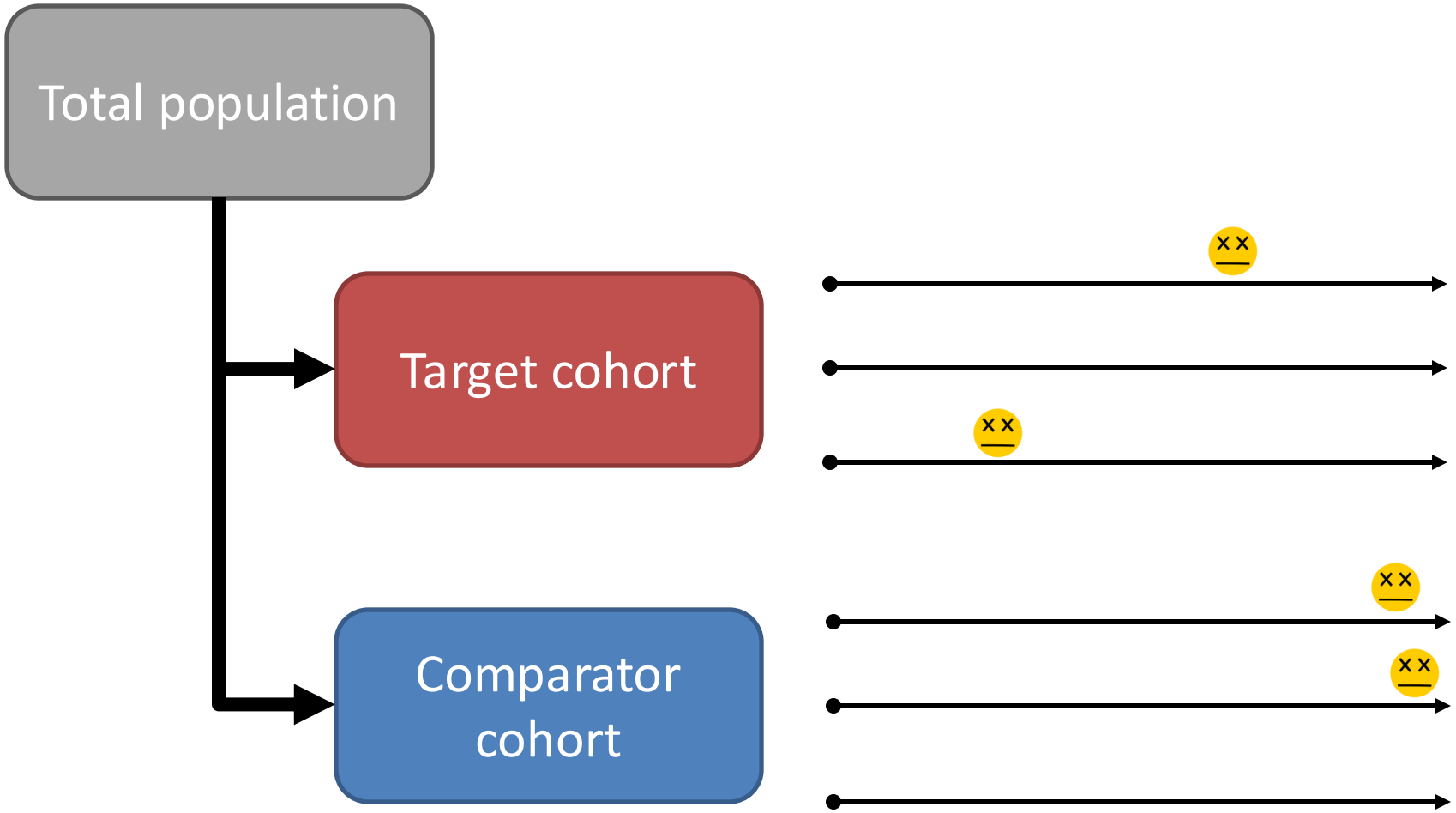
Estimation using CohortMethod in Strategus

George Hripcsak

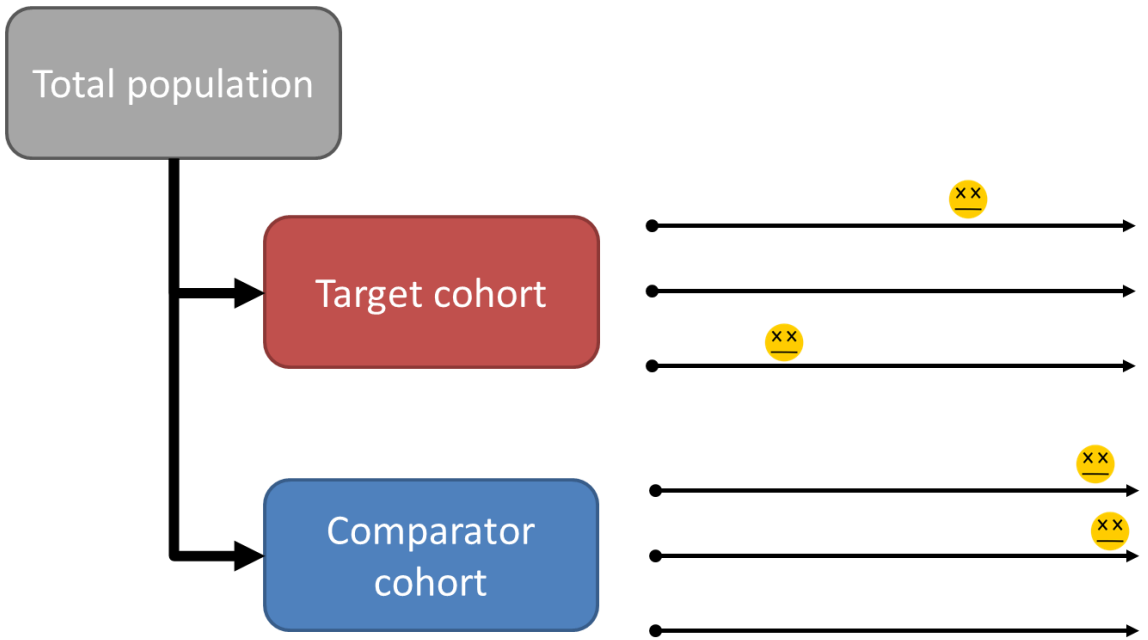
(Slides thanks to Martijn Schuemie and Patrick Ryan)

