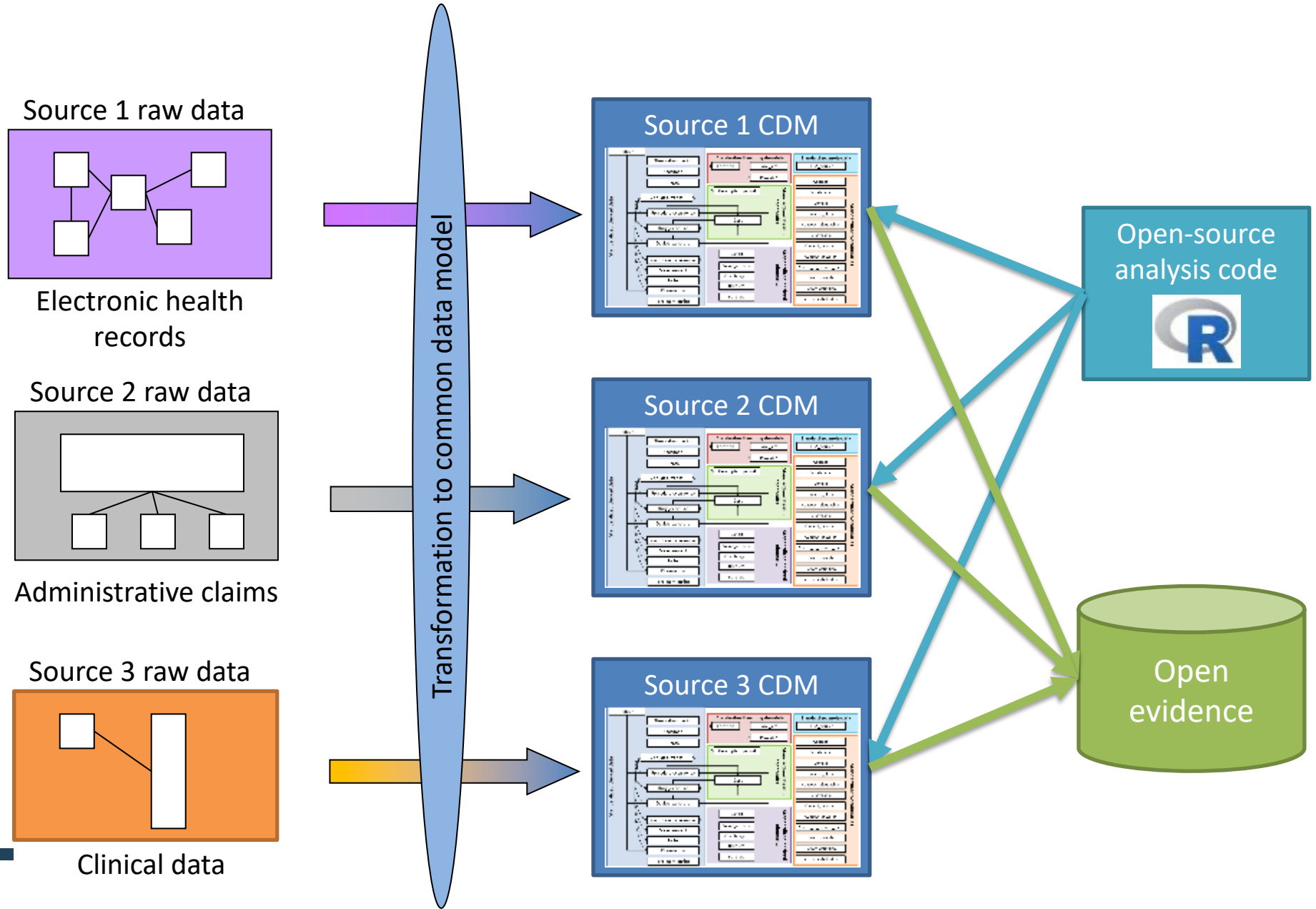
Cohort Building

Patrick Ryan, PhD

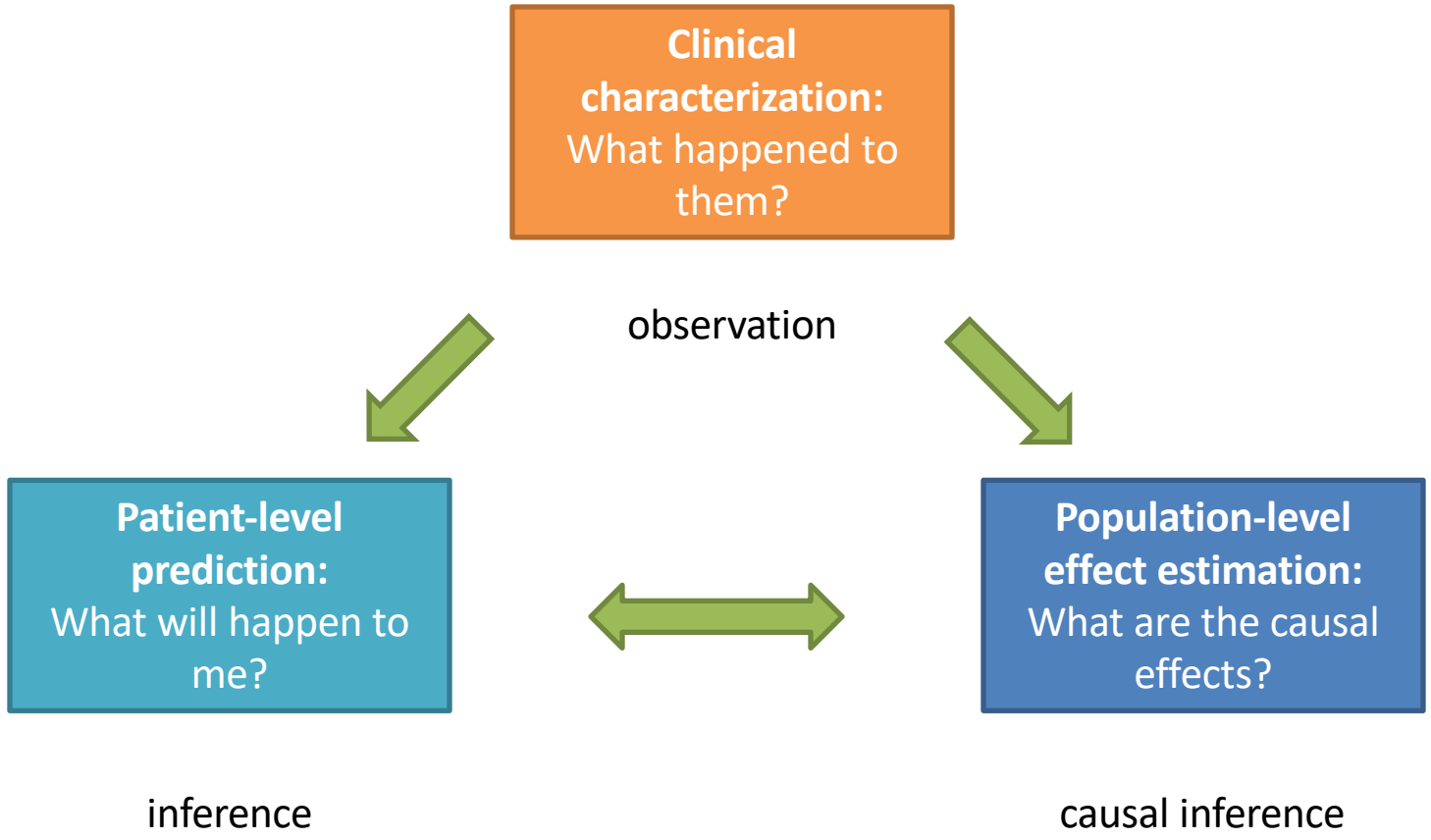
Vice President, Observational Health Data Analytics, Janssen Research and Development

