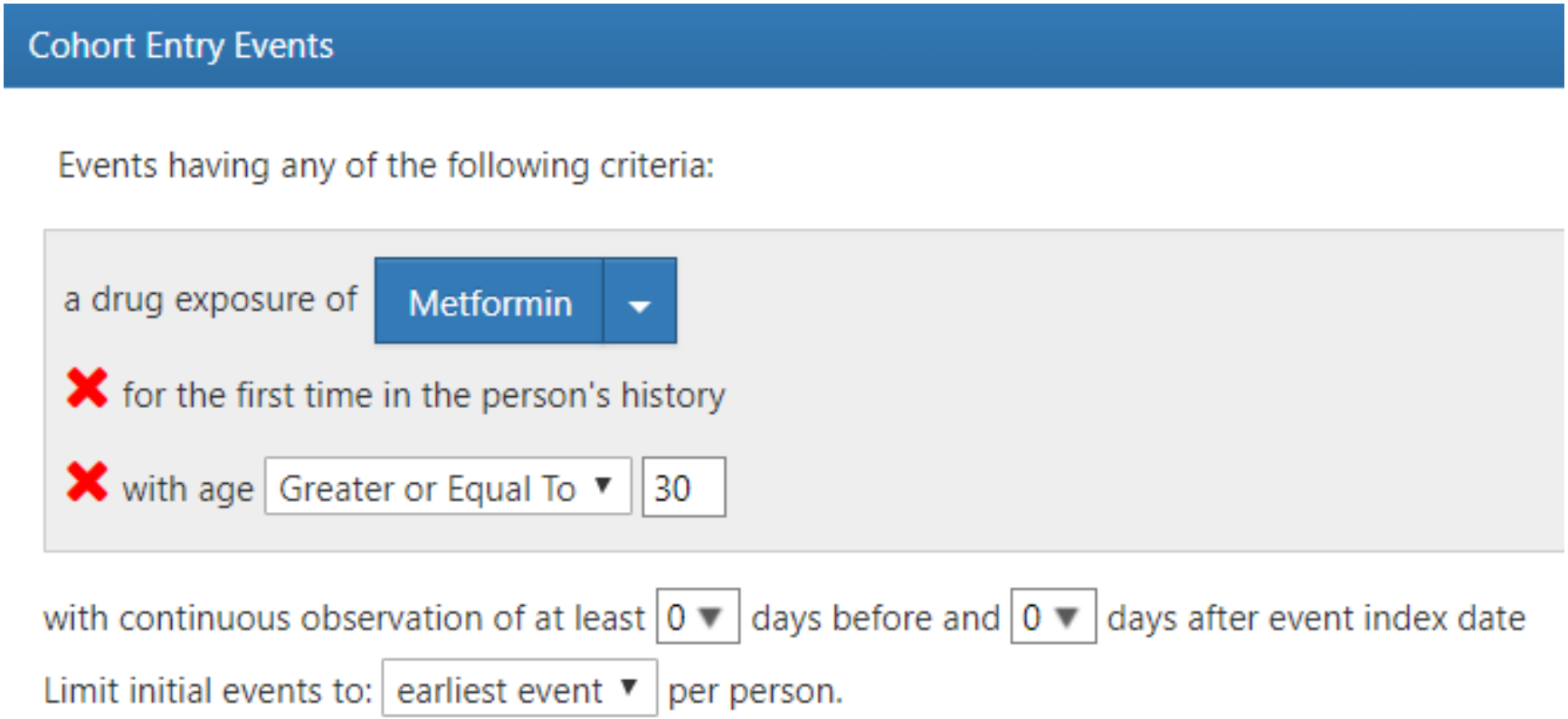


Define cohort: Process flow for formally defining a cohort in ATLAS

- **Cohort entry criteria**
  - **Initial events**
    - Events are recorded time-stamped observations for the persons, such as drug exposures, conditions, procedures, measurements and visits
    - All events have a start date and end date, though some events may have a start date and end date with the same value (such as procedures or measurements).
  - **Initial event inclusion criteria**
  - **Additional qualifying inclusion criteria**
    - The qualifying cohort will be defined as all persons who have an initial event, satisfy the initial event inclusion criteria, and fulfill all additional qualifying inclusion criteria
    - Each qualifying inclusion criteria will be evaluated to determine the impact of the criteria on the attrition of persons from the initial cohort
- **Cohort exit criteria**



Demo: [Target] Metformin User – initial event



http://www.ohdsi.org/web/atlas/#/cohortdefinition/1769410

Demo: [Target] Metformin User – initial event inclusion criteria

Restrict intial events to:

having all ▼ of the following criteria:

with at least ▼ 1 ▼ using all occurrences of:

a condition occurrence of Type 2 Diabetes ▼

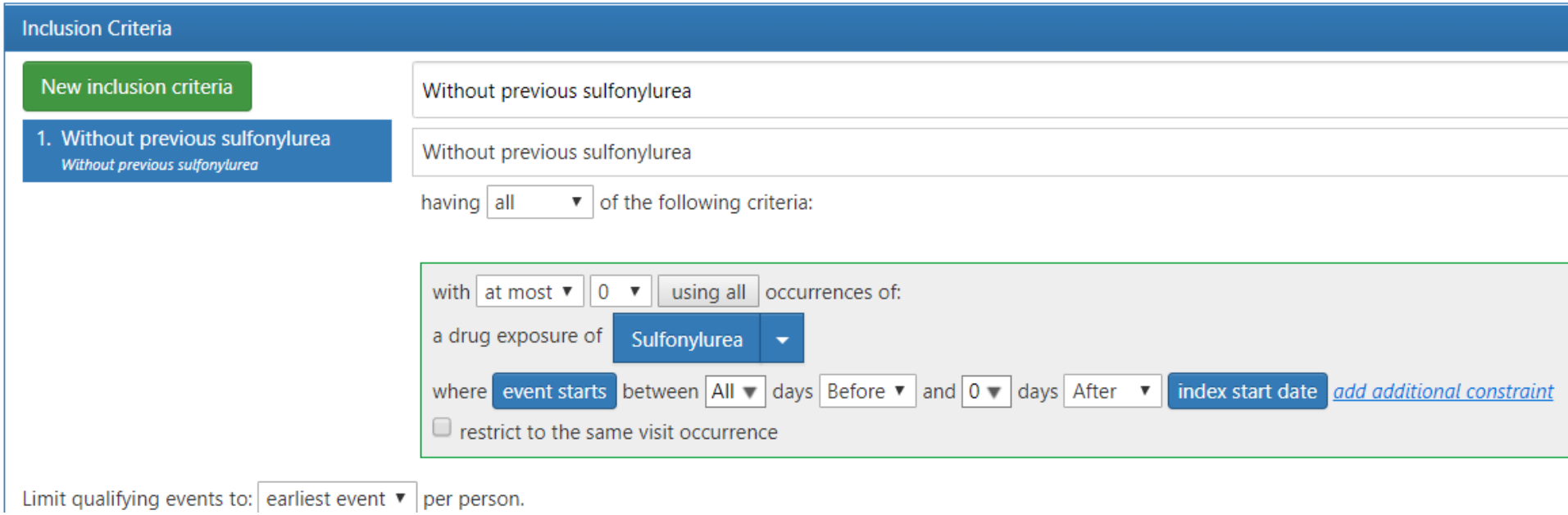
where event starts between All ▼ days Before ▼ and 0 ▼ days After ▼ index start date *add additional constraint*

☐ restrict to the same visit occurrence

Limit initial events to: earliest event ▼ per person.

http://www.ohdsi.org/web/atlas/#/cohortdefinition/1769410

Demo: [Target] Metformin User – additional qualifying inclusion criteria



http://www.ohdsi.org/web/atlas/#/cohortdefinition/1769410

Demo: [Target] Metformin User – cohort exit criteria and censoring event

Cohort Exit

Event Persistence:

Event will persist until: end of a continuous drug exposure

Continuous Exposure Persistence:

Specify a concept set that contains one or more drugs. A drug era will be derived from all drug exposure events for any of the drugs within the conce
successive exposure events and adding a specified surveillance window to the final exposure event. If no exposure event end date is provided, then a
supply is available or event start date + 1 day otherwise. This event persistence assures that the cohort end date will be no greater than the drug era

Concept set containing the drug(s) of interest: Metformin

- Persistence window: allow for a maximum of 30 days between exposure records when inferring the era of persistence exposure
- Surveillance window: add 0 days to the end of the era of persistence exposure as an additional period of surveillance prior to cohort exit.

Censoring Events:

Exit Cohort based on the following criteria:

a drug exposure of Sulfonylurea

http://www.ohdsi.org/web/atlas/#/cohortdefinition/1769410

Demo: [Target] Metformin User – Result

Source Name	Generation Status	People	Records	Generated	Generation Duration	
AUMC_CDM_v5.3	COMPLETE	12,264	12,264	08/07/2019 8:12 AM	00:00:46	View Reports
AUMC_DB_TEST	n/a	n/a	n/a	n/a	n/a	
Dolphin_IQVIA_CDM_v5.2.2	n/a	n/a	n/a	n/a	n/a	
ICARUS_CDM_v5.3	n/a	n/a	n/a	n/a	n/a	
MIMIC3_CDM_v5.3	COMPLETE	0	...	08/07/2019 8:12 AM	00:00:07	View Reports
NHIS_NSC_CDM_2019_v5.3.1	n/a	n/a	n/a	n/a	n/a	
NHIS_NSC_CDM_v5.3	COMPLETE	19,595	19,595	05/23/2019 9:33 AM	00:01:51	View Reports
SynPUF_CDM_v5.2.2	COMPLETE	7,661	7,661	08/07/2019 8:13 AM	00:00:07	View Reports
SynPuf_1k_CDM_v5.2.2	n/a	n/a	n/a	n/a	n/a	
SynPuf_5%_CDM_v5.2.2	COMPLETE	5,474	5,474	08/07/2019 8:13 AM	00:00:05	View Reports

	Match Rate	Matches	Total Events
Summary Statistics:	57.03%	12,264	21,505

Inclusion Rule	N	% Remain	% Diff
1. Without previous sulfonylurea	12,264	57.03%	42.97%

Switch to intersect view

Attrition Visualization

http://www.ohdsi.org/web/atlas/#/cohortdefinition/1769410

Demo: [Comparator] Glyburide User

	Source Name	Generation Status	People	Records	Generated	Generation Duration	
Generate	AUMC_CDM_v5.3	COMPLETE	10,971	10,971	05/23/2019 9:41 AM	00:00:17	View Reports
Generate	AUMC_DB_TEST	n/a	n/a	n/a	n/a	n/a	
Generate	Dolphin_IQVIA_CDM_v5.2.2	n/a	n/a	n/a	n/a	n/a	
Generate	ICARUS_CDM_v5.3	n/a	n/a	n/a	n/a	n/a	
Generate	MIMIC3_CDM_v5.3	n/a	n/a	n/a	n/a	n/a	
Generate	NHIS_NSC_CDM_2019_v5.3.1	n/a	n/a	n/a	n/a	n/a	

By Events | By Person

Inclusion Report for AUMC_CDM_v5.3

	Match Rate	Matches	Total Events
Summary Statistics:	58.57%	10,971	18,732

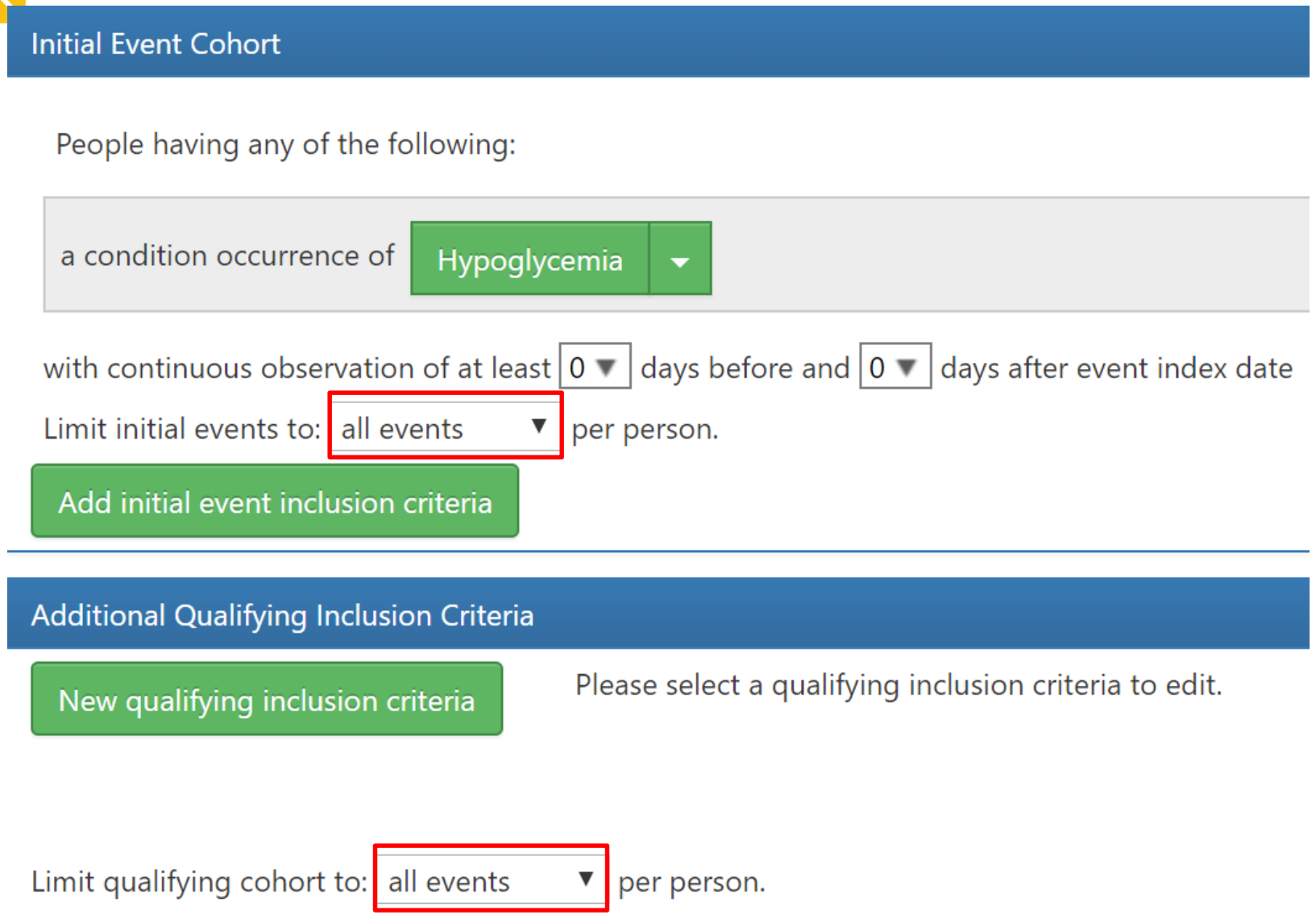
Inclusion Rule	N	% Remain	% Diff
1. Without previous metformin	10,971	58.57%	41.43%