New-user cohort design

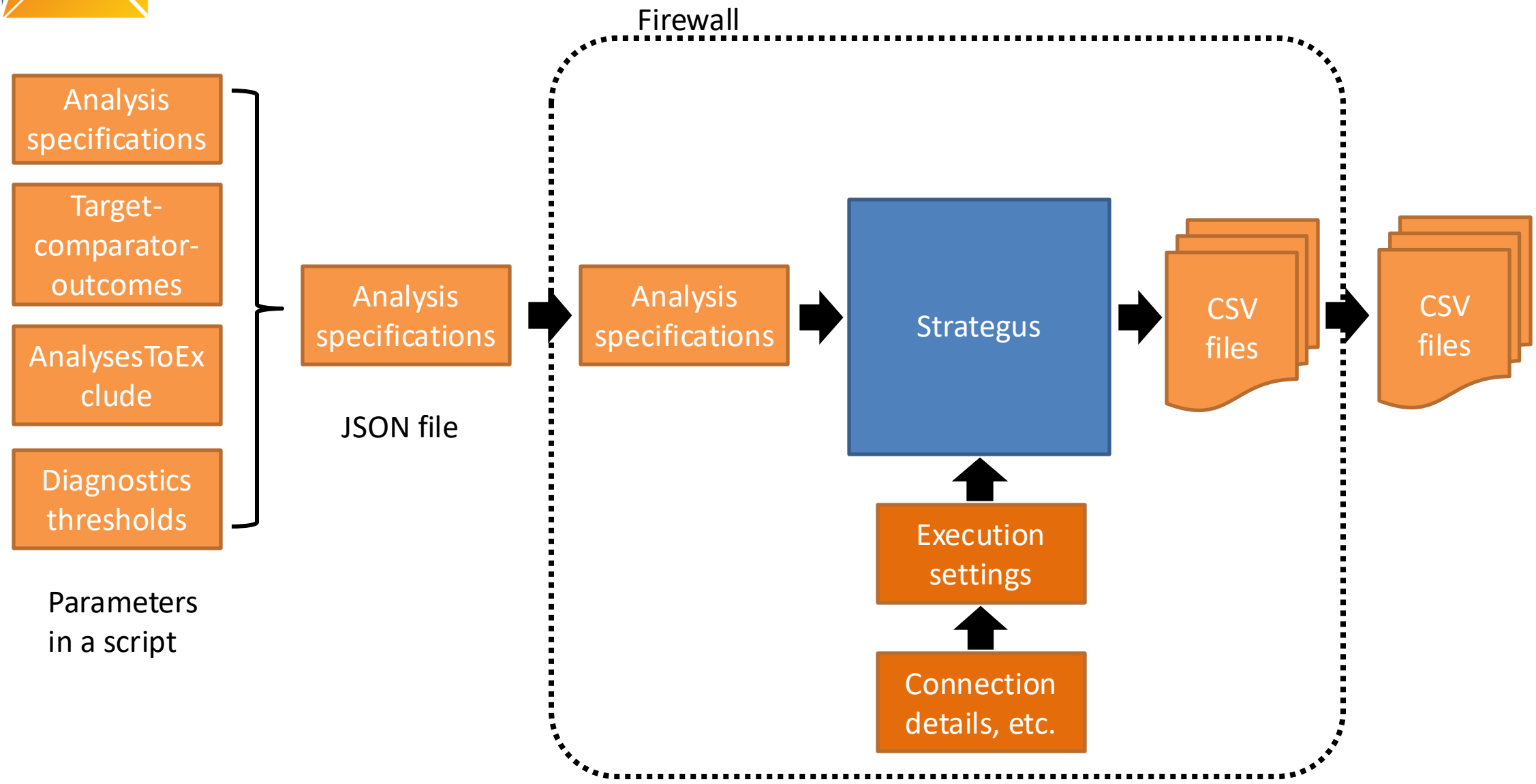


What do we need?

- Target and comparator cohorts
- Indication (included in above cohorts or separate)
- Outcome cohorts
- Timing of targets and outcomes
- Method to adjust for confounding and its parameters