Assistant Professor, Adjunct, Department of Biomedical Informatics, Columbia University Medical Center

Common data model can enable standardized analytics across a distributed data network



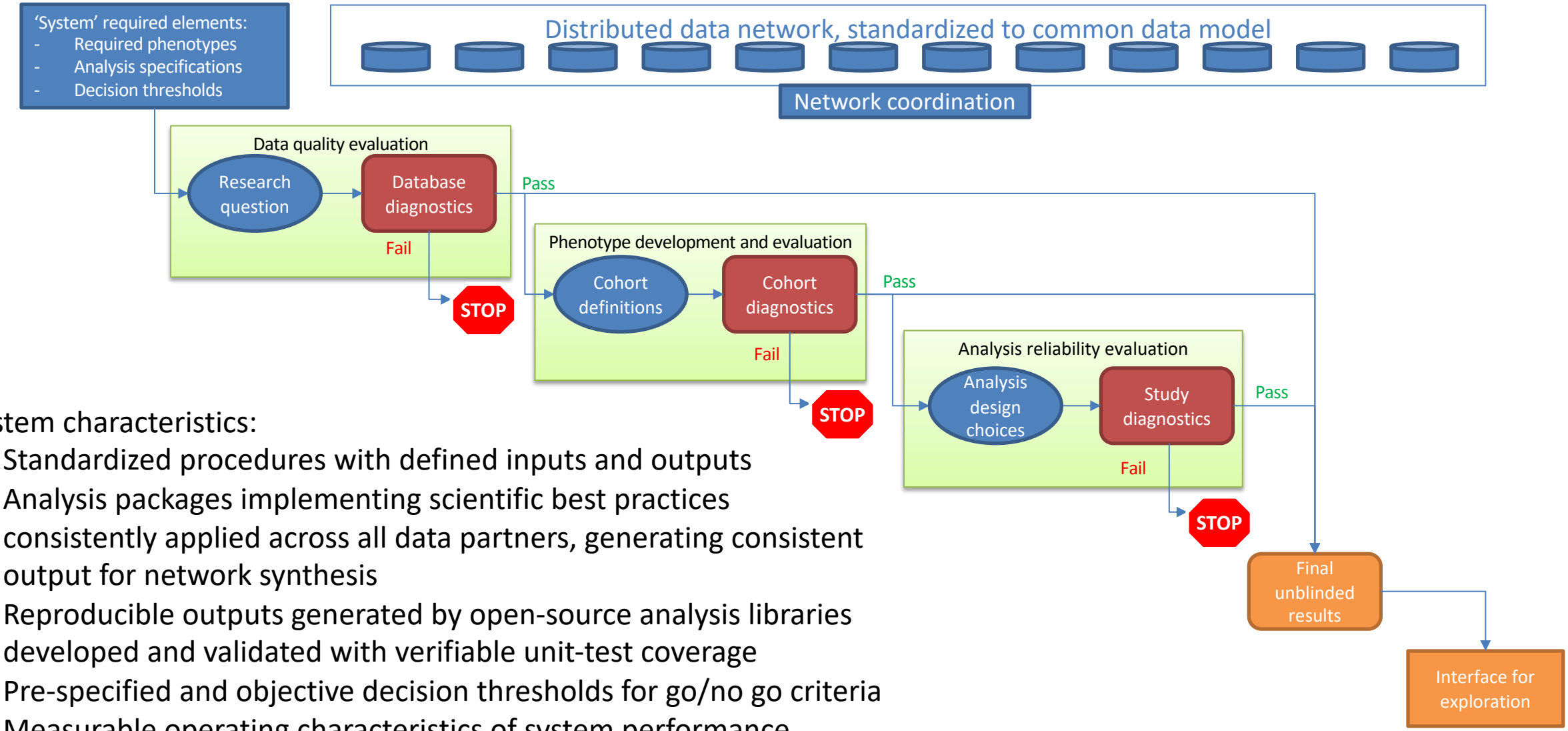
Complementary evidence to inform the patient journey



Standardizing the question makes it possible to standardize the analysis and standardize the evidence

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with <insert disease of interest>, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have <insert disease of interest>, which treatments were patients exposed to amongst <list of treatments for disease> and in which sequence?
	Outcome incidence	Amongst patients who are new users of <insert drug of interest> among the population with <insert indication of interest>, how many patients experienced <insert outcome of interest> within <time horizon following exposure start>?
Population-level effect estimation	**Safety surveillance**	Does exposure to <insert drug of interest> increase the risk of experiencing <insert an adverse event> within <time horizon following exposure start>, among the population with <insert indication of interest>?
	Comparative effectiveness	Does exposure to <insert drug of interest> have a different risk of experiencing <insert any outcome (safety or benefit) > within <time horizon following exposure start>, relative to <insert comparator treatment>, among the population with <insert indication of interest>?
Patient level prediction	**Disease onset and progression**	For a given patient who is diagnosed with <insert your favorite disease>, what is the probability that they will go on to have <another disease or related complication> within <time horizon from diagnosis>?
	Treatment response	For a given patient who is a new user of <insert drug of interest> for <insert indication of interest>, what is the probability that they will <insert desired effect> in <time window>?
	Treatment safety	For a given patient who is a new user of <insert drug of interest> for <insert indication of interest>, what is the probability that they will experience <insert adverse event> within <time horizon following exposure>?