Switch to intersect view

Attrition Visualization

http://www.ohdsi.org/web/atlas/#/cohortdefinition/1769411

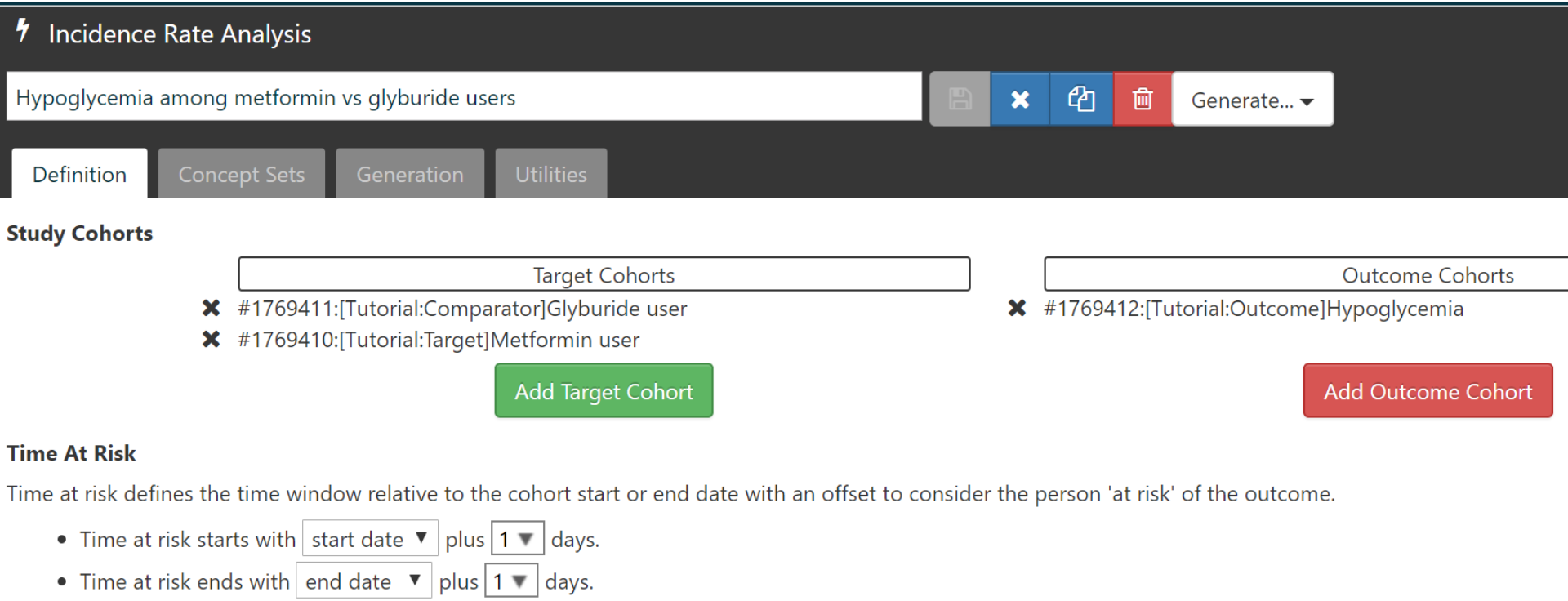
Demo: [Outcome] Hypoglycemia



http://www.ohdsi.org/web/atlas/#/cohortdefinition/1769412

Incidence Rate

What proportion of patients with diabetes experience disease-related complications



http://www.ohdsi.org/web/atlas/#/iranalysis/1747089

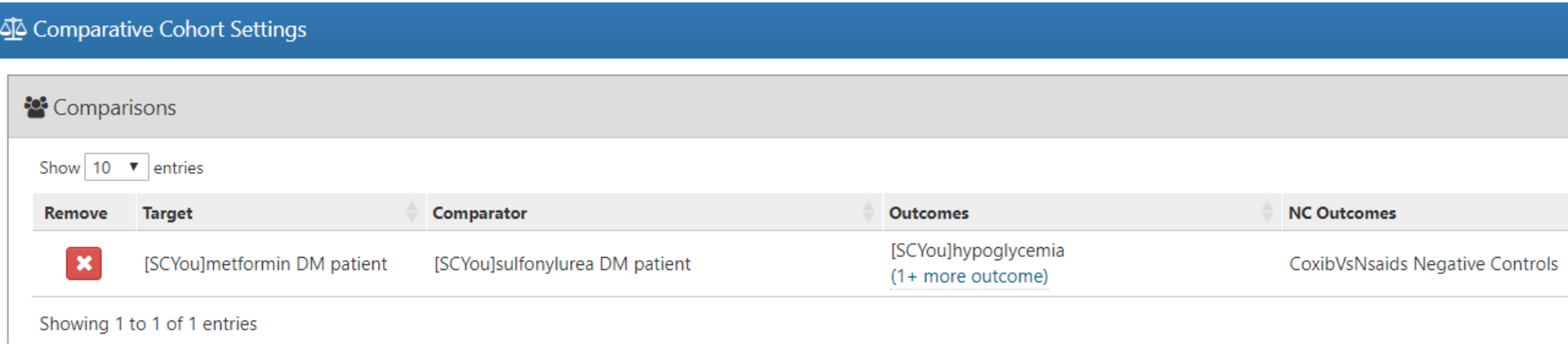
What evidence does OHDSI seek to generate from observational data?

- **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 hypoglycemia?
  - **Comparative effectiveness**: Does metformin cause hypoglycemia more than glyburide?
- **Patient-level prediction**
  - **Precision medicine**: 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 hypoglycemia during the first 30 days?
  - **Disease interception**: Given everything you know about me, what is the chance I will develop diabetes?

3. Define statistical model: The choice of the outcome model defines your research question

	Logistic regression	Poisson regression	Cox proportional hazards
How the outcome cohort is used	Binary classifier of presence/ absence of outcome during the fixed time-at-risk period	Count the number of occurrences of outcomes during time-at-risk	Compute time-to-event from time-at-risk start until earliest of first occurrence of outcome or time-at-risk end, and track the censoring event (outcome or no outcome)
'Risk' metric	Odds ratio	Rate ratio	Hazard ratio
Key model assumptions	Constant probability in fixed window	Outcomes follow Poisson distribution with constant risk	Proportionality – constant relative hazard

Population Level Estimation: Comparative Cohort Settings



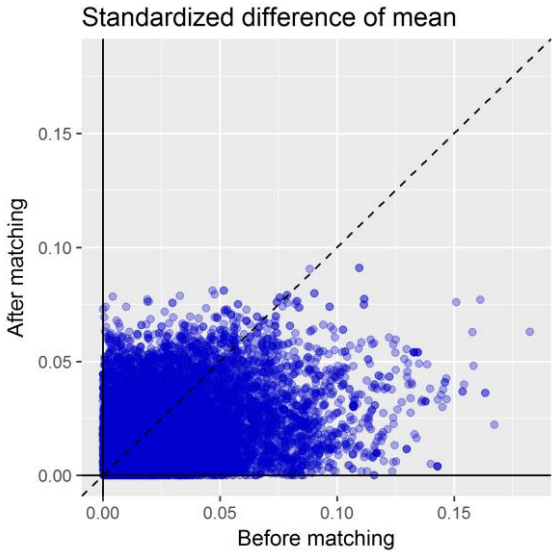
Best practices (new-user cohort design)

- Use **propensity scores** (PS)
- Build PS model using **regularized regression** and a **large set of candidate covariates** (as implemented in the CohortMethod package)
- Use either **variable-ratio matching** or **stratification** on the PS
- **Compute covariate balance** after matching, and terminate study if a covariate has standardized difference > 0.2

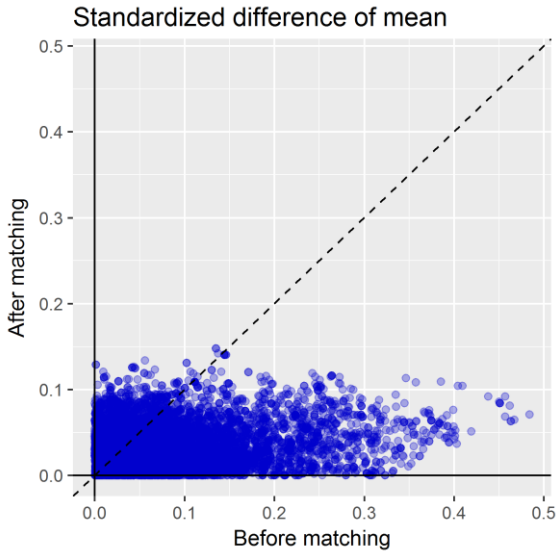
Open questions:
- Terminate study if there's insufficient overlap in PS distributions?
- Require outcome model to be conditioned on matched sets?
- Prescribe Cox models over Poisson and logistic?
- Is there any merit to the dogma stating PS models mustn't include instrumental variables?

3. Define the statistical model and Compare: large scale propensity matching

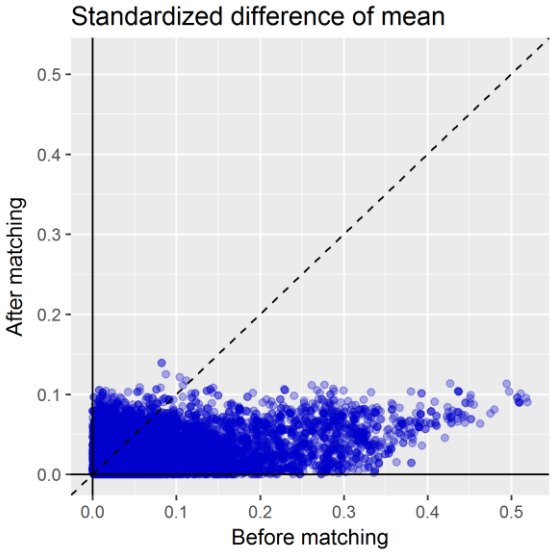
AC vs AD



CD vs AC



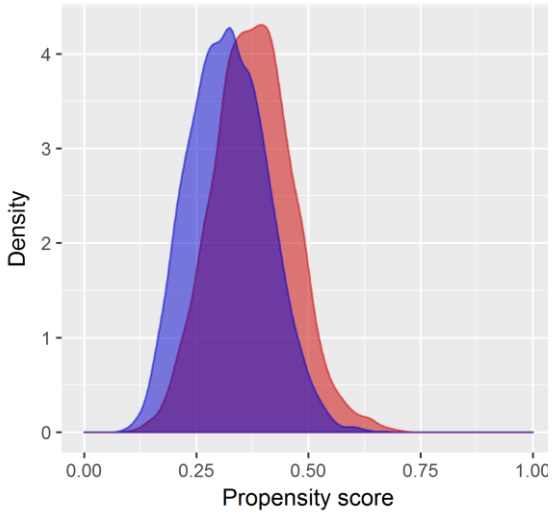
CD vs AD



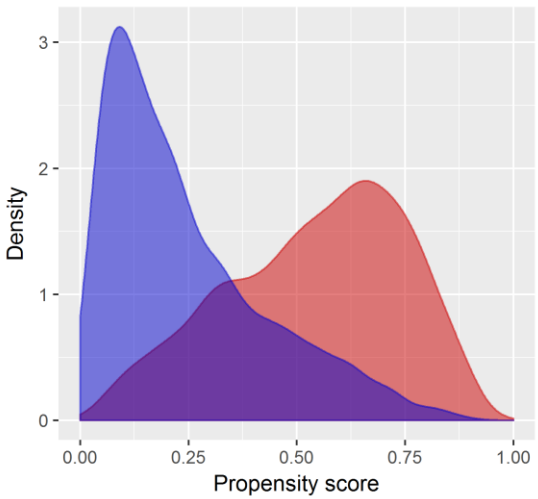
You et al., ESC Congress[Abstract], 2017

3. Define the statistical model and Compare: large scale propensity matching

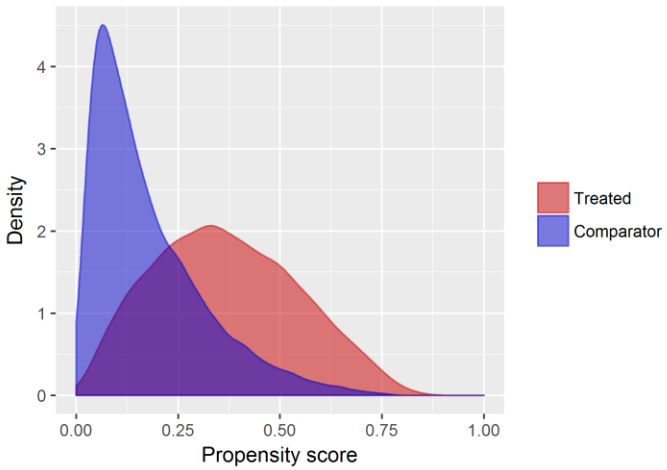
AC vs AD



CD vs AC

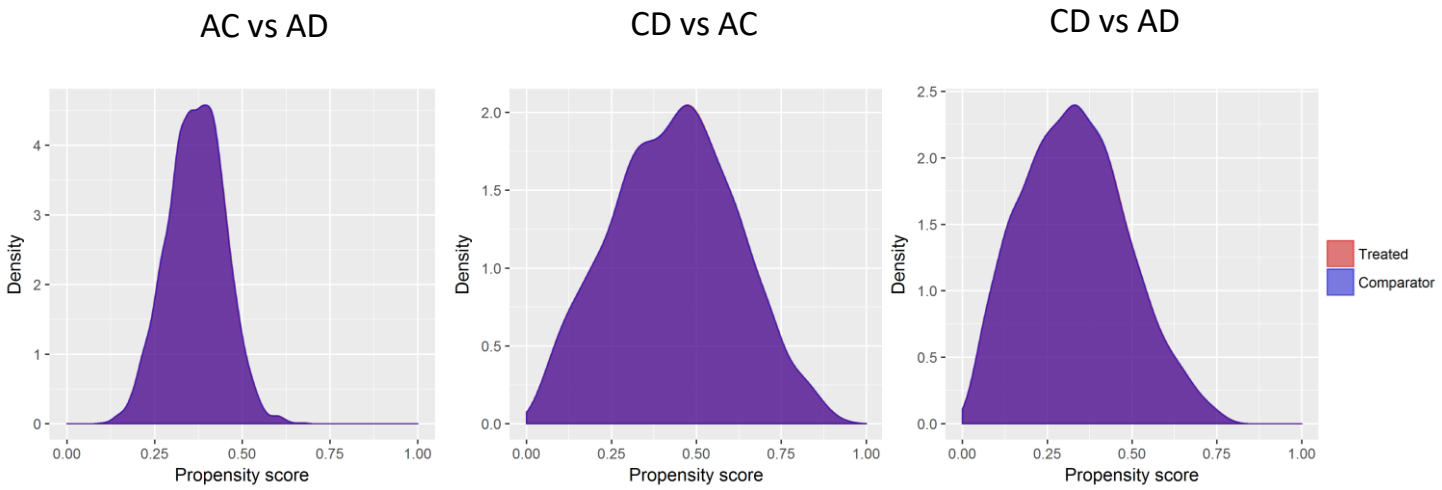


CD vs AD



You et al., ESC Congress[Abstract], 2017

3. Define the statistical model and Compare: large scale propensity matching



You et al., ESC Congress[Abstract], 2017

3. Define the statistical model and Compare: large scale propensity matching

Guertin *et al. BMC Medical Research Methodology* (2016) 16:22
DOI 10.1186/s12874-016-0119-1

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Head to head comparison of the propensity score and the high-dimensional propensity score matching methods

Jason R. Guertin[1,2], Elham Rahme[3,4], Colin R. Dormuth[5] and Jacques LeLorier[6*]

Abstract

Background: Comparative performance of the traditional propensity score (PS) and high-dimensional propensity score (hdPS) methods in the adjustment for confounding by indication remains unclear. We aimed to identify which method provided the best adjustment for confounding by indication within the context of the risk of diabetes among patients exposed to moderate versus high potency statins.