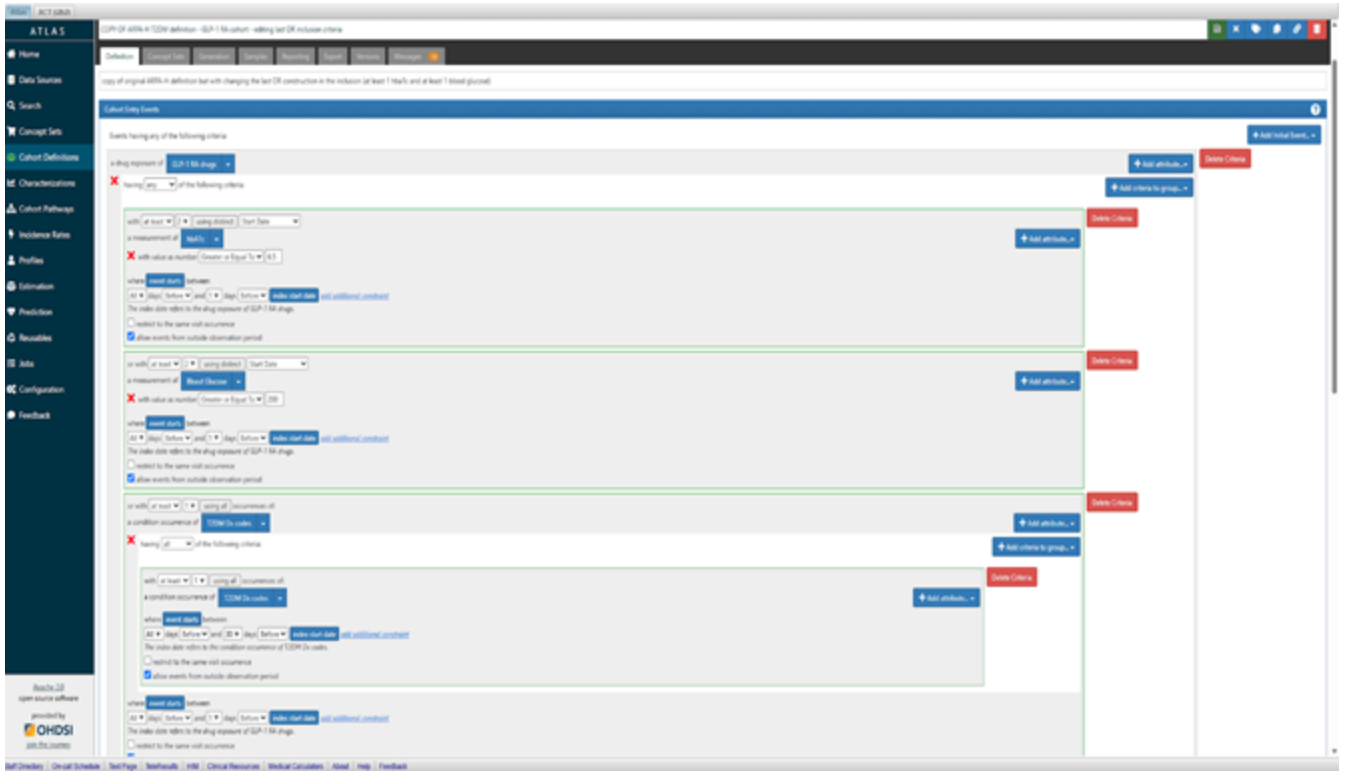


Strategus workflow



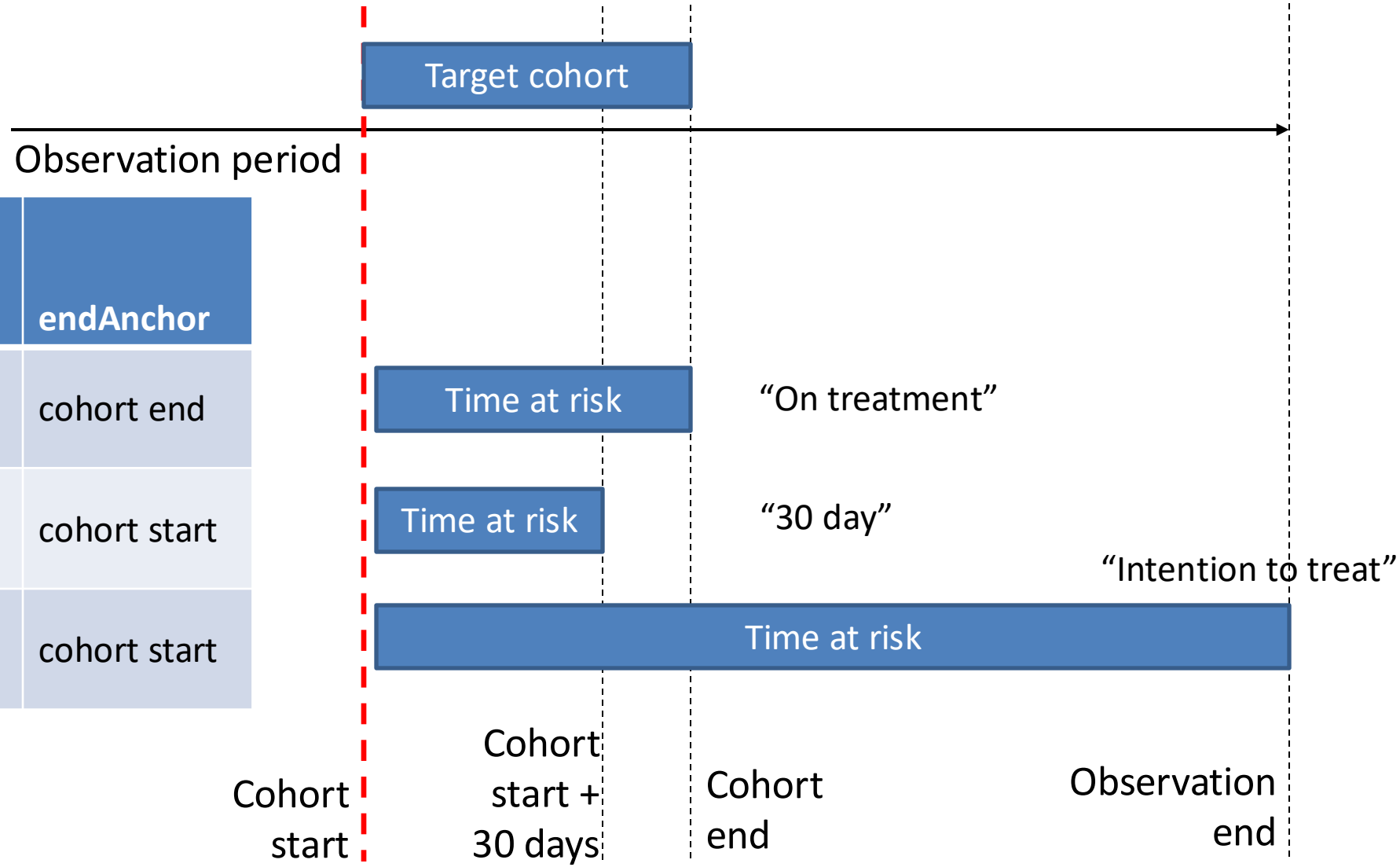
Create cohorts using ATLAS or CapR

- Use all of ATLAS's features
- Import these cohort definitions into the Strategus specification



Time-at-risk: when might the outcome occur?

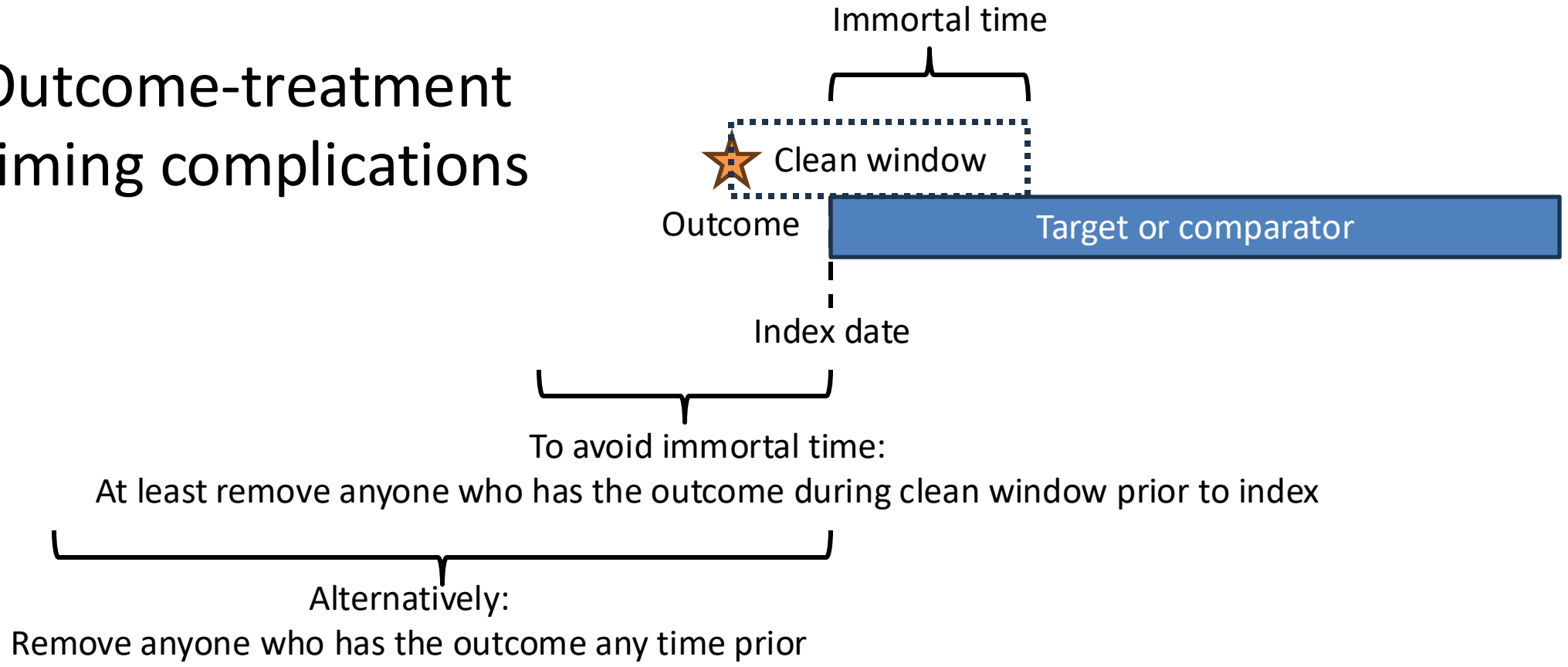
Risk Window start	startAnchor	Risk Window end	endAnchor
1	cohort start	0	cohort end
1	cohort start	30	cohort start
1	cohort start	99999	cohort start