Engineering open science systems that build trust into the real-world evidence generation and dissemination process

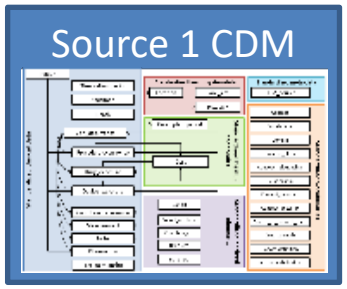


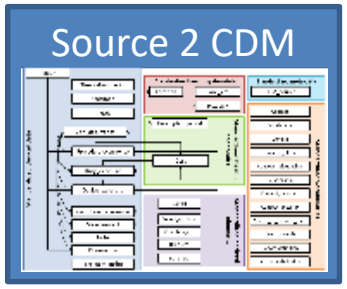
System characteristics:

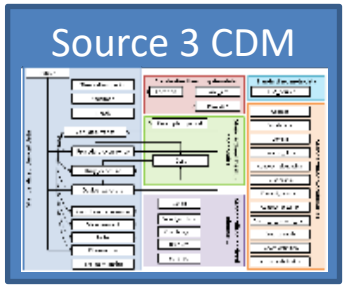
- Standardized procedures with defined inputs and outputs
- Analysis packages implementing scientific best practices consistently applied across all data partners, generating consistent output for network synthesis
- Reproducible outputs generated by open-source analysis libraries developed and validated with verifiable unit-test coverage
- Pre-specified and objective decision thresholds for go/no go criteria
- Measurable operating characteristics of system performance

The journey to evidence

Standardized data







Cohort definition: a specification to identify the set of persons satisfying one or more criteria for a duration of time

Standardized analytics

T, {E} → Treatment pathways

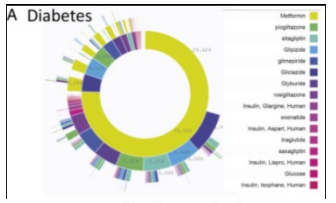
T, O → Incidence rate analysis

T, C, I, O → Comparative cohort design

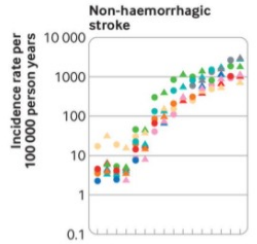
T, I, O → Self-controlled case series

T, O → Patient-level prediction

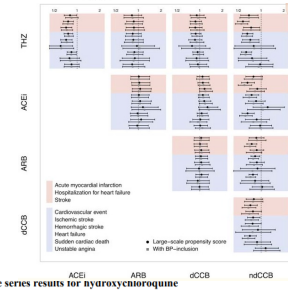
Impactful results



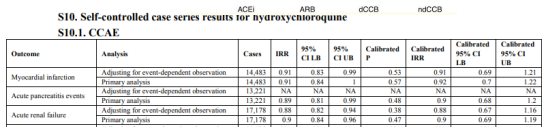
Hripcsak et al
PNAS 2016



Li et al
BMJ 2021



Suchard et al
Lancet 2019



S10. Self-controlled case series results for hydroxychloroquine

S10.1. CCAE

Outcome	Analysis	Cases	IRR	95% CI LB	95% CI UB	Calibrated P	Calibrated IRR	Calibrated 95% CI LB	Calibrated 95% CI UB
Myocardial infarction	Adjusting for event-dependent observation	14,483	0.91	0.83	0.99	0.53	0.91	0.69	1.21
	Primary analysis	14,483	0.91	0.84	1	0.57	0.92	0.7	1.22
Acute pancreatitis events	Adjusting for event-dependent observation	13,221	NA	NA	NA	NA	NA	NA	NA
	Primary analysis	13,221	0.89	0.81	0.99	0.48	0.9	0.68	1.2
Acute renal failure	Adjusting for event-dependent observation	17,178	0.88	0.82	0.94	0.38	0.88	0.67	1.16
	Primary analysis	17,178	0.9	0.84	0.96	0.47	0.9	0.69	1.19

Lane et al Lancet Rheumatology 2020

Williams et al
BMC MRM 2022

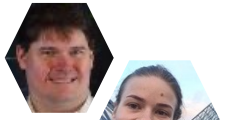
OHDSI's definition of 'cohort'

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

Cohort era = a continuous period during which a person has satisfied a cohort's inclusion criteria

Cohort definition = the specification for how to identify a cohort

OHDSI open-source community tools to support phenotype development and evaluation process




Phenotype definition tools:

- ATLAS
  - Concept set expressions – with recommendations from PHOEBE2.0
  - Cohort Definitions – to design a rule-based cohort definition
  - Profiles – to review individual cases

- CapR - cohort definition application programming in R, to design rule-based cohort definitions consistent with CIRCE JSON specifications

- APHRODITE - to develop a probabilistic phenotype by training a prediction model using noisy labels