Method: A cohort of diabetes-free incident statins users was identified from the Quebec's publicly funded medico-administrative database (*Full Cohort*). We created two matched sub-cohorts by matching one patient initiated on a lower potency to one patient initiated on a high potency either on patients' PS or hdPS. Both methods' performance were compared by means of the absolute standardized differences (ASDD) regarding relevant characteristics and by means of the obtained measures of association.

Results: Eight out of the 18 examined characteristics were shown to be unbalanced within the *Full Cohort*. Although matching on either method achieved balance within all examined characteristic, matching on patients' hdPS created the most balanced sub-cohort. Measures of associations and confidence intervals obtained within the two matched sub-cohorts overlapped.

Conclusion: Although ASDD suggest better matching with hdPS than with PS, measures of association were almost identical when adjusted for either method. Use of the hdPS method in adjusting for confounding by indication within future studies should be recommended due to its ability to identify confounding variables which may be unknown to the investigators.

Keywords: Confounding by indication, Propensity scores, High-dimensional propensity scores

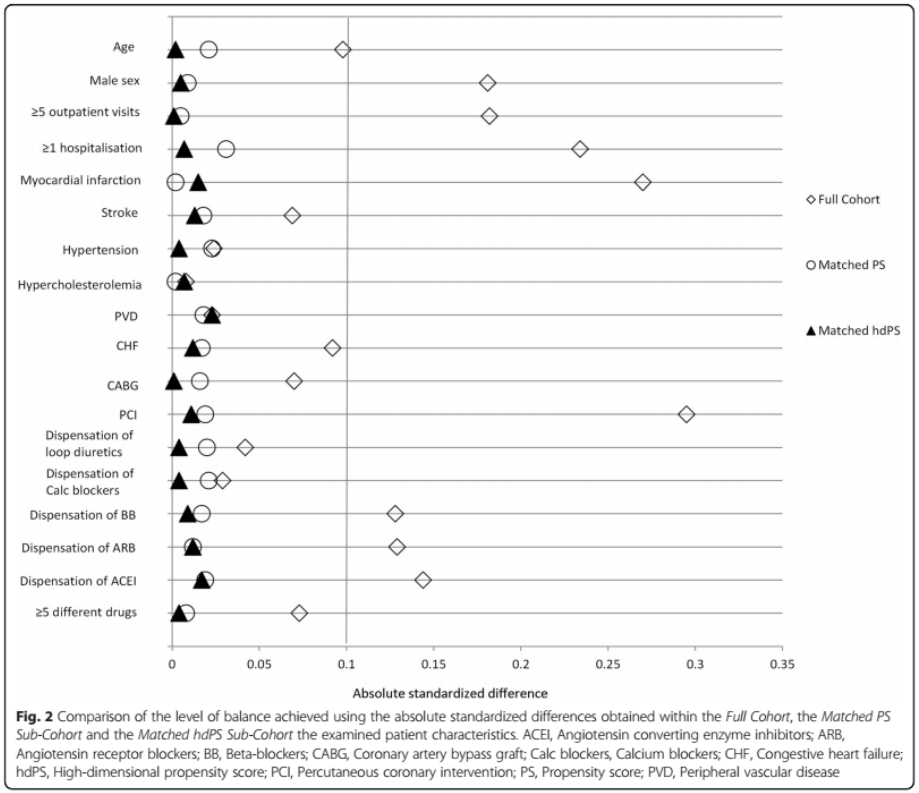


Fig. 2 Comparison of the level of balance achieved using the absolute standardized differences obtained within the *Full Cohort*, the *Matched PS Sub-Cohort* and the *Matched hdPS Sub-Cohort* the examined patient characteristics. ACEI, Angiotensin converting enzyme inhibitors; ARB, Angiotensin receptor blockers; BB, Beta-blockers; CABG, Coronary artery bypass graft; Calc blockers, Calcium blockers; CHF, Congestive heart failure; hdPS, High-dimensional propensity score; PCI, Percutaneous coronary intervention; PS, Propensity score; PVD, Peripheral vascular disease

Guertin et al., *BMC Medical Research Methodology*, 2016

4. Validation: We can check for correctness

- We can review the study code
- We should make the study code publicly available as part of the paper
- Large parts of the study are automatically checked using unit tests

```
test_that("Simple 1-on-1 matching", {
  rowId <- 1:5
  treatment <- c(1, 0, 1, 0, 1)
  propensityScore <- c(0, 0.1, 0.3, 0.4, 1)
  data <- data.frame(rowId = rowId, treatment = treatment, propensityScore = propensityScore)
  result <- matchOnPs(data, caliper = 0, maxRatio = 1)
  expect_equal(result$stratumId, c(0, 0, 1, 1))
})

test_that("Simple 1-on-n matching", {
  rowId <- 1:6
  treatment <- c(0, 1, 0, 0, 1, 0)
  propensityScore <- c(0, 0.1, 0.12, 0.85, 0.9, 1)
  data <- data.frame(rowId = rowId, treatment = treatment, propensityScore = propensityScore)
  result <- matchOnPs(data, caliper = 0, maxRatio = 100)
  expect_equal(result$stratumId, c(0, 0, 0, 1, 1, 1))
})
```

4. Validation: We can evaluate how well the study worked

- Included 100 negative control outcomes
- Results show little residual confounding when using propensity score matching

Epidemiology. 2010 May ; 21(3): 383–388. doi:10.1097/EDE.0b013e3181d61eeb.

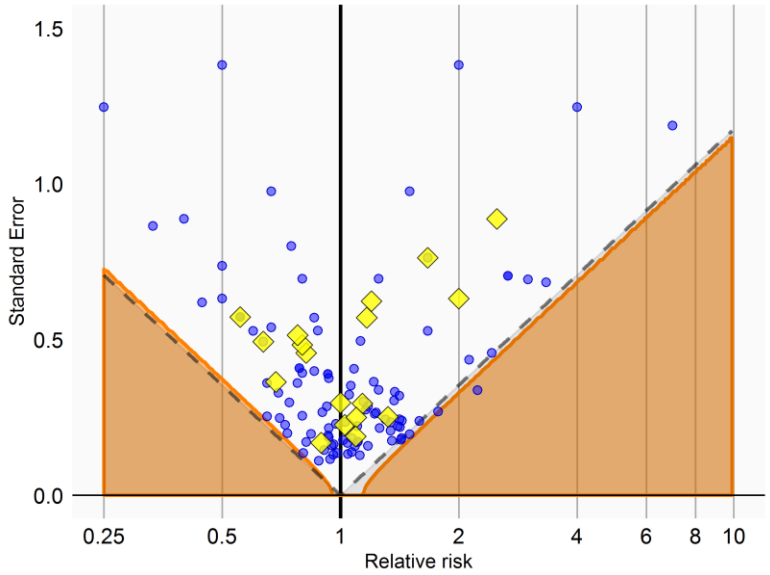
Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies

Marc Lipsitch[1,2,3], **Eric Tchetgen Tchetgen**[1,3,4], and **Ted Cohen**[5,1,3]

[1] Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115

Lipsitch et al., *Epidemiology* (2010)

4. Validation: We can evaluate how well the study worked



Target vs. Comparator	Total negative outcomes	False Positive count	False positive proportion
AC vs AD	38	1	0.026
CD vs AD	38	2	0.053
CD vs AC	37	1	0.027

You et al., ESC Congress[Abstract], 2017

4. Validation: Validation of OHDSI statistical tool

Large-Scale Population-Level Evidence Generation

Objective: - Generate evidence for the comparative effectiveness for each pairwise comparison of depression treatments for a set of outcomes of interest..

NHIS_NSC (20170911)		Crude	Adjusted
Negative Control	TrueNeg	1383	1512
	TotalNeg	1646	1637
	Proportion	0.84	0.92
Positive Control	TruePos	1424	1251
	TotalPos	1815	1815
	Proportion	0.78	0.69

True_Positive_Proportion (20170919)		
Effect_Size	Crude	Adjusted
1.5	0.53	0.38
2.0	0.84	0.70
4.0	0.99	0.99

5. Recruiting data partners and Aggregating their results

- Recruiting data partners
  - Posting on Forum

Study on comparison of combination treatment in hypertension

Researchers

SCYou Seng Chan You — Sep '16

The new study below is planned to post to the OHDSI Research Network.

Comparison of combination treatment in hypertension

Objective: The objective of this study is to compare the effectiveness and adverse events between the combination treatments in hypertension

Rationale: The goal of antihypertensive therapy is to reduce cardiovascular end points including stroke, myocardial infarction, and heart failure by lowering blood pressure. Although it is evident that BP reduction per se is the primary determinant of CV risk reduction, the choice of initial drug therapy can exert some effect on long-term outcomes. Many large randomized trials have shown that two or more antihypertensive agents are required for reaching their treatment goals. Furthermore, recent data have suggested that the use of combination therapy in patients with hypertension may be beneficial for blood-pressure-lowering efficacy, obtaining blood pressure goals earlier, and reducing major adverse cardiovascular events. To date, However, the best combination treatment in hypertension have not been demonstrated. The evidence through OHDSI network can help clinicians to select the combination treatment for their patients.

Project Leads: Seng Chan You, Sungjae Jung, and Rae Woong Park from Ajou university

Please provide any comments or suggestions

5. Recruiting data partners and Aggregating their results

- Whole process of analysis was packaged as a software in R and released for reproducible research
  - https://github.com/OHDSI/StudyProtocolSandbox/tree/master/HypertensionCombination

Execute the following code:

```
library(HypertensionCombination)

cdmDatabaseSchema<-"OMOP CDM DATABASE SCHEMA"
resultsDatabaseSchema<-"RESULT DATABASE SCHEMA"

connectionDetails<-DatabaseConnector::createConnectionDetails(dbms="DBMS",
                                                              server="SERVER IP",
                                                              user="ID",
                                                              password="PW")

execute(connectionDetails,
        cdmDatabaseSchema = cdmDatabaseSchema,
        resultsDatabaseSchema = resultsDatabaseSchema,
        exposureTable = "exposureTable",
        outcomeTable = "outcomeTable",
        cdmVersion = 5,
        outputFolder = "output",
        createCohorts = TRUE,
        runAnalyses = TRUE,
        maxCores = 4,
        packageResults = TRUE,
        createTableAndFigures=TRUE,
        writeReport = TRUE,
        compressResults = TRUE,
        submitResults = TRUE,
        localName = "YOUR LOCAL NAME")
```

5. Recruiting data partners and Aggregating their results

- Whole process of analysis was packaged as a software in R and released for reproducible research
  - https://github.com/OHDSI/StudyProtocolSandbox/tree/master/HypertensionCombination
- The whole analytic process was prespecified before conduction
- Only pre-specified aggregated results absent of patient-level information is collected for meta-analysis and interpretation
  - Meta-analysis: Random-effect model

5. Recruiting data partners and Aggregating their results

Study on comparison of combination treatment in hypertensior

Researchers

SCYou Seng Chan You — Sep '16

The new study below is planned to post to the OHDSI Research Network.

Comparison of combination treatment in hypertension

Hi Chan:

You did a great job presenting at the OHDSI Symposium.

I just got internal J&J permission to participate in your study, so we are set to go. On Monday, I have my next team meeting, where executing your study will be a new project that will get assigned within the team. Martijn has already expressed interest in participating, and I suspect that at least one other in may team will want to join in as well. In terms of database contributions, I suspect we should have sufficient information in a few different US databases (Truven CCAE, Truven MDCR, Optum SES), in the IMS Germany and JMDC databases. We could consider applying for ISAC approval to use CPRD for UK EHR. It could be worth applying, but our experience is that can be quite time-consuming so we may not want to wait for the results before moving forward.

It was great seeing you and Rae at the Symposium. Thank you for all your valuable contributions to the community.

Cheers,

Patrick

5. Recruiting data partners and Aggregating their results

Study on comparison of combination treatment in hypertensior

Researchers

SCYou Seng Chan You — Sep '16

The new study below is planned to post to the OHDSI Research Network.

Comparison of combination treatment in hypertension

Greetings and congrats on your award. I'd like to join. I'm at the Univ of Colorado. We have not completed our medication mappings to OMOP- and obviously this will need to be done. Do you have an expected date that the data pull?