Outcome lookback window and cohort clean window

- Outcome cohort clean window: how soon can outcome recur versus just repeat code

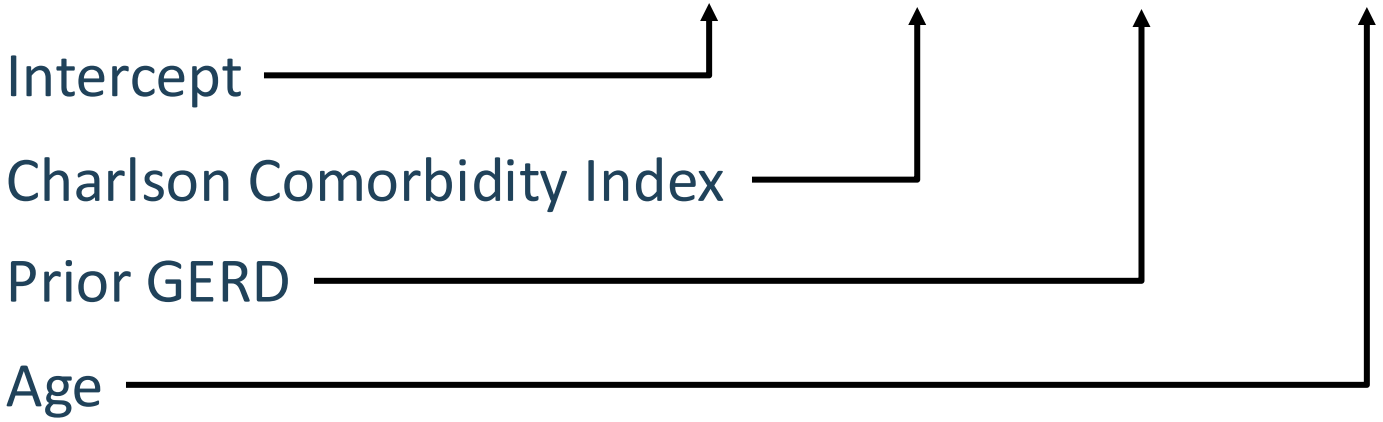
Outcome-treatment timing complications



Propensity score (PS)

Adjust for confounding using the propensity score, the probability of receiving the treatment conditional on a set of baseline characteristics

$$P(\,treatment \mid X) = f(\,\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\,)$$

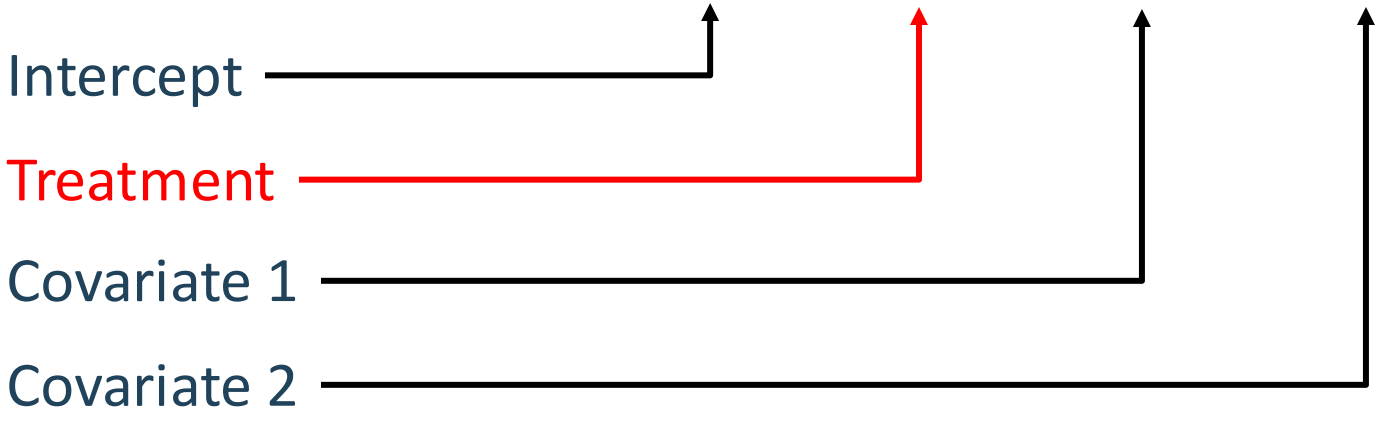


Using the PS, excluding codes

- Balance the two treatment groups so that any difference in outcome must be due to treatment
  - Matching, Stratification, Weighting
- Match 1-to-1 or 1-to-N (best if comparators have difference sample sizes)
- Need to exclude treatment and comparator from the PS model, or else perfectly predict treatment and cannot balance
  - Other administratively linked concepts (injection devices for GLP-1)

Outcome modeling

$$P(\,outcome|\,X) = f(\,\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\,)$$



- Logistic: Did the outcome occur yes/no?
- Poisson: How many times did the outcome occur?
- Cox: What was the time to the first outcome or end of observation?
- Conditional or non-conditional (Logistic, Poisson, Cox): stratify by PS strata or matched sets

Typical design

- Use another exposure as comparator
  - Could use https://data.ohdsi.org/ComparatorSelectionExplorer/ to choose
- Use LSPS propensity score
  - Variable ratio matching if C >> T, else 1-on-1 matching
  - Stratification
  - Remove exposure and comparator from propensity model (+ standard list of exclusion concept)
- Use both on-treatment TAR and 0-30-day TAR
- Cox model
  - Must be stratified when using variable ratio matching or stratification
- Include negative control outcomes