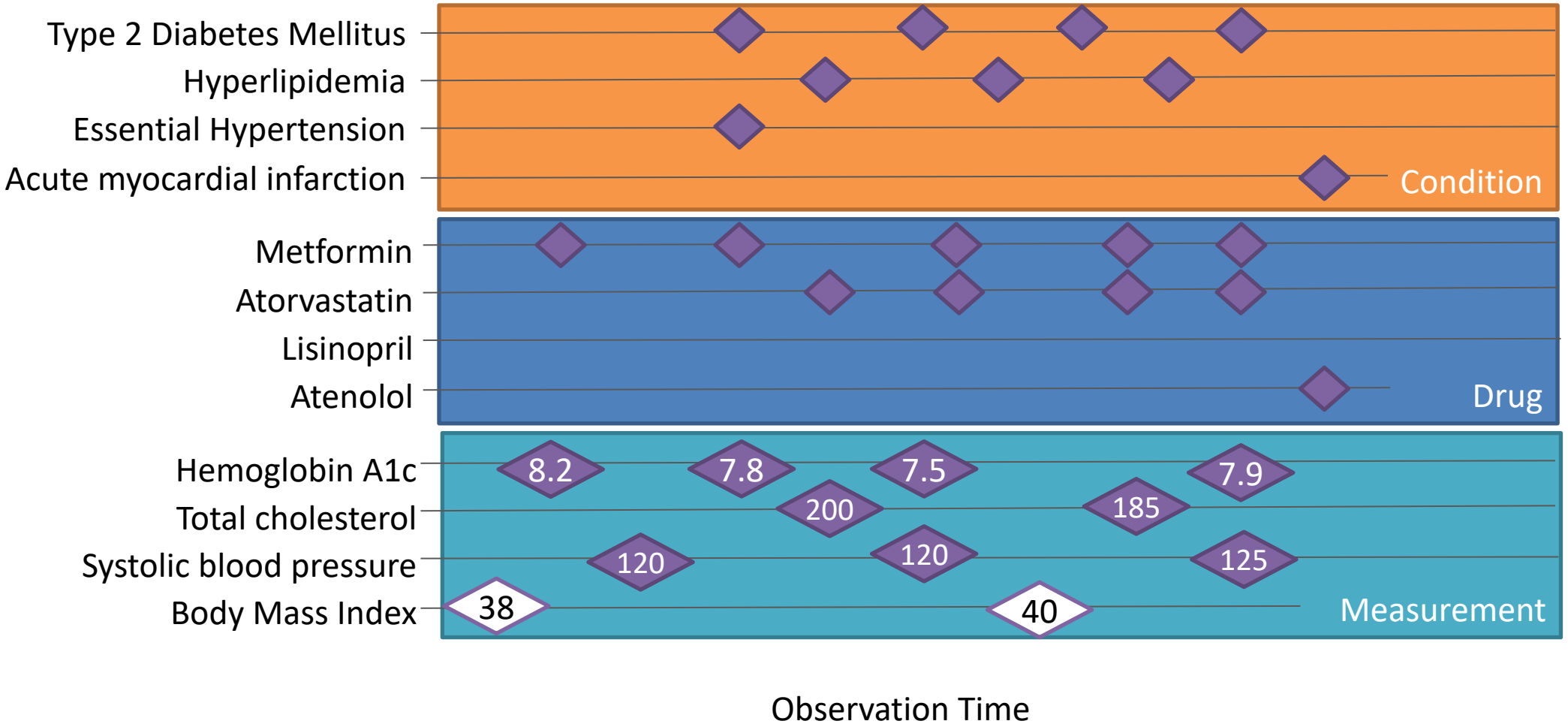


Phenotype evaluation tools:

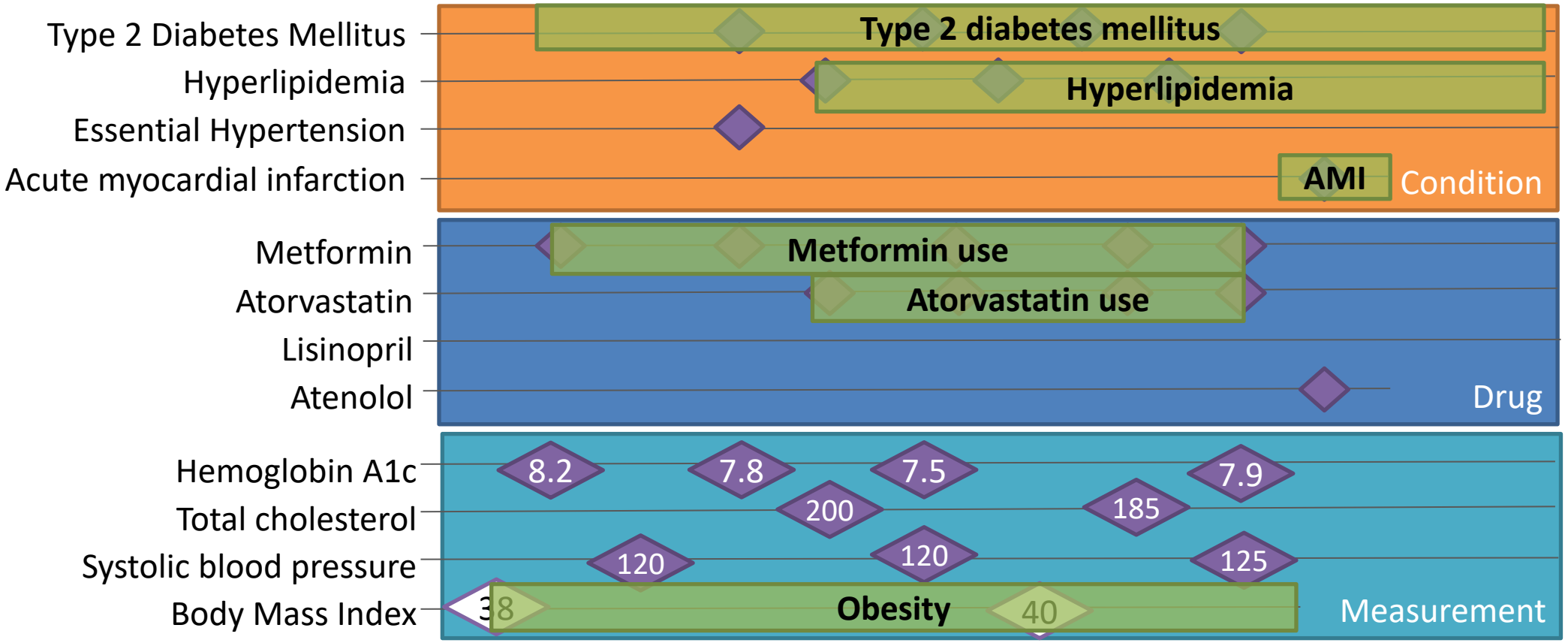
- CohortExplorer – to review individual cases
- CohortDiagnostics – to evaluate phenotype algorithms using population-level characterization to identify sensitivity/specificity errors and index date misspecification
- PheValuator - to evaluate a phenotype algorithm (estimate sensitivity/specificity/PPV) by training a prediction model and creating a probabilistic reference standard