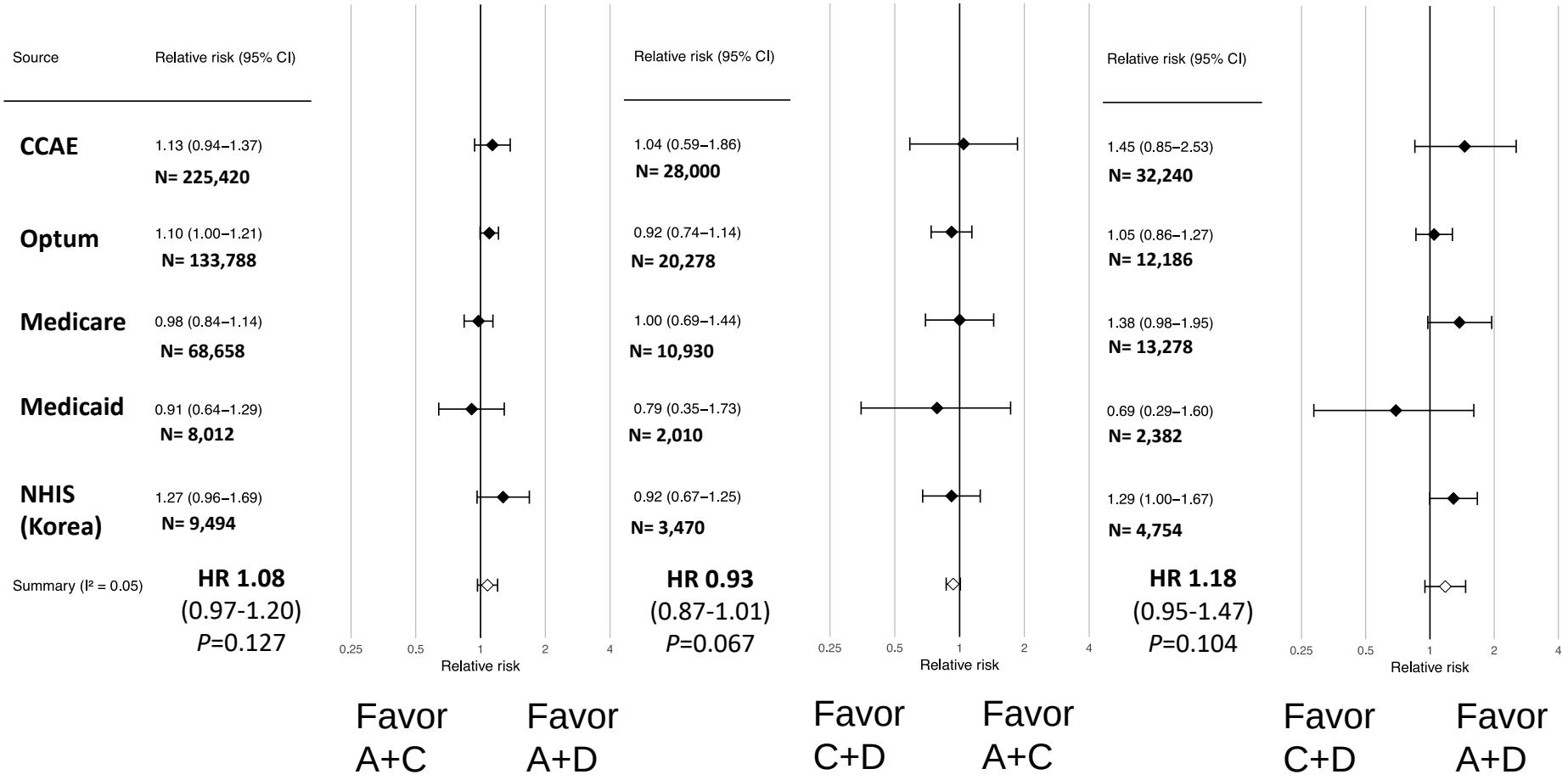
I'm also with the University of California Medical Centers, and I'll solicit their involvement too.

When I have chance, I'll review the protocol, etc

Lisa

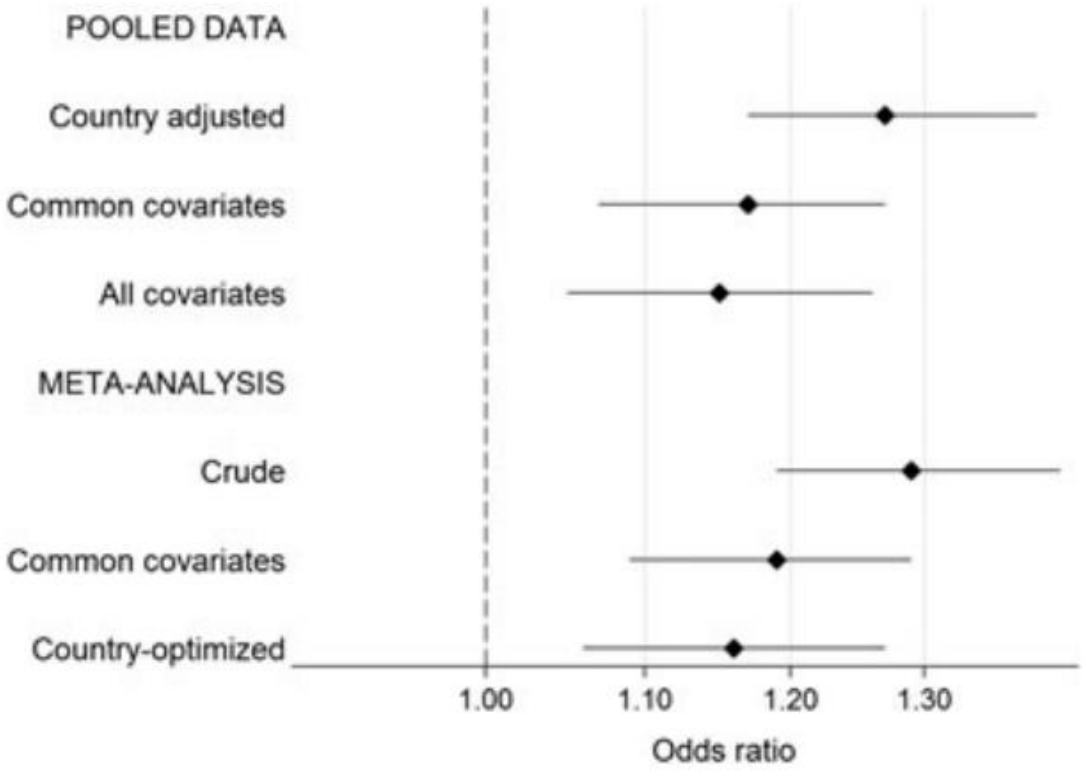
Lisa Schilling, MD, MSPH
Professor of Medicine
Department of Medicine, Division of General Internal Medicine
University of Colorado, School of Medicine

5. Recruiting data partners and Aggregating their results



5. Recruiting data partners and Aggregating their results

Individual-based versus aggregate meta-analysis in multi-database studies of pregnancy outcomes: the Nordic example of selective serotonin reuptake inhibitors and venlafaxine in pregnancy



Selmer et al., *Pharmacoepidemiology and Drug Safety* (2016)

5. Recruiting data partners and Aggregating their results

Biometrika. 2015 June ; 102(2): 281–294. doi:10.1093/biomet/asv011.

On random-effects meta-analysis

D. ZENG and **D. Y. LIN**
Department of Biostatistics, CB #7420, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A
D. ZENG: dzeng@bios.unc.edu; D. Y. LIN: lin@bios.unc.edu

Summary

Meta-analysis is widely used to compare and combine the results of multiple independent studies. To account for between-study heterogeneity, investigators often employ random-effects models, under which the effect sizes of interest are assumed to follow a normal distribution. It is common to estimate the mean effect size by a weighted linear combination of study-specific estimators, with the weight for each study being inversely proportional to the sum of the variance of the effect-size estimator and the estimated variance component of the random-effects distribution. Because the estimator of the variance component involved in the weights is random and correlated with study-specific effect-size estimators, the commonly adopted asymptotic normal approximation to the meta-analysis estimator is grossly inaccurate unless the number of studies is large. When individual participant data are available, one can also estimate the mean effect size by maximizing the joint likelihood. We establish the asymptotic properties of the meta-analysis estimator and the joint maximum likelihood estimator when the number of studies is either fixed or increases at a slower rate than the study sizes and we discover a surprising result: the former estimator is always at least as efficient as the latter. We also develop a novel resampling technique that improves the accuracy of statistical inference. We demonstrate the benefits of the proposed inference procedures using simulated and empirical data.

maximizing the joint likelihood. We establish the asymptotic properties of the meta-analysis estimator and the joint maximum likelihood estimator when the number of studies is either fixed or increases at a slower rate than the study sizes and we discover a surprising result: the former estimator is always at least as efficient as the latter. We also develop a novel resampling technique that improves the accuracy of statistical inference. We demonstrate the benefits of the proposed inference procedures using simulated and empirical data.

Zeng and Lin. *Biometrika* (2015)

5. Recruiting data partners and Aggregating their results

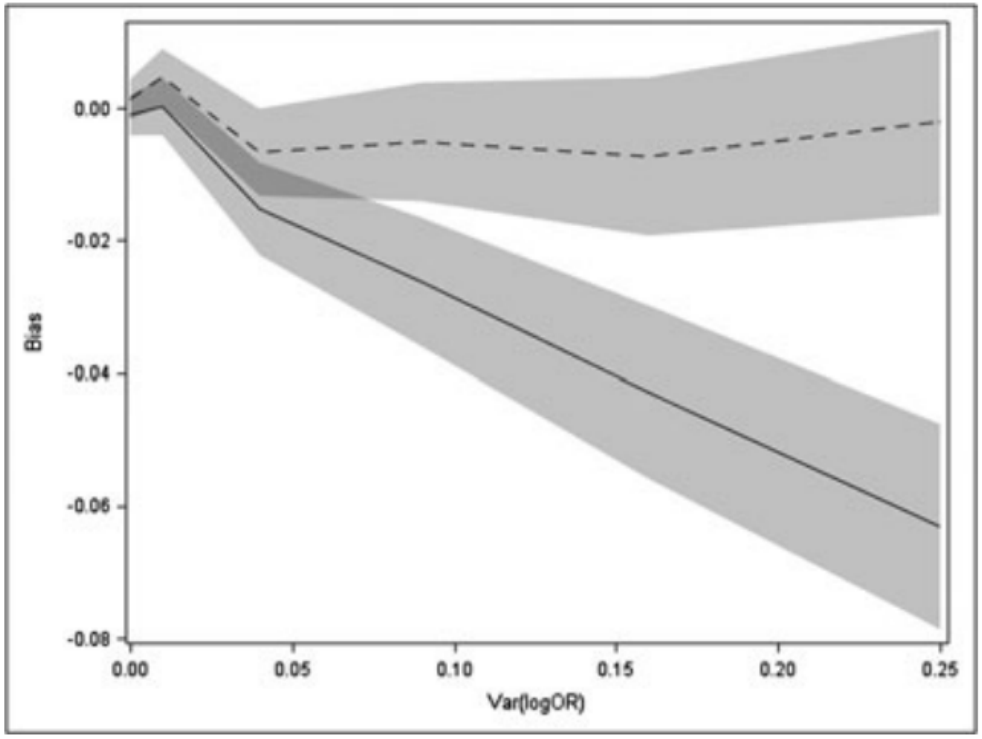


FIGURE 1 Trends in bias of two-stage (dashed line) and one-stage (continuous line) meta-analytic estimators (and 95% confidence bounds) according to increasing levels of between-database exposure effect heterogeneity (scenario 2). Bias is the difference between log $\left(\widehat{OR}_S\right)$, ie, estimated beta coefficient for exposure effect, and the true $\log(OR_{EO})$, ie, the assigned odds ratio for exposure-outcome

- **When the effect of interest is heterogeneous**, a **one-stage meta-analysis** ignoring clustering **gives biased estimates**.
- **Two-stage meta-analysis** generates estimates **at least as accurate and precise as one-stage meta-analysis**.

La Gamba et al., *Pharmacoepidemiology and Drug Safety* (2017)

6. Writing: Writing the study was very efficient

- Reuse of R code in CohortMethod, DatabaseConnector, SqlRender, EmpiricalCalibration, etc.
- Implementation took days instead of months
- Next study will be faster

Complexity is not a problem

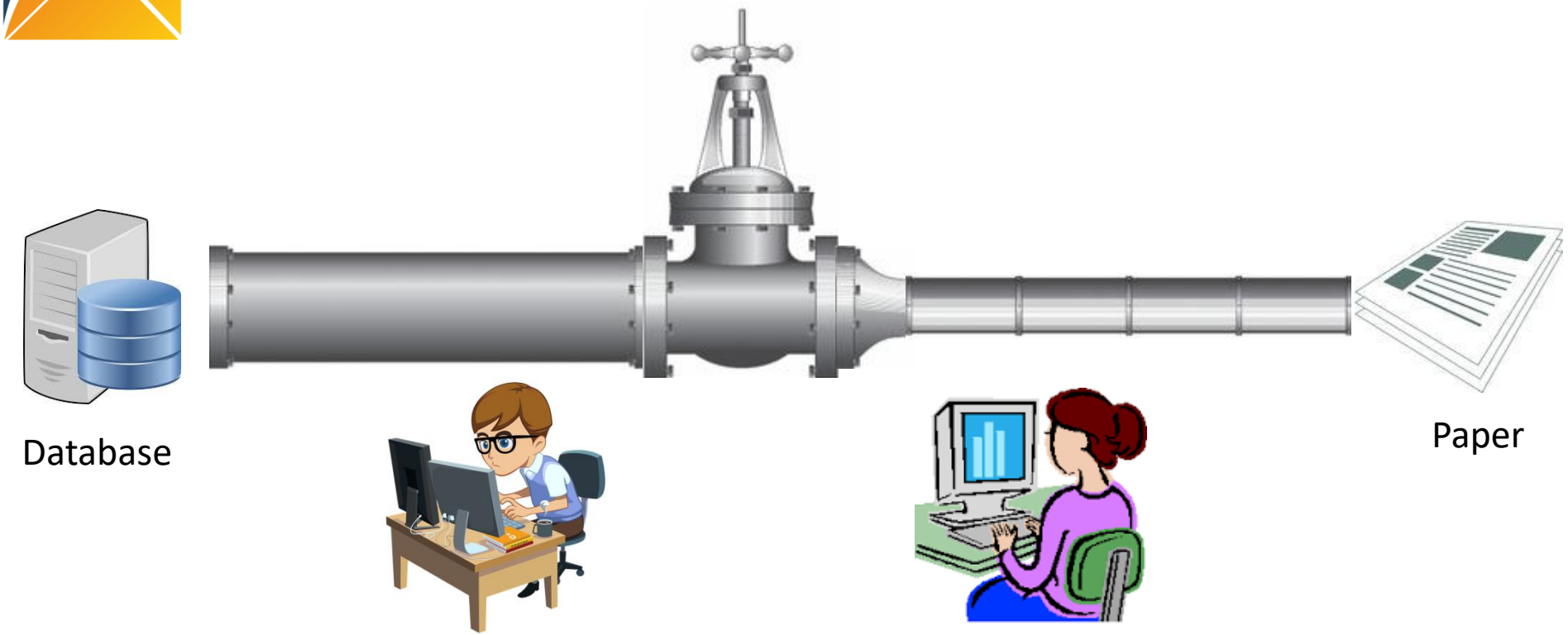
Use software engineering approaches to deal with complexity:

- Abstraction
- Encapsulation
- Writing clear code
- Re-use

Viewing a study as a pipeline has many advantages

- Full traceability
- Ability to check for correctness
- Ability to evaluate using controls
- More efficient
- Ability to deal with complexity
- Ability to work with several people on one analysis
- Easy to rerun on different data

What should OHDSI studies look like?