Script to generate JSON

```
################################################################################
# INSTRUCTIONS: Make sure you have downloaded your cohorts using
# DownloadCohorts.R and that those cohorts are stored in the "inst" folder
# of the project. This script is written to use the sample study cohorts
# located in "inst/sampleStudy" so you will need to modify this in the code
# below.
#
# See the Create analysis specifications section
# of the UsingThisTemplate.md for more details.
#
# More information about Strategus HADES modules can be found at:
# https://ohdsi.github.io/Strategus/reference/index.html
# omop-cdm-hades-modules.
# This help page also contains links to the corresponding HADES package that
# further details.
################################################################################

library(dplyr)
library(Strategus)

######################################################
# Above the line - MODIFY ----------------------------
######################################################
```

Get previously defined cohorts

```
# Get the list of cohorts - NOTE: you should modify this for your
# study to retrieve the cohorts you downloaded as part of
# DownloadCohorts.R
cohortDefinitionSet <- CohortGenerator::getCohortDefinitionSet(
  settingsFileName = "inst/sampleStudy/Cohorts.csv",
  jsonFolder = "inst/sampleStudy/cohorts",
  sqlFolder = "inst/sampleStudy/sql/sql_server"
)
```

Define each T-C analysis

```
tcis <- list(
  #standard analyses that would be performed during routine signal detection
  list(
    targetId = 20126, # Ace inhibitor
    comparatorId = 20127, # Diuretic
    indicationId = 20128, # Hypertensive disorder
    excludedCovariateConceptIds = c(
      21601783,
      21601461
    )
  ),
  list(
  # … next
```

Define outcomes and their clean windows

```
outcomes <- tibble(
  cohortId = c(20129, 20130), # AMI, Angioedema
  cleanWindow = c(365, 365)
)
```

Define the time-at-risk for each T-C pair

```
# Time-at-risks (TARs) for the outcomes of interest in your study
timeAtRisks <- tibble(
  label = c("On treatment", "On treatment"),
  riskWindowStart  = c(1, 1),
  startAnchor = c("cohort start", "cohort start"),
  riskWindowEnd  = c(0, 0),
  endAnchor = c("cohort end", "cohort end")
)
```

Study dates if needed

```
# If you are not restricting your study to a specific time window,
# please make these strings empty
studyStartDate <- '20200101' #YYYYMMDD
studyEndDate <- '20241231'   #YYYYMMDD
# Some of the settings require study dates with hyphens
studyStartDateWithHyphens <- gsub("(\\d{4})(\\d{2})(\\d{2})", "\\1-\\2-\\3", studyStartDate)
studyEndDateWithHyphens <- gsub("(\\d{4})(\\d{2})(\\d{2})", "\\1-\\2-\\3", studyEndDate)
```

Subgroup analysis

```
# These are the cohorts we'd like to used as subsets for all T/C's
cohortSubsets <- c(20226, 20227)
ageGroups <- list(
  list(
    minAge = 0,
    maxAge = 20
  ),
  list(
    minAge = 21,
    maxAge = 60
  ),
  list(
    minAge = 61,
    maxAge = 80
  )
)
```

Additional parameters

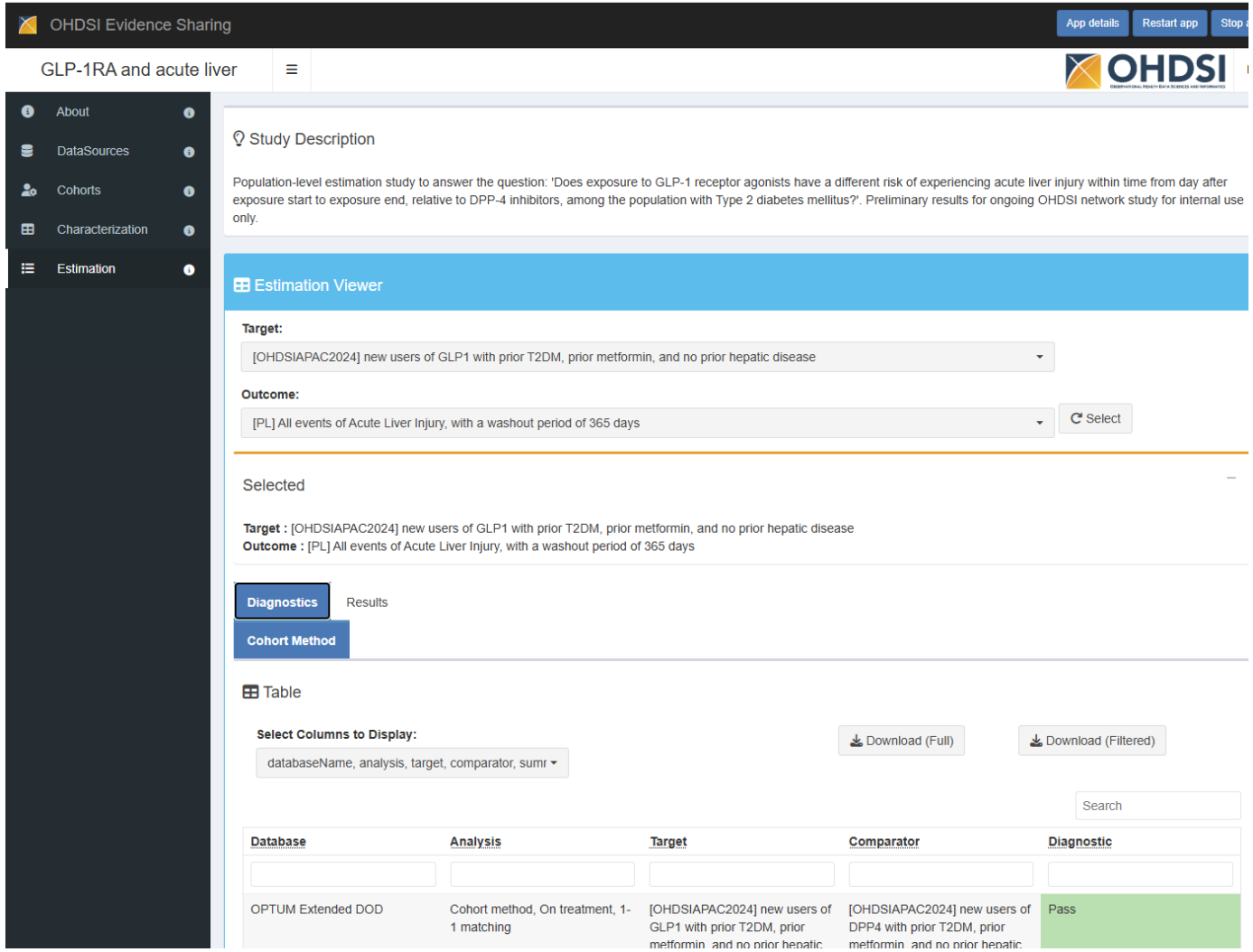
```
# Consider these settings for estimation  ---------------------------------
useCleanWindowForPriorOutcomeLookback <- FALSE # If FALSE, lookback window is all time prior, i.e., including only first events
psMatchMaxRatio <- 1 # If bigger than 1, the outcome model will be conditioned on the matched set

#############################################################
# Below the line - DO NOT MODIFY ----------------------
#############################################################

# Don't change below this line (unless you know what you're doing) ------------
```