Phenotype Library

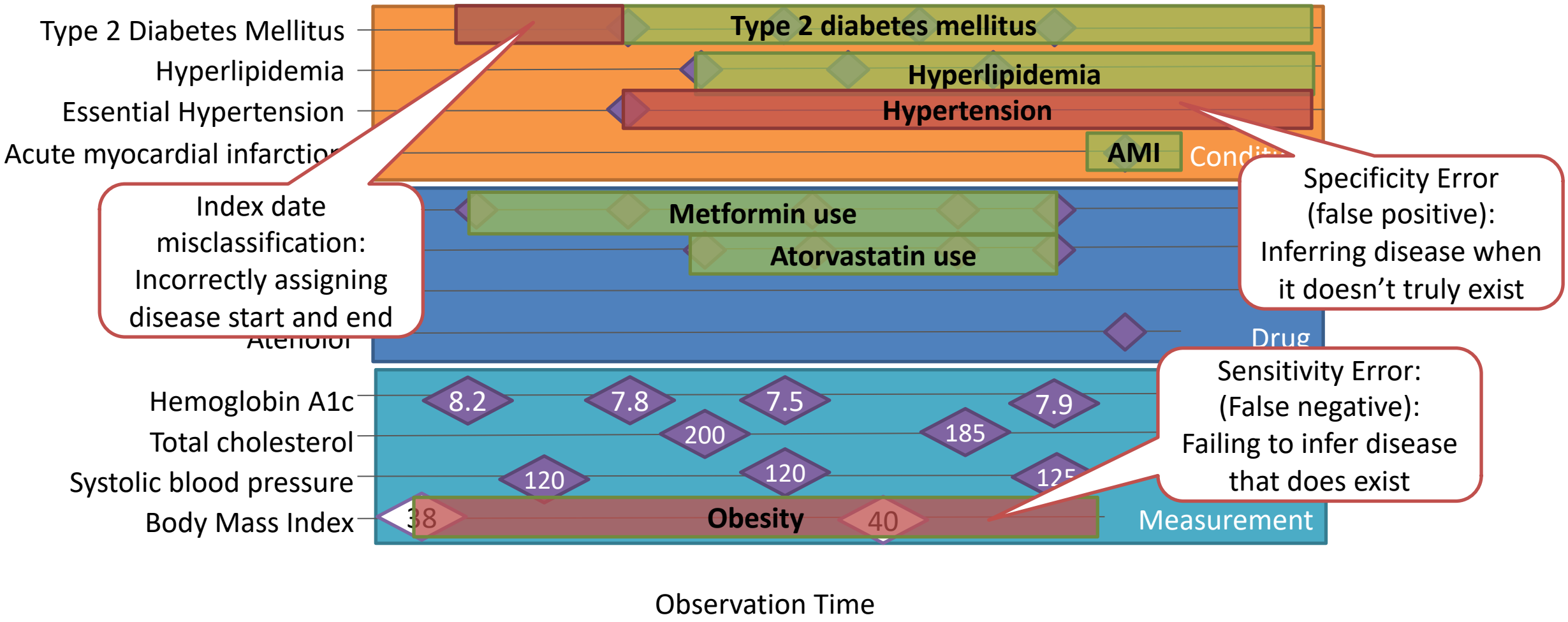
What we *HAVE*?
Observational data for a single person



What we *WANT*?
Longitudinal health status for a single person



Potential errors from inference in disease phenotyping



Evaluating phenotypes

- Objective: estimate the extent to which the inference from the phenotype algorithm consistent with the true health state of the patients?
- Measurement error measures:
  - Sensitivity, specificity, positive predictive value, negative predictive value
- 'A phenotype is fit-for-use' = The measurement error of the phenotype in the dataset is sufficiently small that it will not negatively impact the interpretation of analysis results

Steps for developing phenotypes with evaluation in mind

1. Identify the persons who might have the disease
   - Aim: Increase sensitivity
   - Task: Create inclusive conceptsets used in cohort entry events
2. Restrict persons who likely do not have disease
   - Aim: Increase specificity / positive predictive value
   - Task: Add inclusion criteria
3. Determine the start and end dates for each disease episode
   1. Aim: Reduce index date misspecification
   2. Task: Set exit strategy, refine entry events and inclusion criteria

OHDSI's definition of 'cohort'