A study should be like a pipeline

- A fully automated process from database to paper
- ‘Performing a study’ = building the pipeline

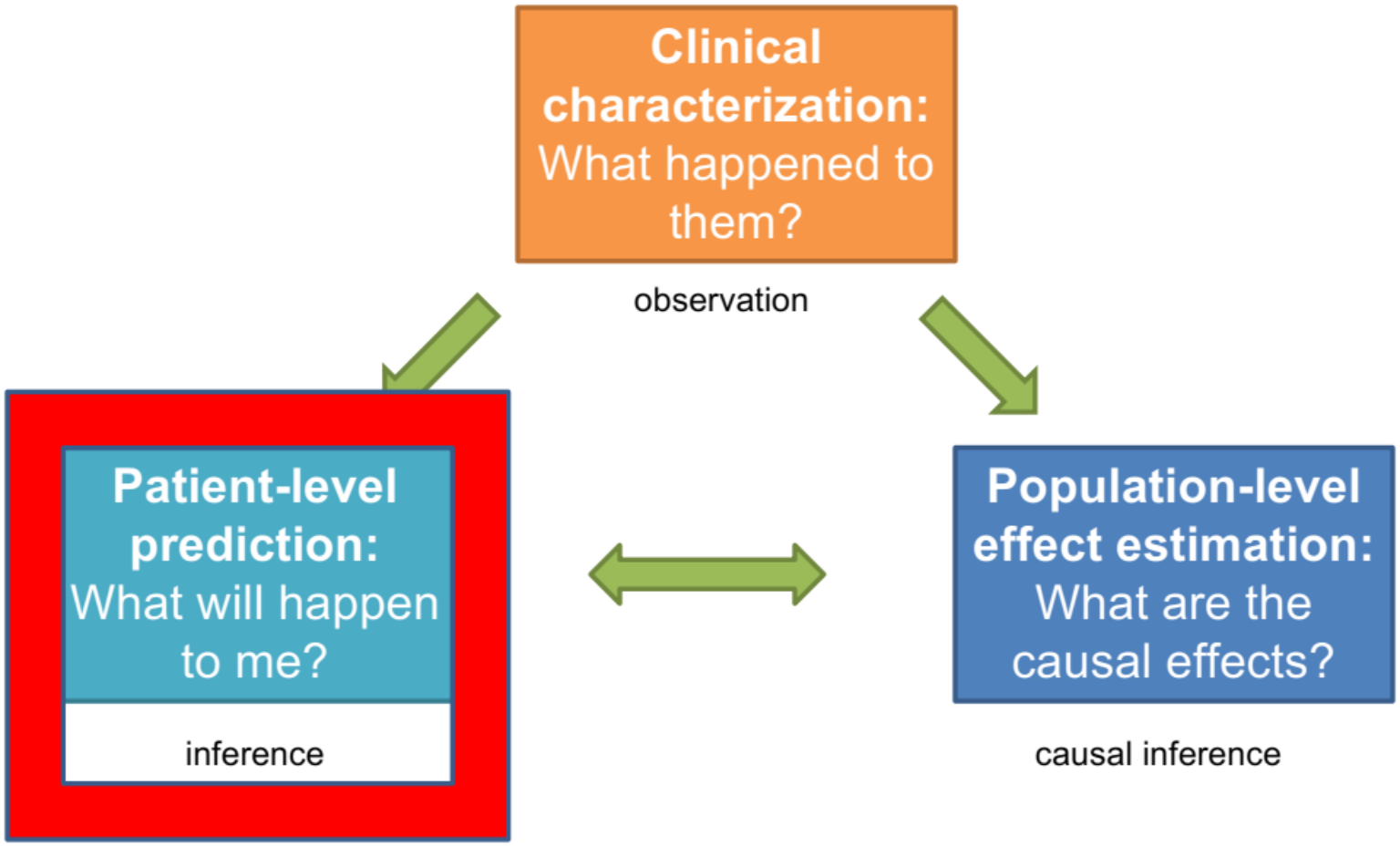
What evidence does OHDSI seek to generate from observational data?

- **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 hypoglycemia?
  - **Comparative effectiveness**: Does metformin cause hypoglycemia more than glyburide?
- **Patient-level prediction**
  - **Precision medicine**: Given everything you know about me and my medical history, if I start taking glyburide, what is the chance that I am going to have hypoglycemia during the first 30 days?
  - **Disease interception**: Given everything you know about me, what is the chance I will develop diabetes?

How to generate evidence by using cohorts? (target / comparator / outcome)

- **Patient-level prediction**
  - **Precision medicine**: Given everything you know about me and my medical history, if I start taking glyburide, what is the chance that I am going to have hypoglycemia during the first 30 days?
    - ➔Target cohort: DM patients using glyburide
    - ➔Outcome cohort: Patients developing hypoglycemia
  - **Disease interception**: Given everything you know about me, what is the chance I will develop diabetes?
    - ➔Target cohort: General population
    - ➔Outcome cohort: Patients developing DM

Complementary evidence to inform the patient journey



Current status of predictive modelling

Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review

RECEIVED 27 October 2015
REVISED 25 January 2016
ACCEPTED 20 February 2016

Benjamin A Goldstein[1,2], Ann Marie Navar[2,3], Michael J Pencina[1,2], John PA Ioannidis[4,5]

ABSTRACT

Objective Electronic health records (EHRs) are an increasingly common data source for clinical risk prediction, presenting both unique analytic opportunities and challenges. We sought to evaluate the current state of EHR based risk prediction modeling through a systematic review of clinical prediction studies using EHR data.

Methods We searched PubMed for articles that reported on the use of an EHR to develop a risk prediction model from 2009 to 2014. Articles were extracted by two reviewers, and we abstracted information on study design, use of EHR data, model building, and performance from each publication and supplementary documentation.

Results We identified 107 articles from 15 different countries. Studies were generally very large (median sample size = 26 100) and utilized a diverse array of predictors. Most used validation techniques ($n = 94$ of 107) and reported model coefficients for reproducibility ($n = 83$). However, studies did not fully leverage the breadth of EHR data, as they uncommonly used longitudinal information ($n = 37$) and employed relatively few predictor variables (median = 27 variables). Less than half of the studies were multicenter ($n = 50$) and only 26 performed validation across sites. Many studies did not fully address biases of EHR data such as missing data or loss to follow-up. Average c-statistics for different outcomes were: mortality (0.84), clinical prediction (0.83), hospitalization (0.71), and service utilization (0.71).

Conclusions EHR data present both opportunities and challenges for clinical risk prediction. There is room for improvement in designing such studies.

Current status of predictive modelling

- Inadequate internal validation
- Small sets of features
- Incomplete dissemination of model and results
- No transportability assessment
- Impact on clinical decision making unknown

Relatively few prediction models are used in clinical practice

OHDSI Mission for Patient-Level Prediction

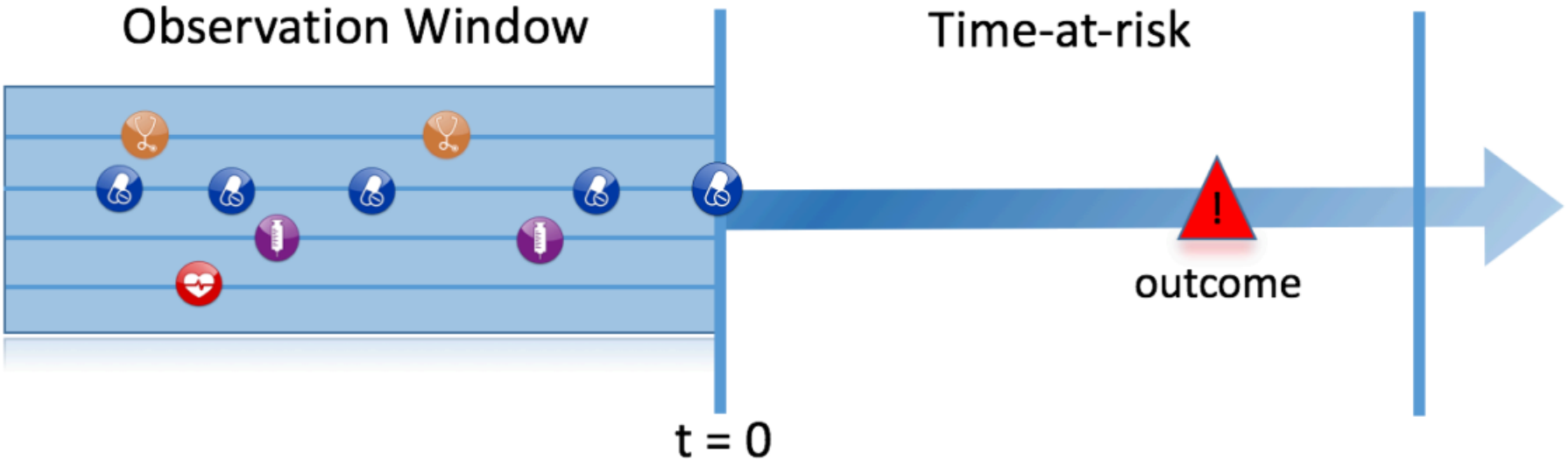
OHDSI aims to develop a systematic process to learn and evaluate large-scale patient-level prediction models using observational health data in a data network

Prediction Model Development

Problem pre-specification. A study protocol should unambiguously pre-specify the planned analyses.

Transparency. Others should be able to reproduce a study in every detail using the provided information. All analysis code should be made available as open source on the OHDSI Github.

Problem definition

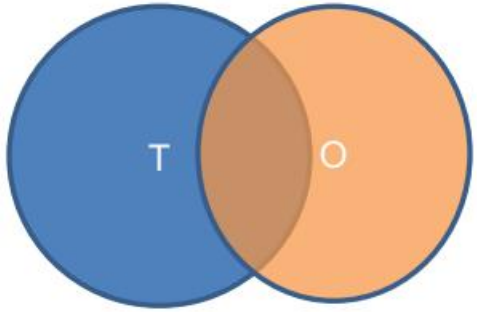


Among a target population (T), we aim to predict which patients at a defined moment in time (t=0) will experience some outcome (O) during a time-at-risk Prediction is done using only information about the patients in an observation window prior to that moment in time.

What are the key inputs to a patient-level prediction study?

Input parameter	Design choice
Target cohort (T)	
Outcome cohort (O)	
Time-at-risk	
Model specification -which model(s)? -which parameters? -which covariates?	

Prediction Model Development



We extract data for the patients in the Target Cohort (T) and we select all patients that experience the outcome (O)

The Target Cohort (T) and Outcome Cohort (O) can be defined using ATLAS or custom code (see later today).

For model development all outcomes (O) of patients in the Target Cohort (T) are used.

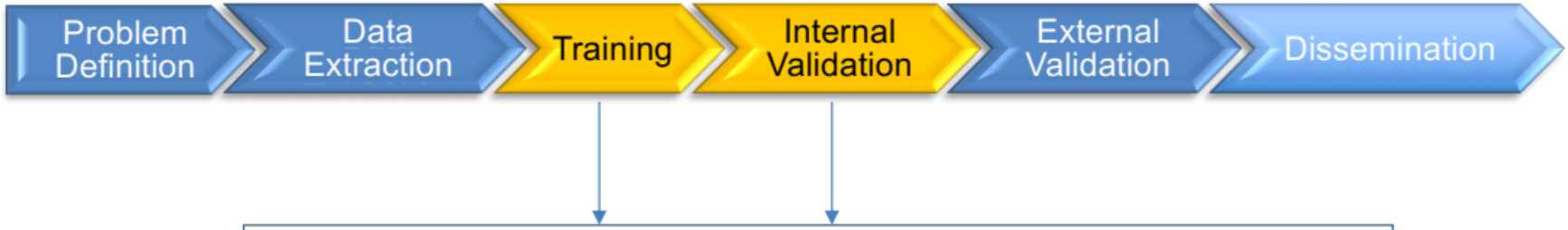
Prediction Model Development

Age	Gender	Var 1	...	Var n	Stroke
34	M	1	0	1	1
50	F	0	0	0	1
24	M	0	1	0	0
67	F	1	1	0	0
34	F	0	0	0	0
65	M	1	0	0	1
34	F	0	0	0	0
40	M	0	1	1	0

Data is extracted from the OMOP CDM using the Feature Extraction R-Package.

This allows for specification of the candidate predictors and time windows.

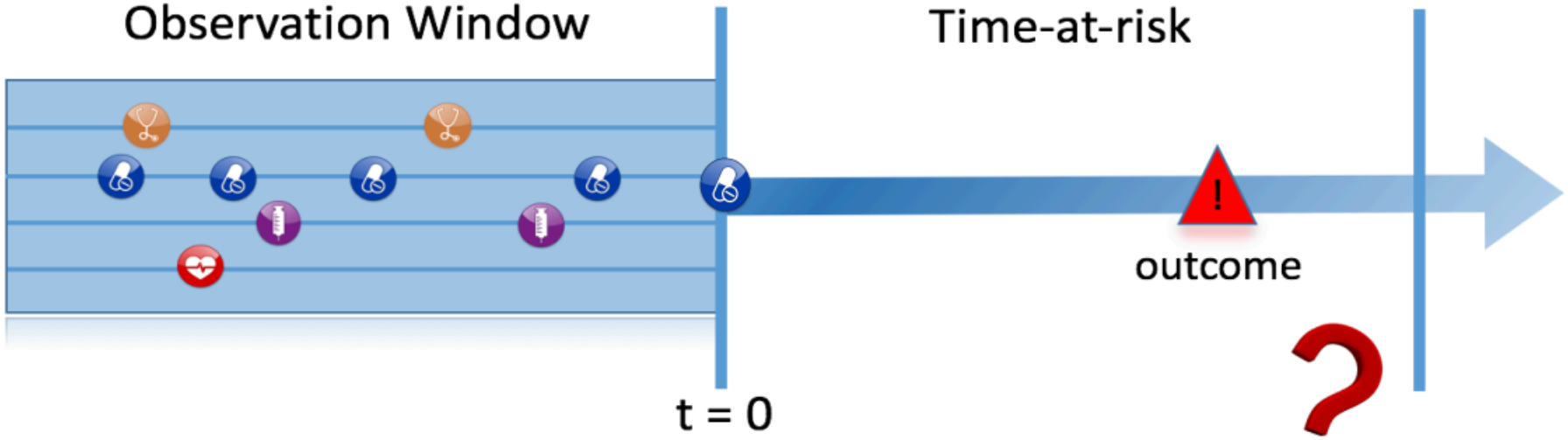
Prediction Model Development



Model training and **Internal validation** is done using a train test split:

1. Person split: examples are assigned randomly to the train or test set, or
2. Time split: a split is made at a moment in time (temporal validation)