Results

- Run analysis on Strategus
  - If no errors…
- Generate results
- Move to results.ohdsi.org
- Run Shiny app



Diagnostics by database

Diagnostics | Results

Cohort Method

Table

Select Columns to Display: databaseName, analysis, target, comparator, sumr

Download (Full) | Download (Filtered)

Search

Database	Analysis	Target	Comparator	Diagnostic
France DA	Cohort method, On treatment, 1-1 matching	[OHDSIAPAC2024] new users of GLP1 with prior T2DM, prior metformin, and no prior hepatic disease	[OHDSIAPAC2024] new users of DPP4 with prior T2DM, prior metformin, and no prior hepatic disease	Fail
German DA	Cohort method, On treatment, 1-1 matching	[OHDSIAPAC2024] new users of GLP1 with prior T2DM, prior metformin, and no prior hepatic disease	[OHDSIAPAC2024] new users of DPP4 with prior T2DM, prior metformin, and no prior hepatic disease	Fail
Healthverity CC	Cohort method, On treatment, 1-1 matching	[OHDSIAPAC2024] new users of GLP1 with prior T2DM, prior metformin, and no prior hepatic disease	[OHDSIAPAC2024] new users of DPP4 with prior T2DM, prior metformin, and no prior hepatic disease	Pass
IBM CCAE	Cohort method, On treatment, 1-1 matching	[OHDSIAPAC2024] new users of GLP1 with prior T2DM, prior metformin, and no prior hepatic disease	[OHDSIAPAC2024] new users of DPP4 with prior T2DM, prior metformin, and no prior hepatic disease	Pass

Power

Power | Attrition | Population characteristics | Propensity model | Propensity scores | Covariate balance | Systematic error | Kaplan-Meier

Power Table | TAR Table

Table

Select Columns to Display:

targetSubjects, comparatorSubjects, targetYears, c

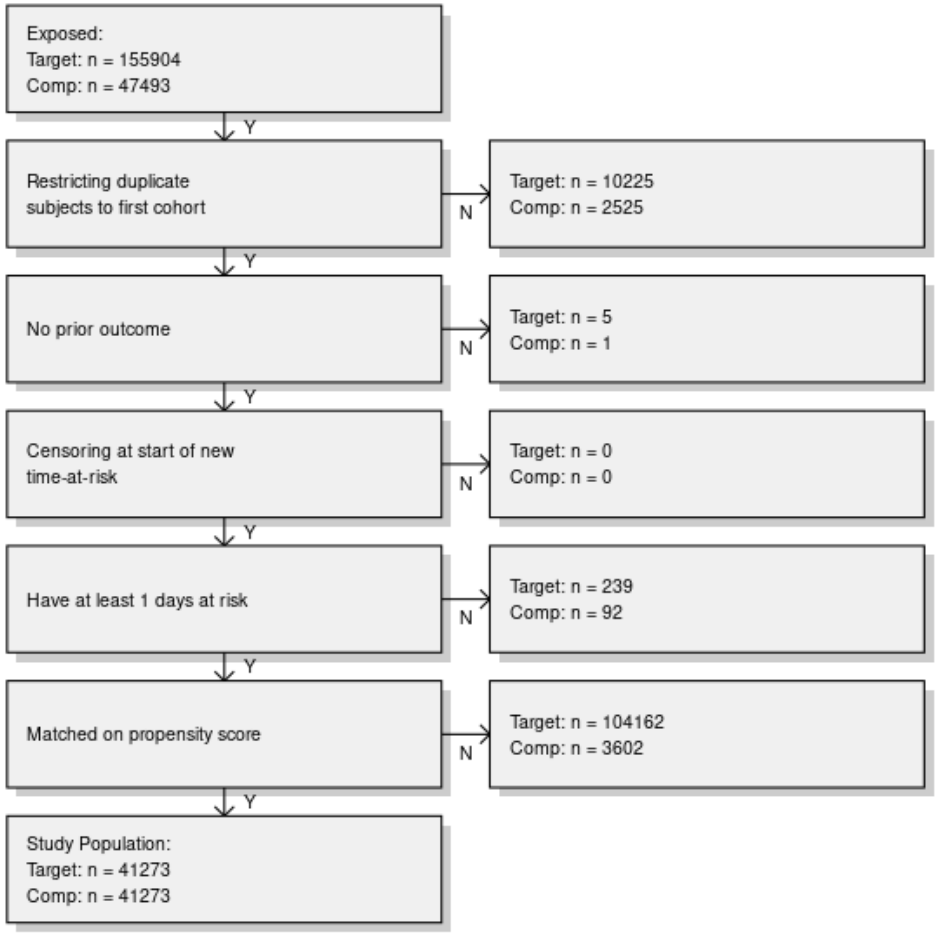
Download (Full) Download (Filtered)

Search

Target Subjects	Comparator Subjects	Target Years	Comparator Years	Target Events	Comparator Events	Target IR (per 1,000 PY)	Comparator IR (per 1,000 PY)	MDRR
41,273	41,273	45,842	31,655	32	47	0.70	1.48	1.88

Attrition

Power | **Attrition** | Population characteristics | Propensity model | Propensity scores