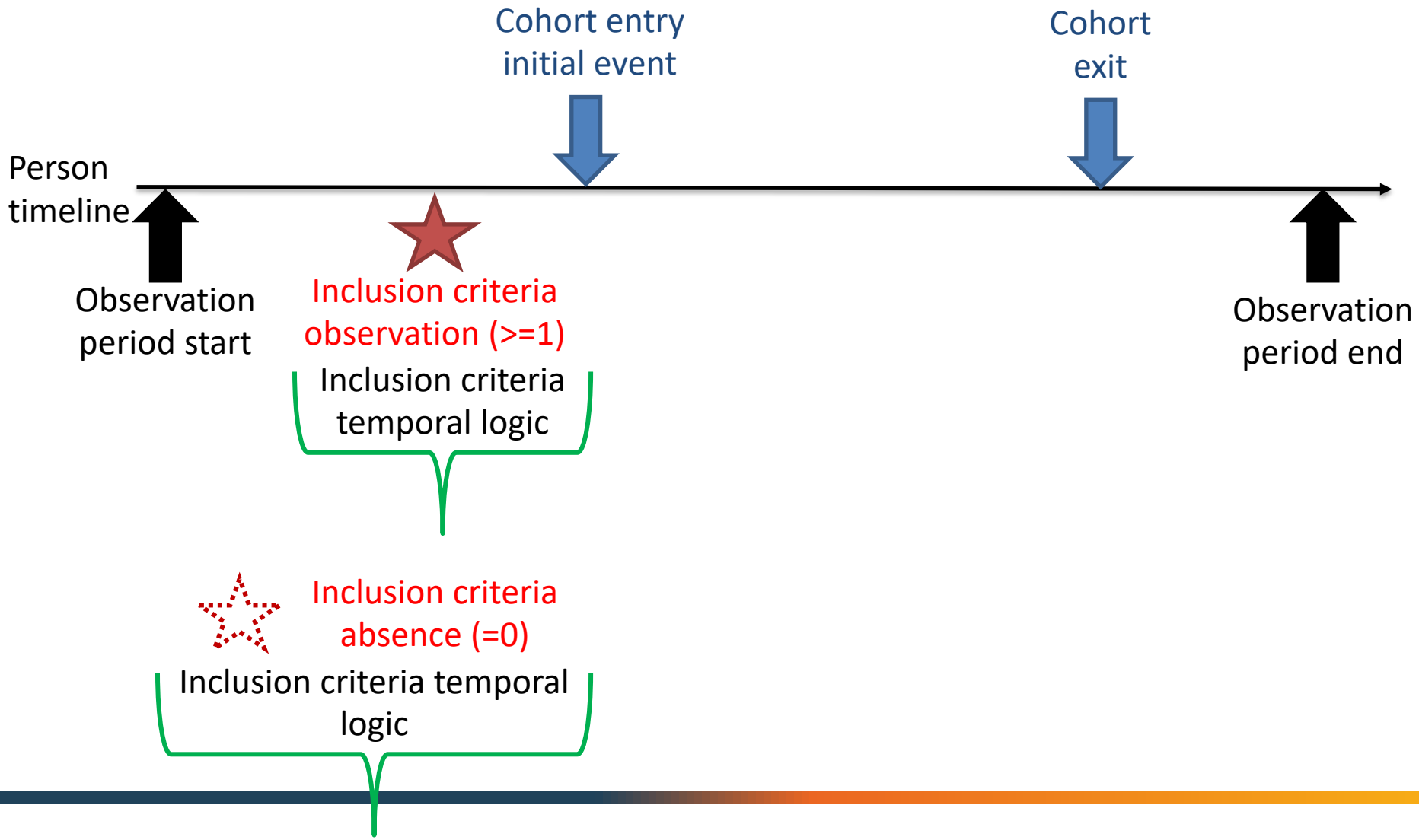
Cohort = a set of persons who satisfy one or more inclusion criteria for a duration of time

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

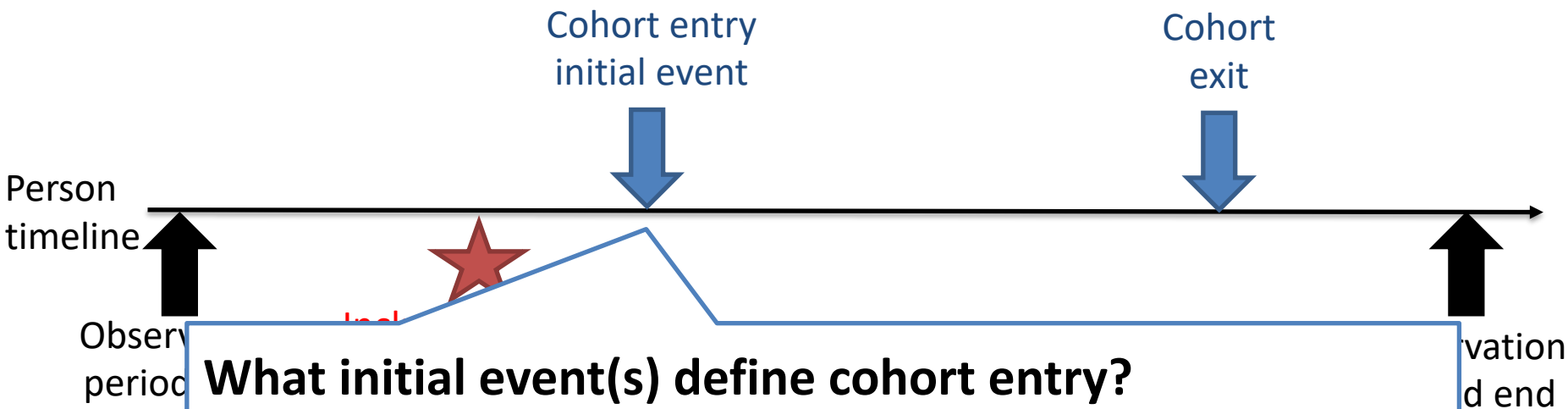
Dissecting the anatomy of a cohort definition



Questions to answer when defining a cohort

- What event(s) let you enter the cohort?
- What inclusion criteria are applied to those events?
- For each event, how long do you satisfy the inclusion criteria?
- How should events be combined into cohort eras?

Dissecting the anatomy of a cohort definition



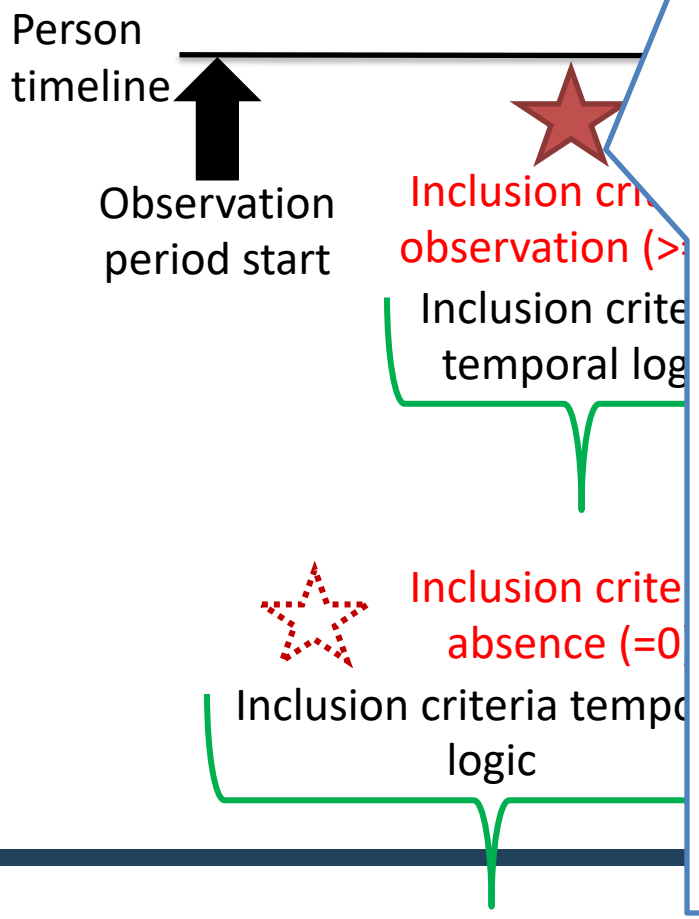
What initial event(s) define cohort entry?