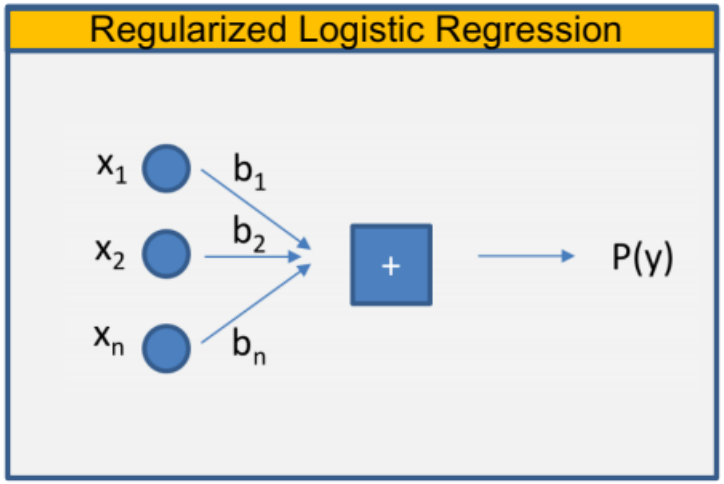
Model Training

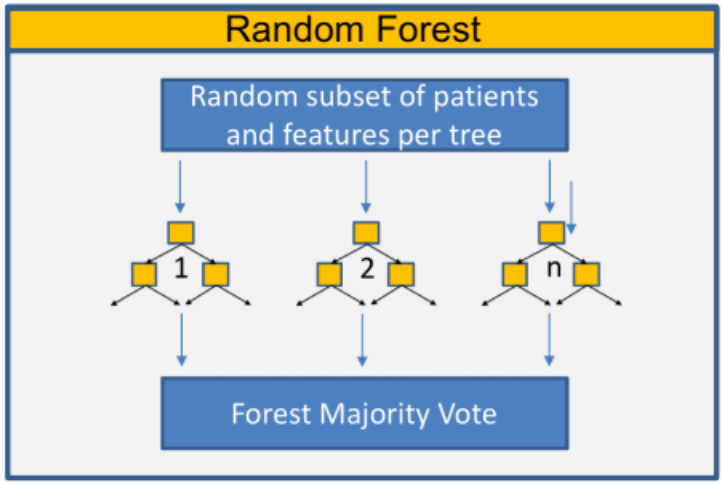


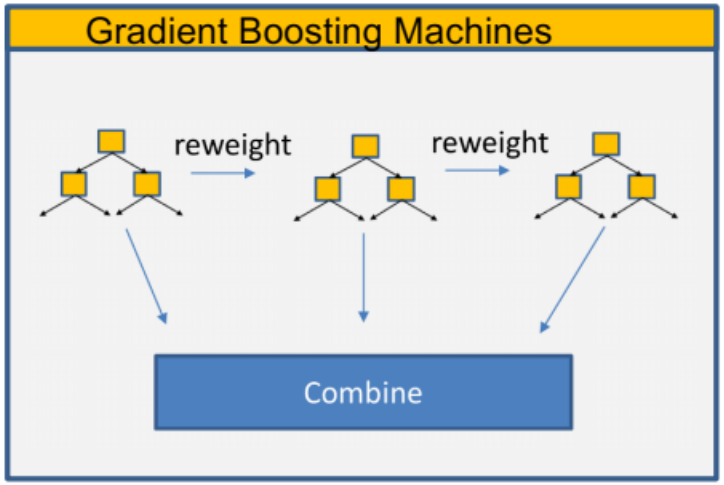
1. Which models?

2. How to evaluate the model?

Models and Algorithms







Many other models for example:

K-nearest neighbors
Naïve Bayes
Decision Tree
Adaboost
Neural Network
Deep Learning
Etc.

Model selection is an empirical process

The **"No Free Lunch" theorem** states that there is not one model that works best for every problem. The assumptions of a great model for one problem may not hold for another problem.

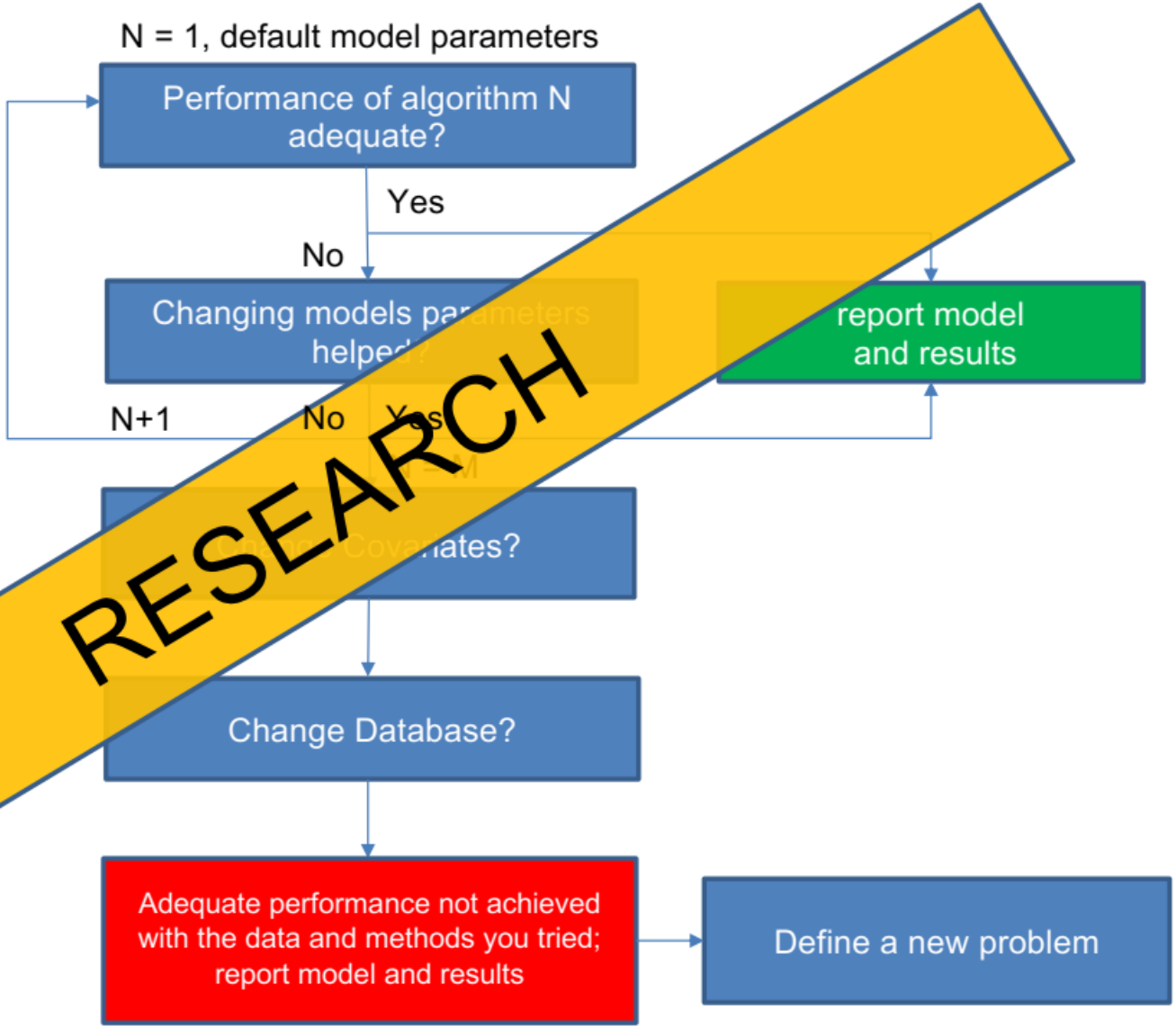
It is common in machine learning to try multiple models and find one that works best for that particular problem.

see no-free-lunch.org

OHDSI Model Selection Strategy

Suggested ordering of available algorithms in PLP package

N	Algorithms
1.	Lasso Logistic Regression
2.	Random Forest
3.	Gradient Boosting Machine
4.	Neural Network
5.	KNN
M.	…



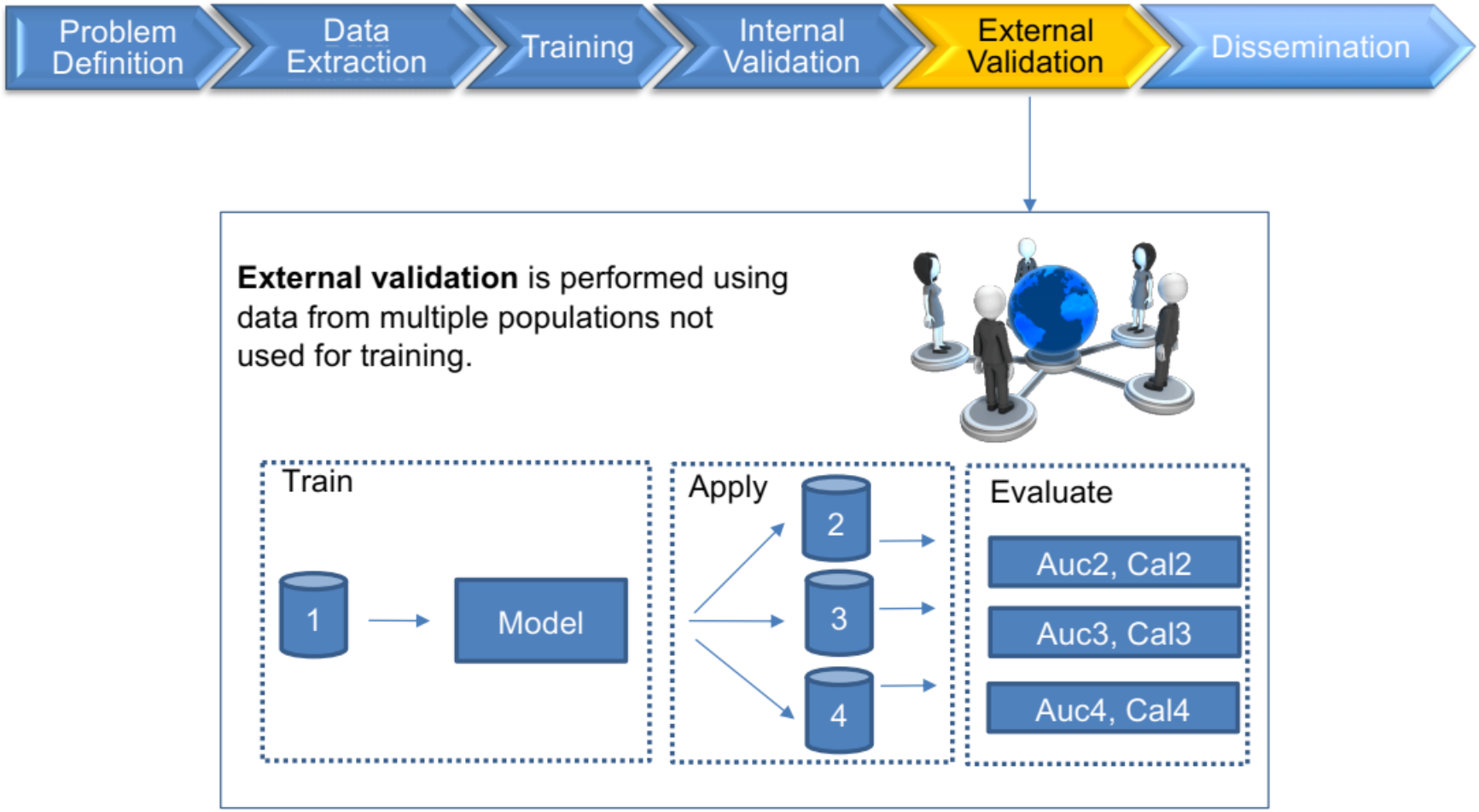
Model Validation

What makes a good model?

Discrimination: differentiates between those with and without the event, i.e. predicts higher probabilities for those with the event compared to those who don't experience the event

Calibration: estimated probabilities are close to the observed frequency

External Validation



Dissemination

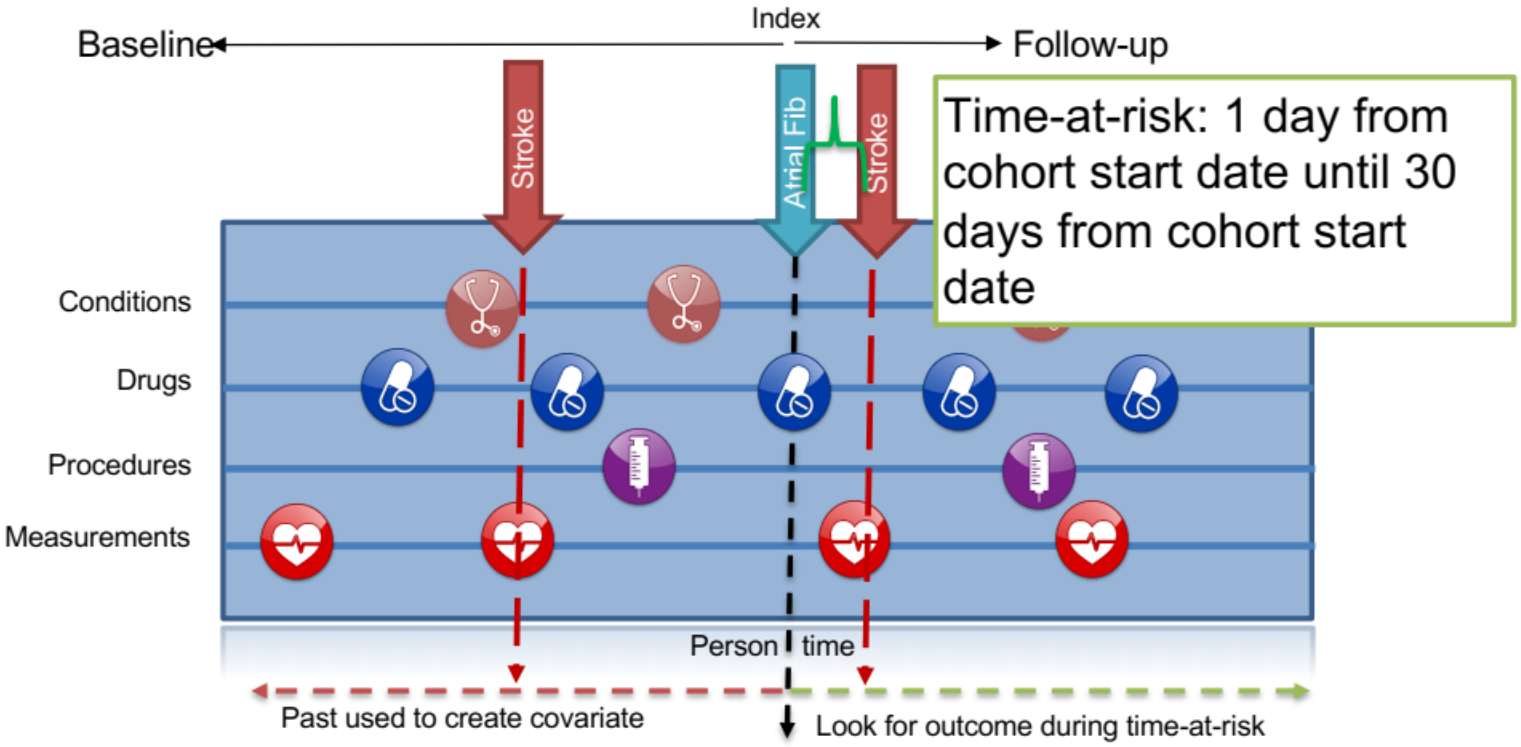
Dissemination of study results should follow the minimum requirements as stated in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [1].