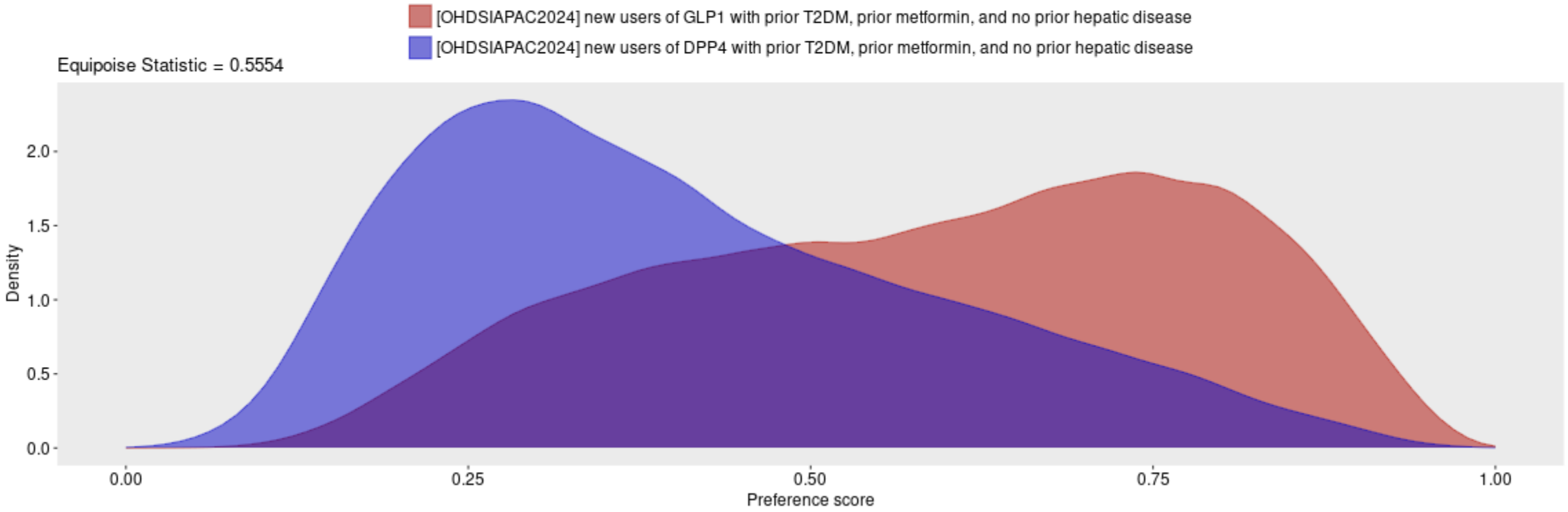


Propensity model

Covariate ↑	Beta	Beta (Absolute Value)
age group: 25 - 29	+0.123	0.123
age group: 30 - 34	+0.114	0.114
age group: 35 - 39	+0.11	0.110
age group: 40 - 44	+0.114	0.114
age group: 45 - 49	+0.099	0.099
age group: 50 - 54	+0.056	0.056
age group: 60 - 64	-0.117	0.117
age group: 65 - 69	-0.211	0.211
Charlson index - Romano adaptation	-0.225	0.225
condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Abdominal distension, gaseous	-0.09	0.090

Propensity scores

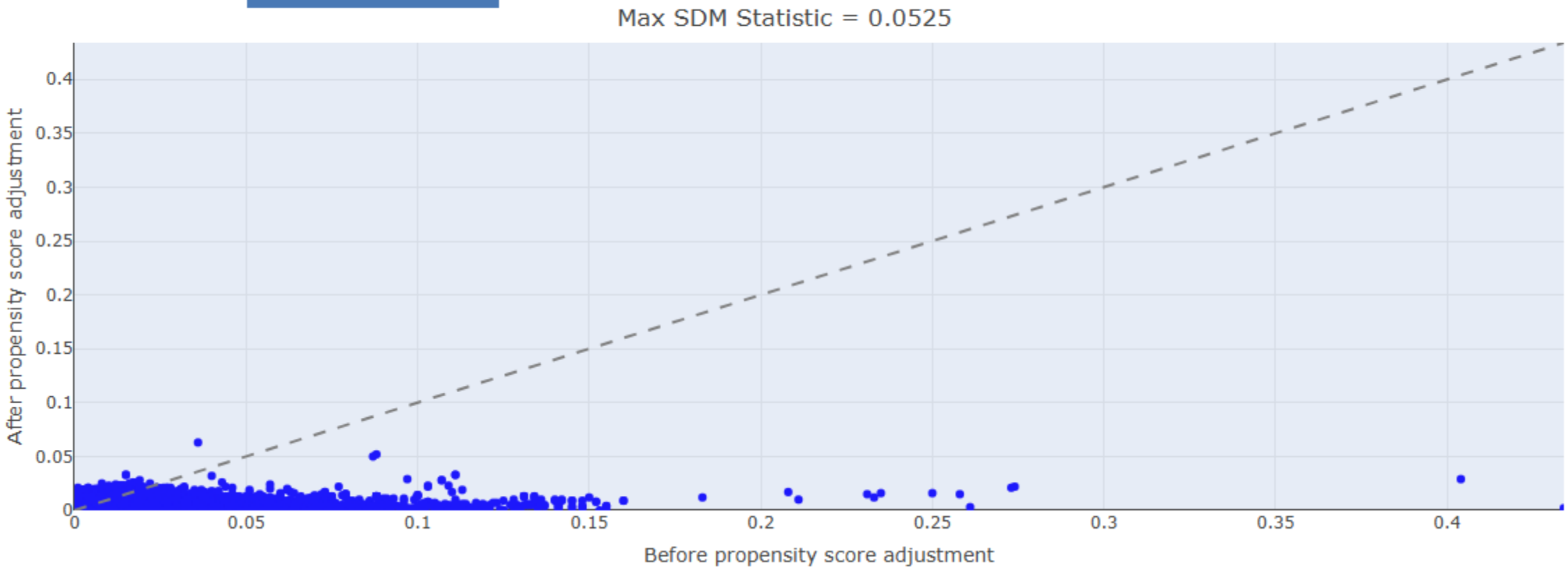
Power | Attrition | Population characteristics | Propensity model | **Propensity scores** | Covariate balance | Systematic error | Kaplan-Meier