- Events are recorded time-stamped observations for the persons, such as drug exposures, conditions, procedures, measurements and visits.
- The event index date is set to be equal to the event start date
- Initial events defined by a domain, conceptset, and any domain-specific attributes required

What initial event(s) define cohort entry?

- Do:
  - Define by existence of any observation in any domain
- Don’t:
  - Define by absence of an observation - when does absence occur?
  - Define by age- year of birth is constant, but requires index date to anchor age calculation
- Caution:
  - Defining a cohort by calendar date can cause observation bias, since that date unlikely to be at point of health service utilization, ex: cases matched to controls. Consider instead defining by a visit that occurs within a calendar timeframe.

Dissecting the anatomy of a cohort definition



What inclusion criteria are applied to the initial events?

- The qualifying cohort will be defined as all persons who have an initial event and satisfy all qualifying inclusion criteria.
- Each inclusion criteria is defined by domain(s), conceptset(s), domain-specific attributes, and the temporal logic relative to initial events
- Each qualifying inclusion criteria can be evaluated to determine the impact of the criteria on the attrition of persons from the initial cohort

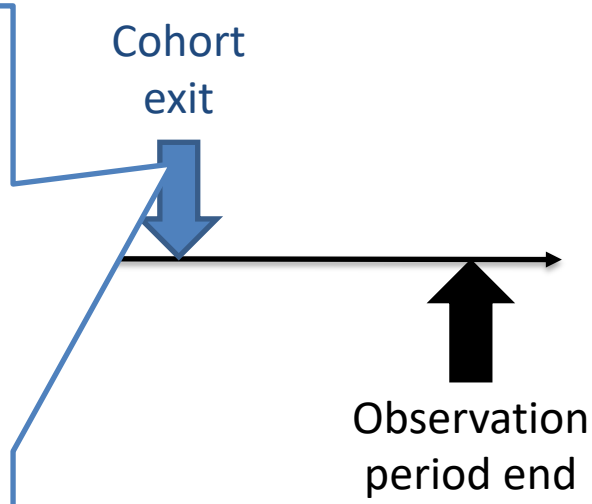
What inclusion criteria are applied to the initial events?

- Do:
  - Specify all criteria as inclusion criteria to avoid confusion of Boolean logic around inclusion vs. exclusion
  - use information on or before index event
    (think like a randomized trial: index event is study start, can't predict future)
- Don't:
  - Assume temporal logic, but always provide relative time window to evaluate criteria
- Caution:
  - There's a difference between 'first time in history with >365d prior observation' vs. 'no prior observation in last 365 days'
  - One person may have multiple initial events, criteria are applied to each event (not person)

Dissecting the anatomy of a cohort definition

What defines a person's cohort exit?

- Cohort exit signifies when a person no longer qualifies for cohort membership
- Cohort exit can be defined in multiple ways:
  - End of observation period
  - Fixed time interval relative to initial event
  - Last event in a sequence of related observations (ex: persistent drug exposure)
  - Censoring observations
- Cohort exit strategy will impact whether a person can belong to the cohort multiple times during different time intervals



What defines a person's cohort exit?

- Do:
  - Specify a cohort exit, even if you are not intending to use it for your analytic use case
- Don't:
  - Confuse censoring for analytical purposes with cohort definition (which can be analysis-independent)…ex: censoring at time of outcome
- Caution:
  - Time-of-cohort participation can be different from analysis time-at-risk…ex: acute effects can be studied using a fixed window post-exposure start, intent-to-treat analysis can follow person through observation period end

Building cohorts together

- Target: GLP1RA exposures
- Comparator: DPP4i exposures
- Indication: Type 2 diabetes mellitus
- Outcome 1: Acute myocardial infarction
- Outcome 2: Diarrhea

ATLAS instances for us to use during tutorials

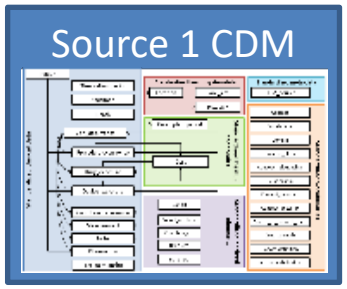
1. http://34.87.31.85
2. http://35.198.228.140
3. http://34.126.162.214
4. http://34.126.174.39
5. http://34.87.47.115

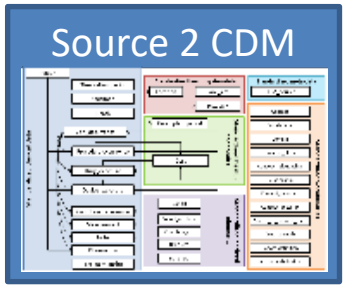
Lessons from building cohorts together

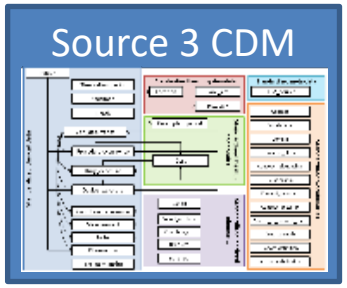
- Target: GLP1RA exposures
- Comparator: DPP4i exposures
  - Navigating ATC -> RxNorm and RxNorm -> ATC
  - All exposures vs. ‘new user’
  - Defining persistent exposure
    - Exercise: Find the errors
- Indication: Type 2 diabetes mellitus
  - Combining multiple entry events
  - Index date correction
  - Inclusion criteria for presence and absence of events
    - Exercise: Find the errors
- Outcome 1: Acute myocardial infarction
- Outcome 2: Diarrhea
  - Using ‘Recommend’ to build conceptset
  - Considering impact of vocabulary versioning
  - Incorporating care setting into event
  - Recurrent events with clean periods
    - Exercise: Find the errors