- Internal and external validation
- Sharing of full model details
- Sharing of all analyses code to allow full reproducibility

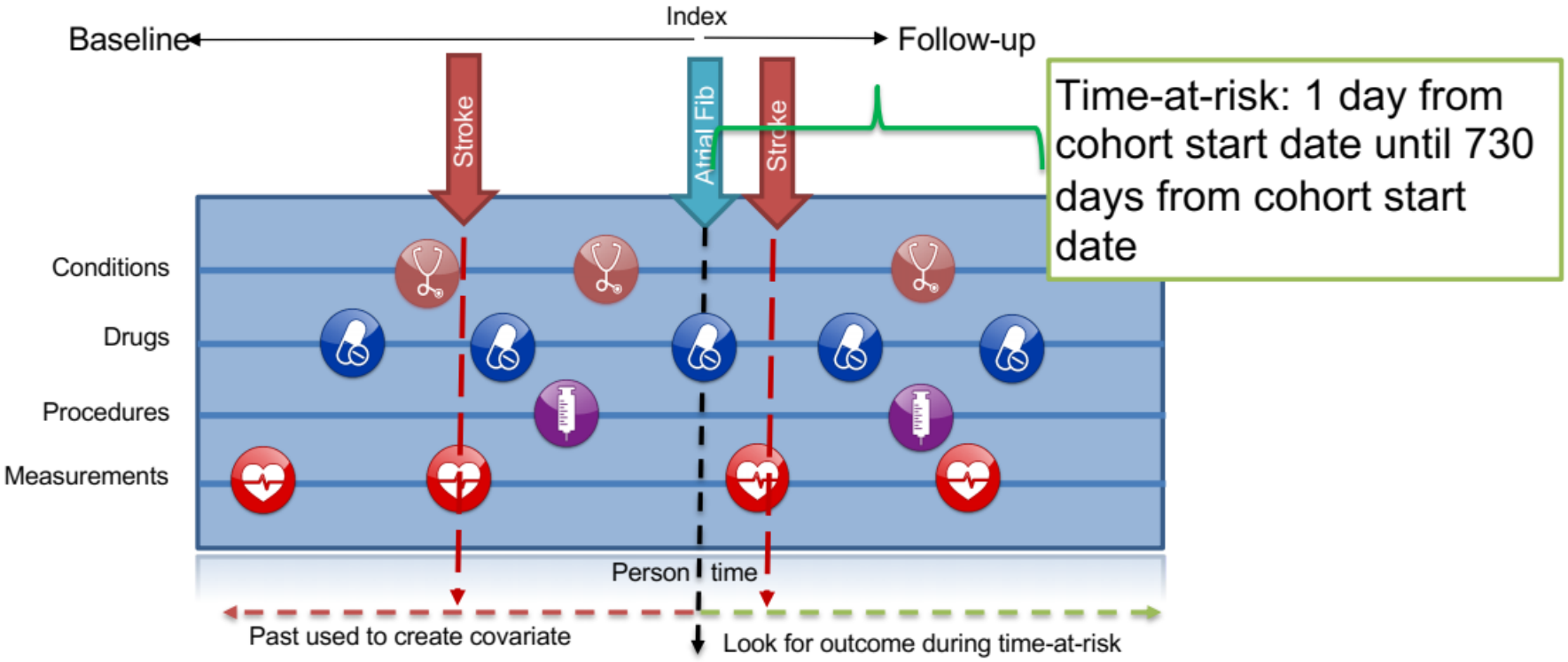
Website to share protocol, code, models and results for all databases

[1] Moons, KG et al. Ann Intern Med. 2015;162(1):W1-73

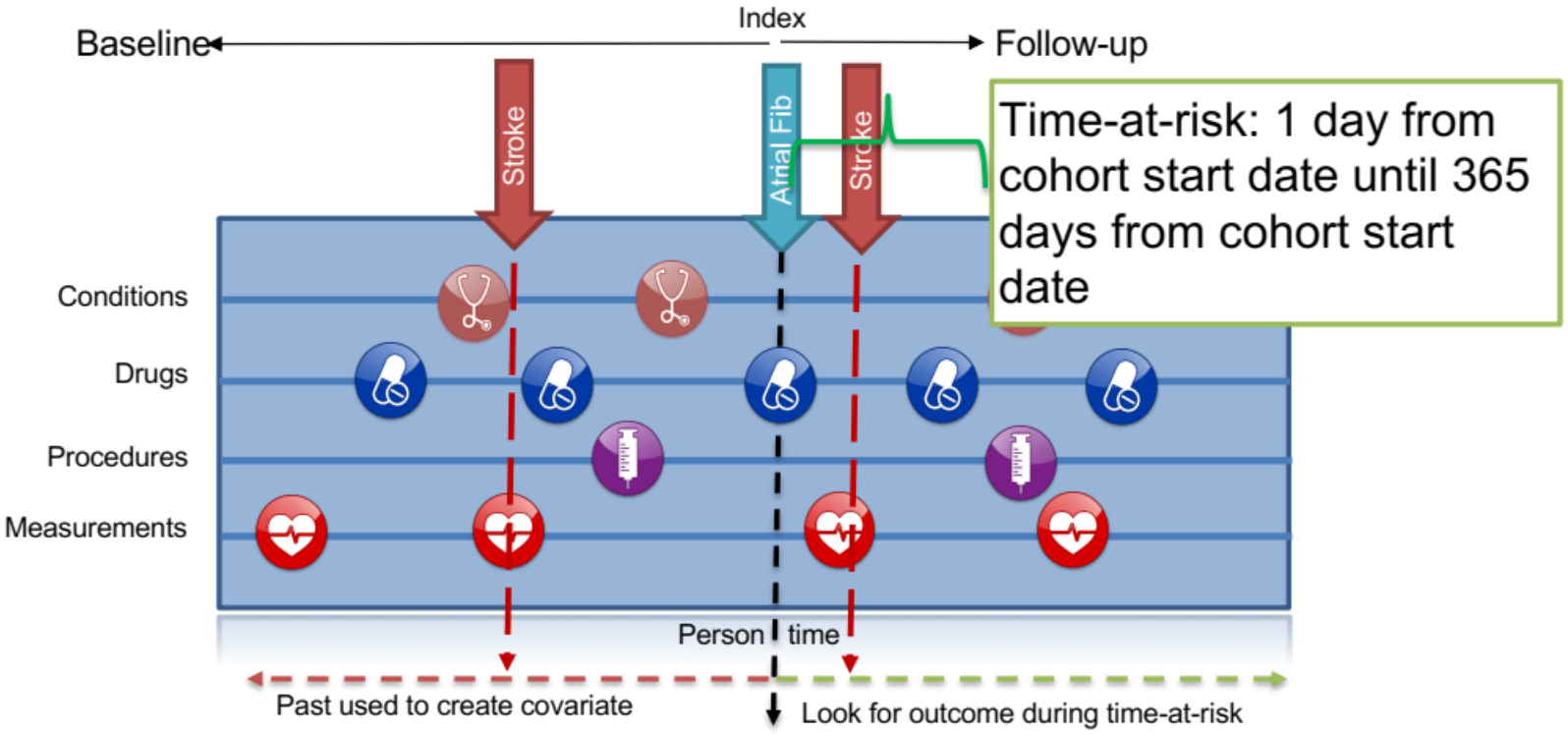
Extracting Labelled Data



Extracting Labelled Data



Extracting Labelled Data



Extracting Labelled Data

Each person corresponds to a row

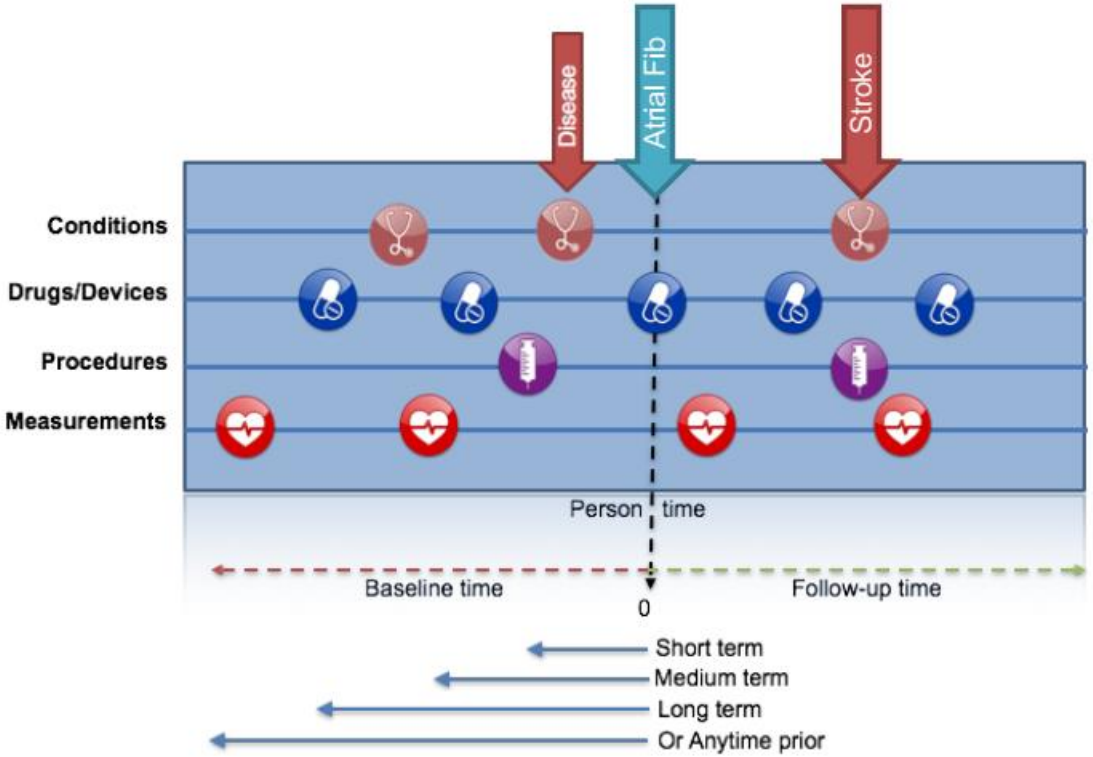
Labelled classification data

Subject_id	Cohort start date	Has outcome during TAR
3454102	2012-01-02	1 (Yes)
105454	2012-08-12	0 (No)
…		…

This gives us our labels for each subject!

Now Use Baseline to Construct Covariates

We create standard features using records prior to the target cohort start date (e.g., the atrial fibrillation)



Covariates

- Can pick three time periods and anytime prior to index (include index is an option)
- Binary indicator variables for conditions, drugs, procedures, measurements and observations
- Values for measurements
- Can use hierarchy to create binary indicators for a code and all children code (grouped covariates)
- Includes record type counts
- Includes some common risk scores
- Can add custom variables

Extracting Labelled Data

We create the covariates using the baseline for each subject

Labelled classification data

Subject_id	Condition A	Condition B	…	Drug N	Has outcome during TAR
3454102	1	1	…	0	1 (Yes)
105454	1	0	…	1	0 (No)
…	…	…	…	…	…

This gives us our label data for each subject!

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior outcome
3454102	1	2012-01-02	0	
105454	1	2012-08-12	0	
1554	1	2009-05-05	0	
56566	1	2011-07-05	0	Yes (-35 days and -999 days)
4346356	1	2011-07-05	1	
342424	1	2010-01-01	1	Yes (-370 days)
…	…	…		

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior outcome
3454102	1	2012-01-02	0	
105454	1	2012-08-12	0	
1554	1	2009-05-05	0	
~~56566~~	~~1~~	~~2011-07-05~~	~~0~~	~~Yes (35 days and~~ 999 days)
4346356	1	2011-07-05	1	
~~342424~~	~~1~~	~~2010-01-01~~	~~1~~	~~Yes (370 days)~~
…	…	…		

- Remove patients who have observed the outcome prior to cohort entry? **[YES]**
- How many days to look back from cohort entry for the outcome? **[99999]** days prior to cohort start

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior outcome
3454102	1	2012-01-02	0	
105454	1	2012-08-12	0	
1554	1	2009-05-05	0	
56566	1	2011-07-05	0	Yes (35 days and - 999 days)
4346356	1	2011-07-05	1	
342424	1	2010-01-01	1	Yes (-370 days)
…	…	…		

- Remove patients who have observed the outcome prior to cohort entry? **[YES]**
- How many days to look back from cohort entry for the outcome? **[365]** days prior to cohort start

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior outcome
3454102	1	2012-01-02	0	
105454	1	2012-08-12	0	
1554	1	2009-05-05	0	
56566	1	2011-07-05	0	Yes (-35 days and -999 days)
4346356	1	2011-07-05	1	
342424	1	2010-01-01	1	Yes (-370 days)
…	…	…		

- Remove patients who have observed the outcome prior to cohort entry? **[No]**

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR
3454102	1	2012-01-02	0
105454	*1*	*2012-08-12*	*0*
105454	*1*	*2013-10-04*	*0*
1554	1	2009-05-05	0
56566	1	2011-07-05	0
4346356	1	2011-07-05	1
342424	1	2010-01-01	1
…	…	…	

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR
3454102	1	2012-01-02	0
105454	*1*	*2012-08-12*	*0*
105454	*1*	*2013-10-04*	*0*
1554	1	2009-05-05	0
56566	1	2011-07-05	0
4346356	1	2011-07-05	1
342424	1	2010-01-01	1
…	…	…	

Should only the first exposure per subject be included? **[YES]**

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR
3454102	1	2012-01-02	0
105454	*1*	*2012-08-12*	*0*
105454	*1*	*2013-10-04*	*0*
1554	1	2009-05-05	0
56566	1	2011-07-05	0
4346356	1	2011-07-05	1
342424	1	2010-01-01	1
…	…	…	

Should only the first exposure per subject be included? **[No]**

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior observation
3454102	1	2012-01-02	0	366
105454	1	2012-08-12	0	2009
1554	1	2009-05-05	0	1098
56566	1	2011-07-05	0	365
4346356	1	2011-07-05	1	4056
342424	1	2010-01-01	1	588
…	…	…		

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior observation
3454102	1	2012-01-02	0	366
105454	1	2012-08-12	0	2009
1554	1	2009-05-05	0	1098
56566	1	2011-07-05	0	365
4346356	1	2011-07-05	1	4056
342424	1	2010-01-01	1	588
…	…	…		