The journey to evidence

Standardized data







Cohort definition: a specification to identify the set of persons satisfying one or more criteria for a duration of time

Standardized analytics

T, {E} → Treatment pathways

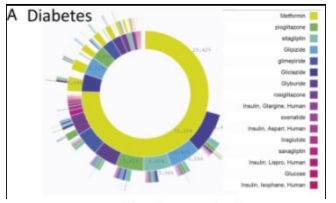
T, O → Incidence rate analysis

T, C, I, O → Comparative cohort design

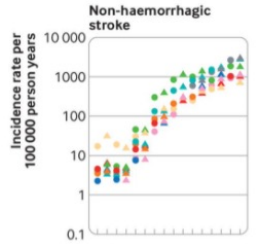
T, I, O → Self-controlled case series

T, O → Patient-level prediction

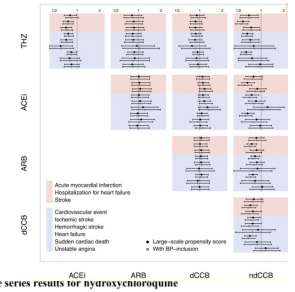
Impactful results



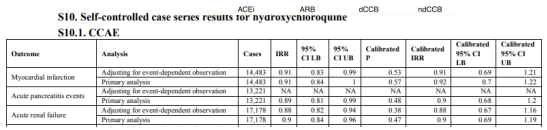
Hripcsak et al
PNAS 2016



Li et al
BMJ 2021



Suchard et al
Lancet 2019



S10. Self-controlled case series results for hydroxychloroquine

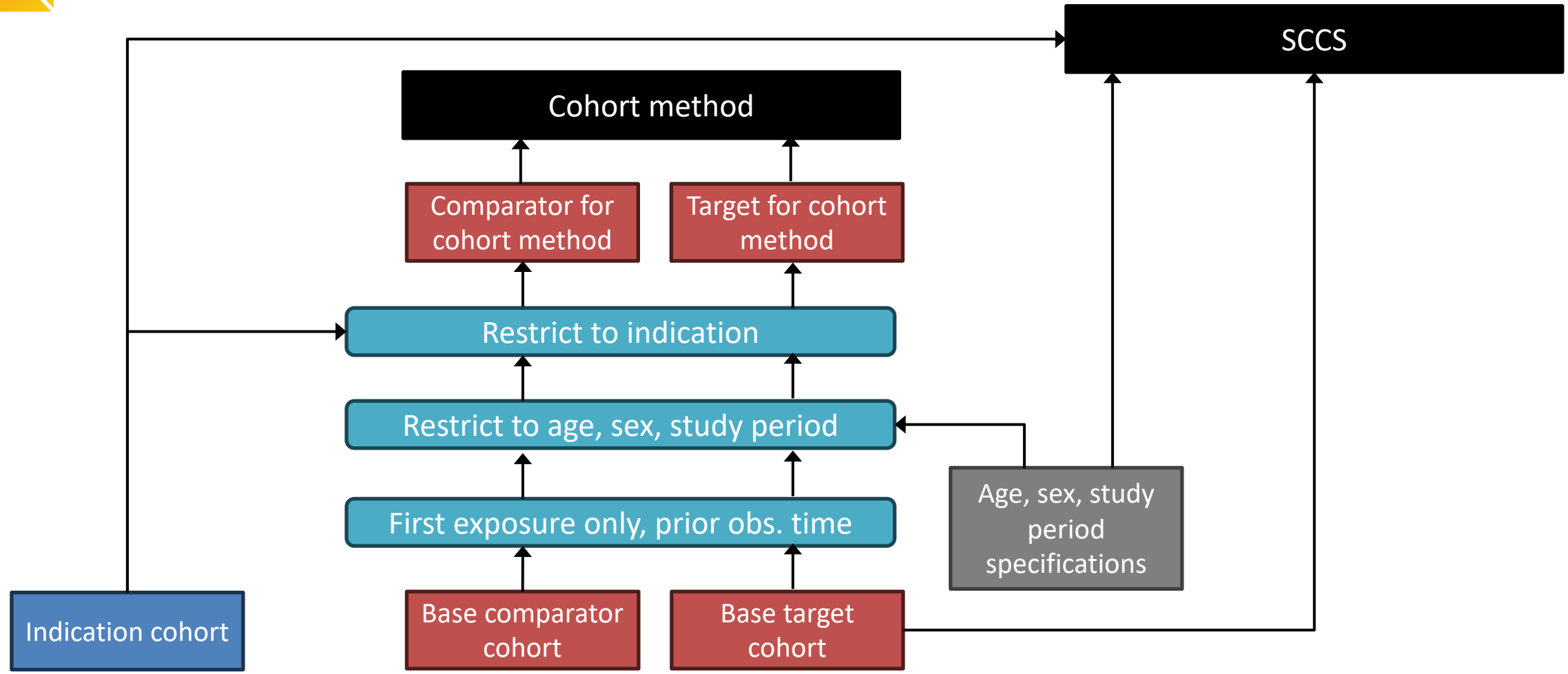
S10.1. CCAE

Outcome	Analysis	Cases	IRR	95% CI LB	95% CI UB	Calibrated P	Calibrated IRR	Calibrated 95% CI LB	Calibrated 95% CI UB
Myocardial infarction	Adjusting for event-dependent observation	14,483	0.91	0.83	0.99	0.53	0.91	0.69	1.21
	Primary analysis	14,483	0.91	0.84	1	0.57	0.92	0.7	1.22
Acute pancreatitis events	Adjusting for event-dependent observation	13,221	NA	NA	NA	NA	NA	NA	NA
	Primary analysis	13,221	0.89	0.81	0.99	0.48	0.9	0.68	1.2
Acute renal failure	Adjusting for event-dependent observation	17,178	0.88	0.82	0.94	0.38	0.88	0.67	1.16
	Primary analysis	17,178	0.9	0.84	0.96	0.47	0.9	0.69	1.19

Lane et al Lancet Rheumatology 2020

Williams et al
BMC MRM 2022

Deriving all exposure cohorts from base cohorts



Indication cohort is usually first diagnosis to end of observation

Target and comparator cohort contain any exposure, no restrictions