Minimum lookback period applied to target cohort: **[730]**

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior observation
105454	1	2012-08-12	0	2009
4346356	1	2011-07-05	1	4056
…	…	…		

Minimum lookback period applied to target cohort: **[1200]**

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior observation
3454102	1	2012-01-02	0	366
105454	1	2012-08-12	0	2009
1554	1	2009-05-05	0	1098
56566	1	2011-07-05	0	365
4346356	1	2011-07-05	1	4056
342424	1	2010-01-01	1	588
…	…	…		

Minimum lookback period applied to target cohort: **[365]**

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Imaging your cohort looks like this:

TAR (time-at-risk) is 1 day to 365 days after cohort start date

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Follow-up observation
3454102	1	2012-01-02	0	50
105454	1	2012-08-12	0	1082
1554	1	2009-05-05	0	366
56566	1	2011-07-05	0	480
4346356	1	2011-07-05	1	40
342424	1	2010-01-01	1	500
…	…	…		

Imaging your cohort looks like this:

TAR (time-at-risk) is 1 day to 365 days after cohort start date

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Follow-up observation
~~3454102~~	~~1~~	~~2012-01-02~~	~~0~~	~~50~~
105454	1	2012-08-12	0	1082
1554	1	2009-05-05	0	366
56566	1	2011-07-05	0	480
~~4346356~~	~~1~~	~~2011-07-05~~	~~1~~	~~40~~
342424	1	2010-01-01	1	500
…	…	…		

- Should subjects without time at risk be removed? **[YES]**
- Minimum time at risk: **[364]** days
- Include people with outcomes who are not observed for the whole at risk period? **[NO]**

Imaging your cohort looks like this:

TAR (time-at-risk) is 1 day to 365 days after cohort start date

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Follow-up observation
~~3454102~~	~~1~~	~~2012-01-02~~	~~0~~	~~50~~
105454	1	2012-08-12	0	1082
1554	1	2009-05-05	0	366
56566	1	2011-07-05	0	480
4346356	1	2011-07-05	1	40
342424	1	2010-01-01	1	500
…	…	…		

- Should subjects without time at risk be removed? **[YES]**
- Minimum time at risk: **[364]** days
- Include people with outcomes who are not observed for the whole at risk period? **[YES]**

Imaging your cohort looks like this:

TAR (time-at-risk) is 1 day to 365 days after cohort start date

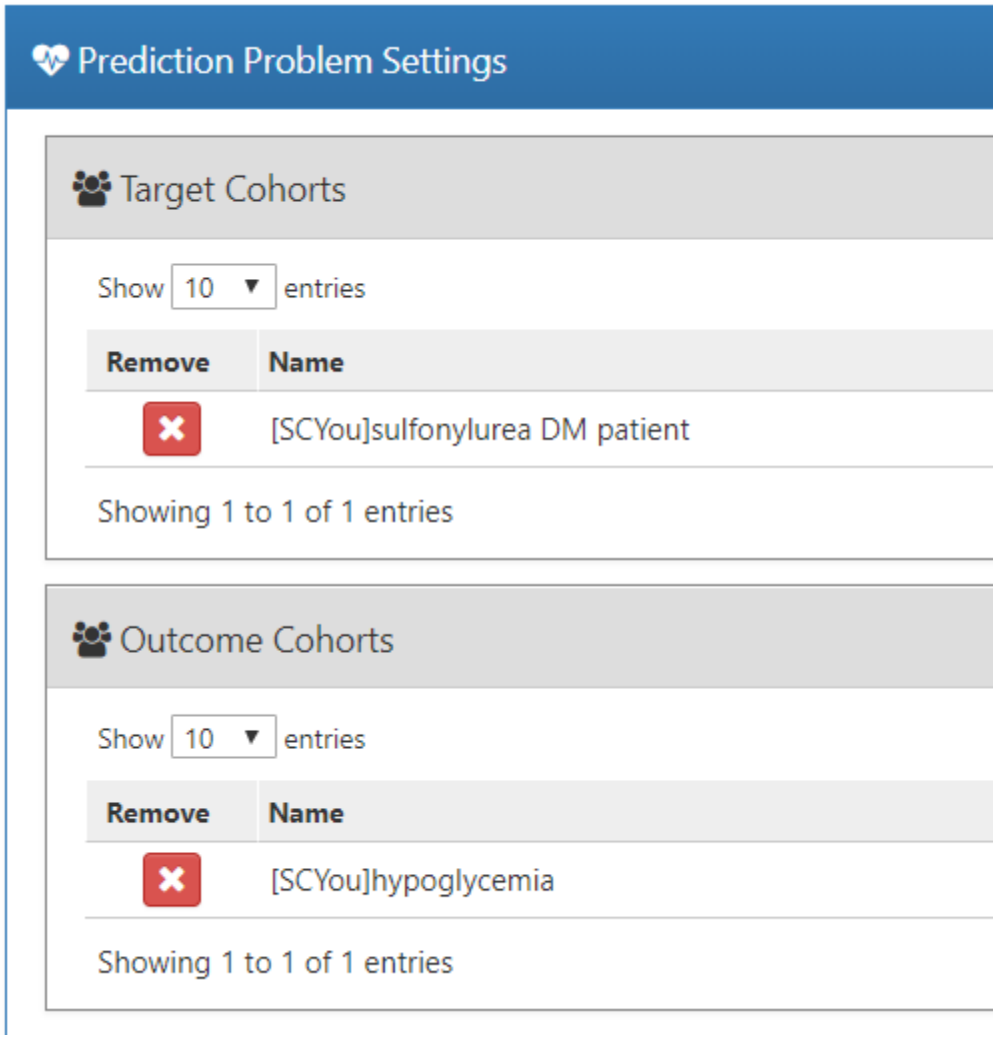
Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Follow-up observation
3454102	1	2012-01-02	0	50
105454	1	2012-08-12	0	1082
1554	1	2009-05-05	0	366
56566	1	2011-07-05	0	480
4346356	1	2011-07-05	1	40
342424	1	2010-01-01	1	500
…	…	…		

- Should subjects without time at risk be removed? **[No]**
- Minimum time at risk: **[1]** days
- Include people with outcomes who are not observed for the whole at risk period? **[No]**

Design an machine-learning algorithm

Input parameter	Design choice
Target cohort (T)	Sulfonylurea user
Outcome cohort (O)	Hypoglycemia
Time-at-risk	1~90 days
Model specification	

Prediction Problem Settings



Model and Covariates Settings

Model Settings

Show 10 entries

Remove	Model	Options
✖	RandomForestSettings	{"mtries":[-1],"ntrees":[500],"maxDepth":[4,10,17],"varImp":[true],"seed":null}
✖	MLPSettings	{"size":[4],"alpha":[0.00001],"seed":null}
✖	LassoLogisticRegressionSettings	{"variance":0.01,"seed":null}

Showing 1 to 3 of 3 entries

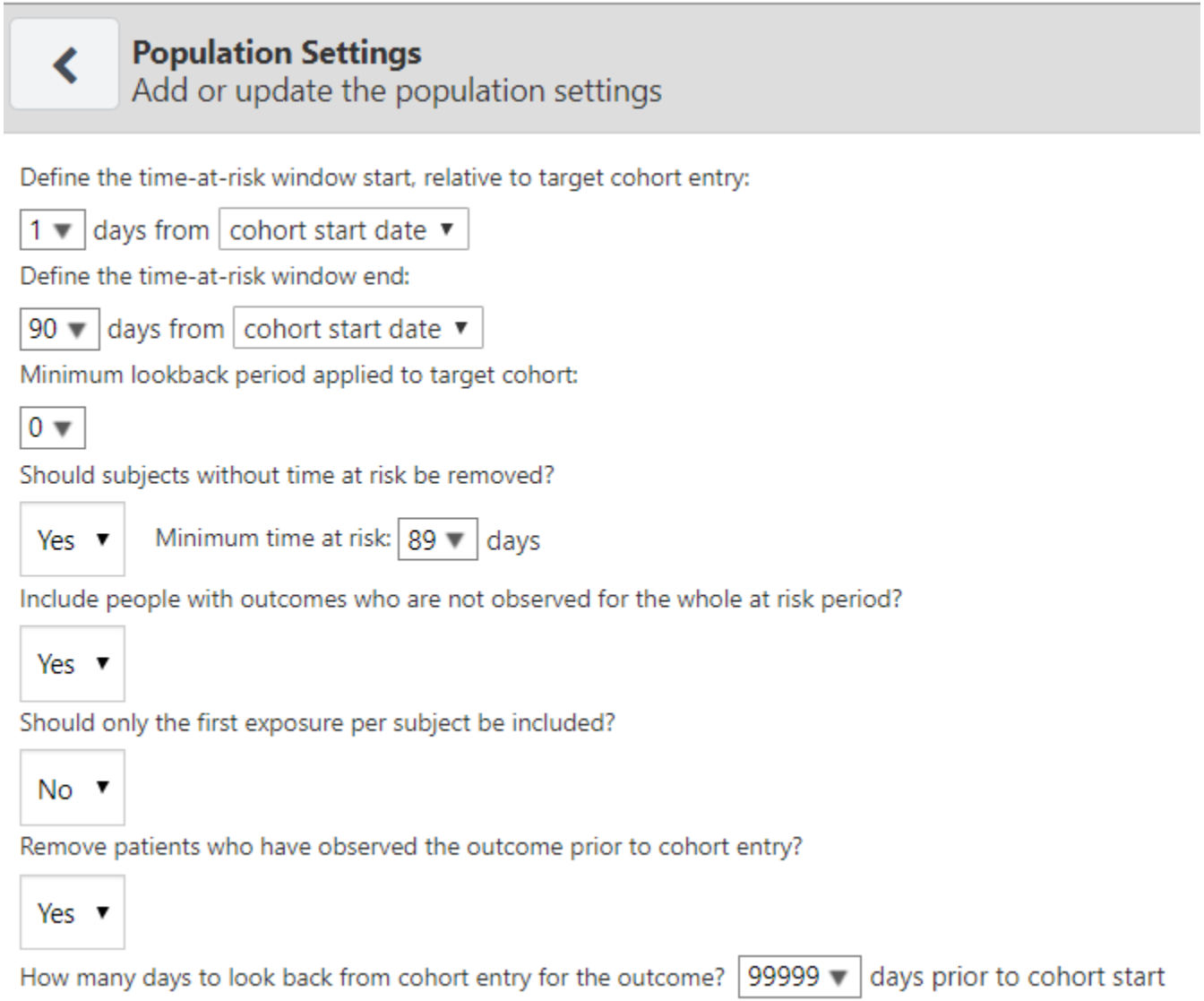
Covariate Settings

Column visibility | Copy | CSV | Show 10 entries

Remove	Options
✖	DemographicsGender, DemographicsAgeGroup, DemographicsRace, ConditionGroupEraLongTerm, DrugGroupEraLongTerm, ProcedureOccurrenceLongTerm (+1 more covariate settings)

Showing 1 to 1 of 1 entries

Population Settings



Summary:

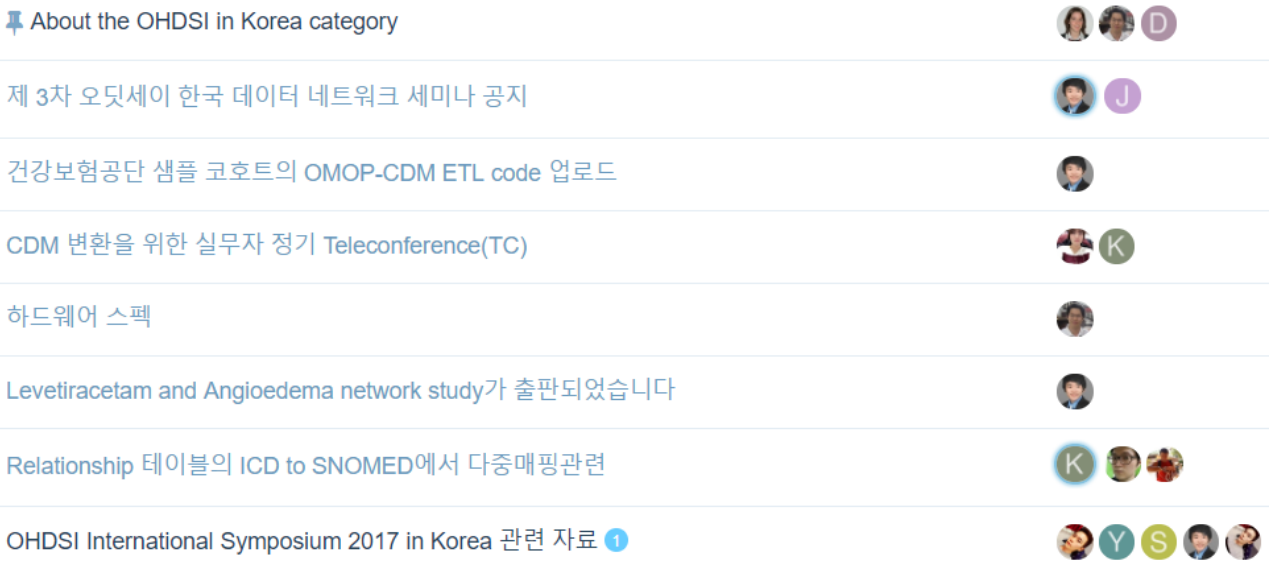
- Need to define prediction problem
- Need to define the target population and outcome cohorts
- Need to specify covariate settings
- Need to specify population settings – this modifies target population and creates labels

Please, Join the Journey

- Main homepage
  - www.ohdsi.org
- OHDSI-KOREA Homepage
  - www.ohdsi-korea.org
- OHDSI Community meeting: 1AM (Korean time), Wed
  - http://www.ohdsi.org/web/wiki/doku.php?id=projects:ohdsi_community
- Workgroup meeting
  - Eastern hemisphere meeting: 4PM (Korean time), Wed,
    - Population-Level estimating workgroup: http://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:est-methods

Please, Join the Journey

- Forum
  - http://forums.ohdsi.org/
- OHDSI in Korea Forum
  - http://forums.ohdsi.org/c/For-collaborators-wishing-to-communicate-in-Korean



Goal of the DataThon

- *행복한 가정은 모두 비슷한 이유로*
  행복하지만 불행한 가정은 저마다의 이유로
  불행하다 (Все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива по-своему)

- 구동이 되는 패키지는 모두 비슷한 이유로 작동하지만, 에러가 나는 패키지는 **모두 저마다의 문제**가 있다.

➔**구동이 되는** PLE or PLP package를 **완성**하고, **결과**를 **발표**, **GitHub Upload**

https://github.com/ohdsi-korea/OhdsiDataThonKorea2019


Thank You
for your time