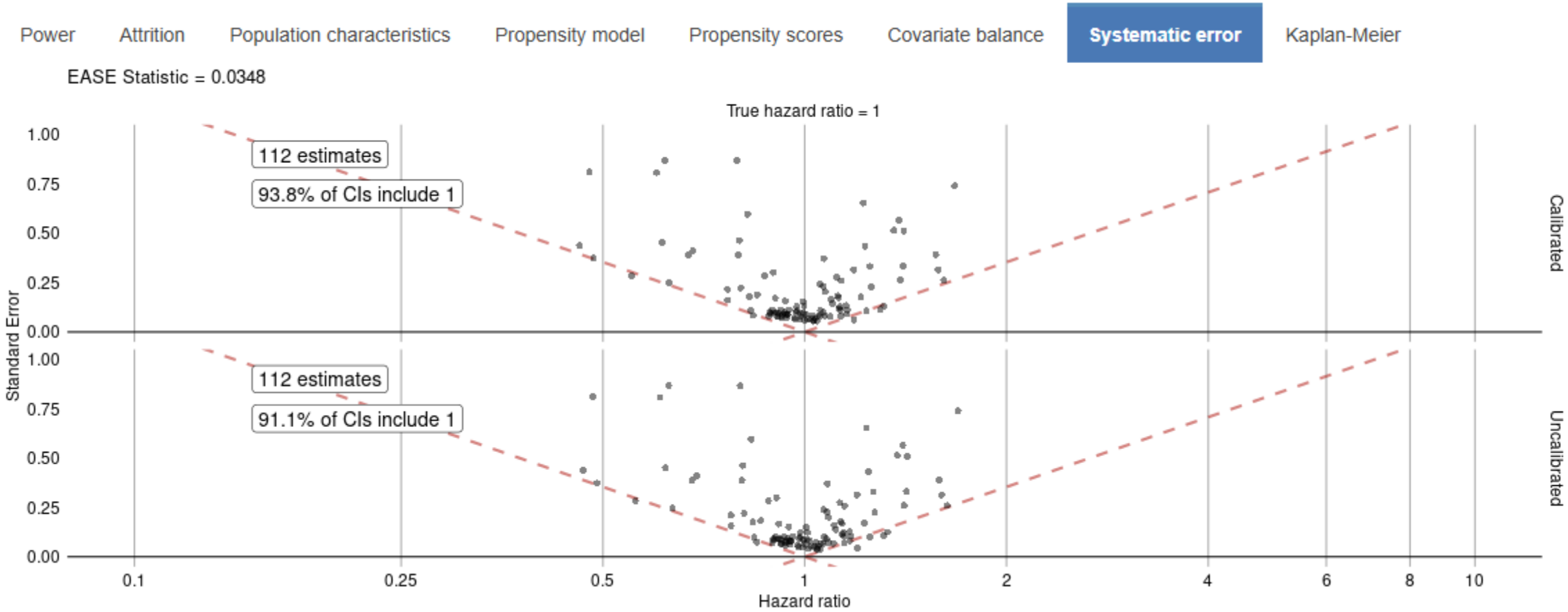
Covariate balance

Power | Attrition | Population characteristics | Propensity model | Propensity scores | **Covariate balance** | Systematic error | Kaplan-Meier

Covariate Balance Table | **Covariate Balance Plot**

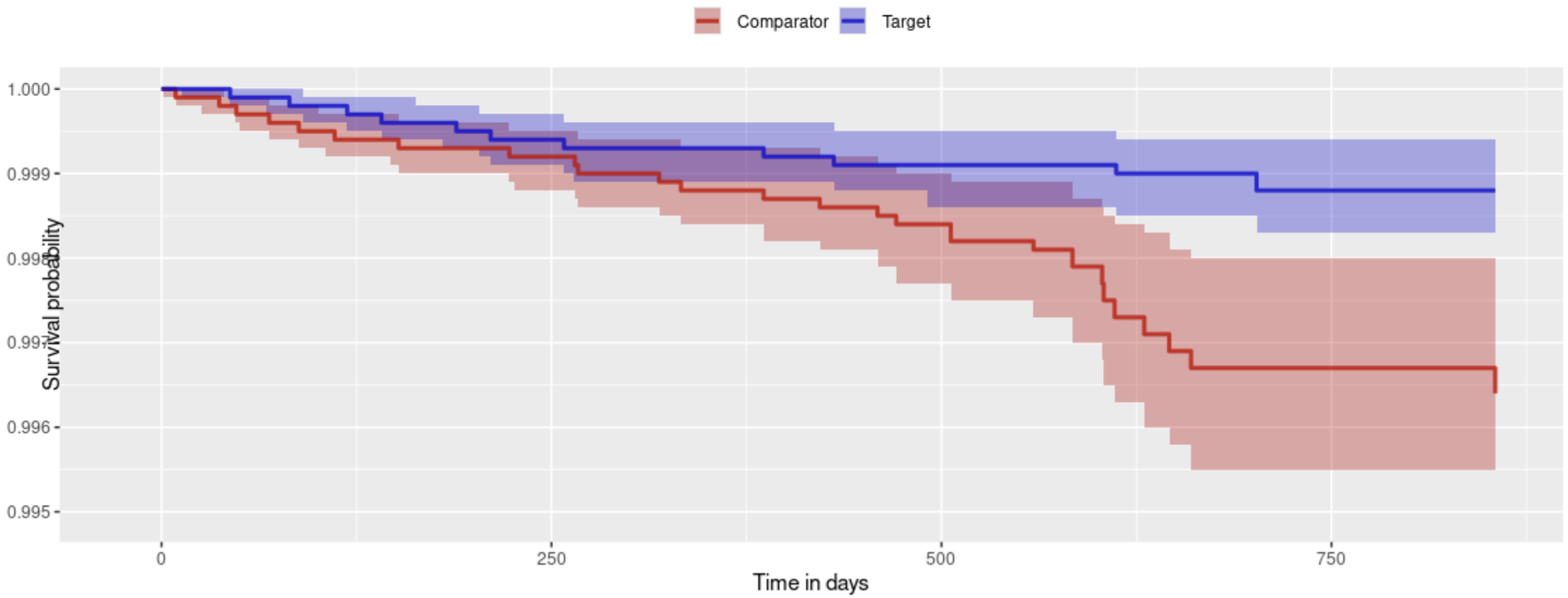


Systematic error



Kaplan-Meier

Power | Attrition | Population characteristics | Propensity model | Propensity scores | Covariate balance | Systematic error | **Kaplan-Meier**

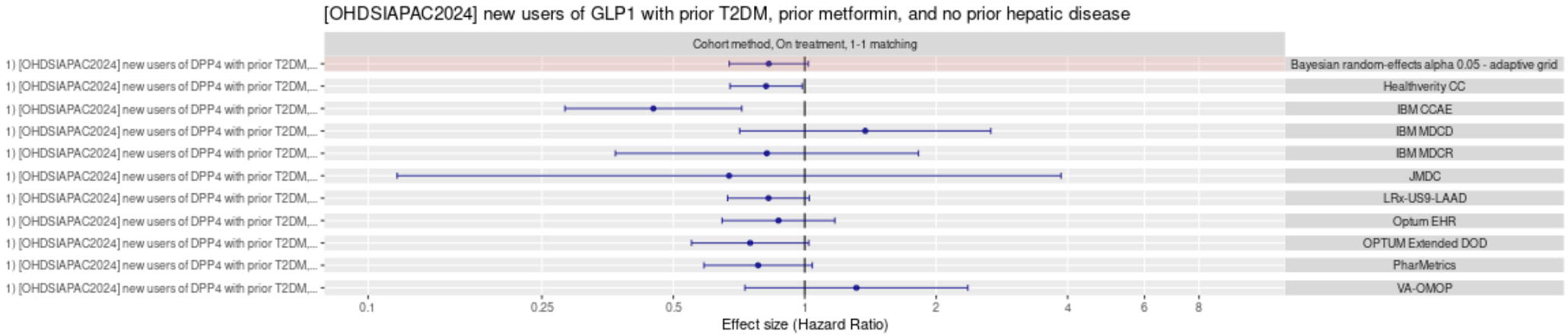


Results

Diagnostics | **Results**

Cohort Method Table | **Cohort Method Plot**

shortName	comparator
1) [OHDSIAPAC2024] new users of DPP4 with prior T2DM,...	[OHDSIAPAC2024] new users of DPP4 with prior T2DM, prior metformin, and no prior hepatic disease



Current practice for the self-controlled case series in Strategus

Marc A Suchard, MD, PhD

Typical use case

Question:

Does **Target** cause **Outcome** [in **Indication**] [in **Age Group**] [in **Sex**] [in **Time Period**]?

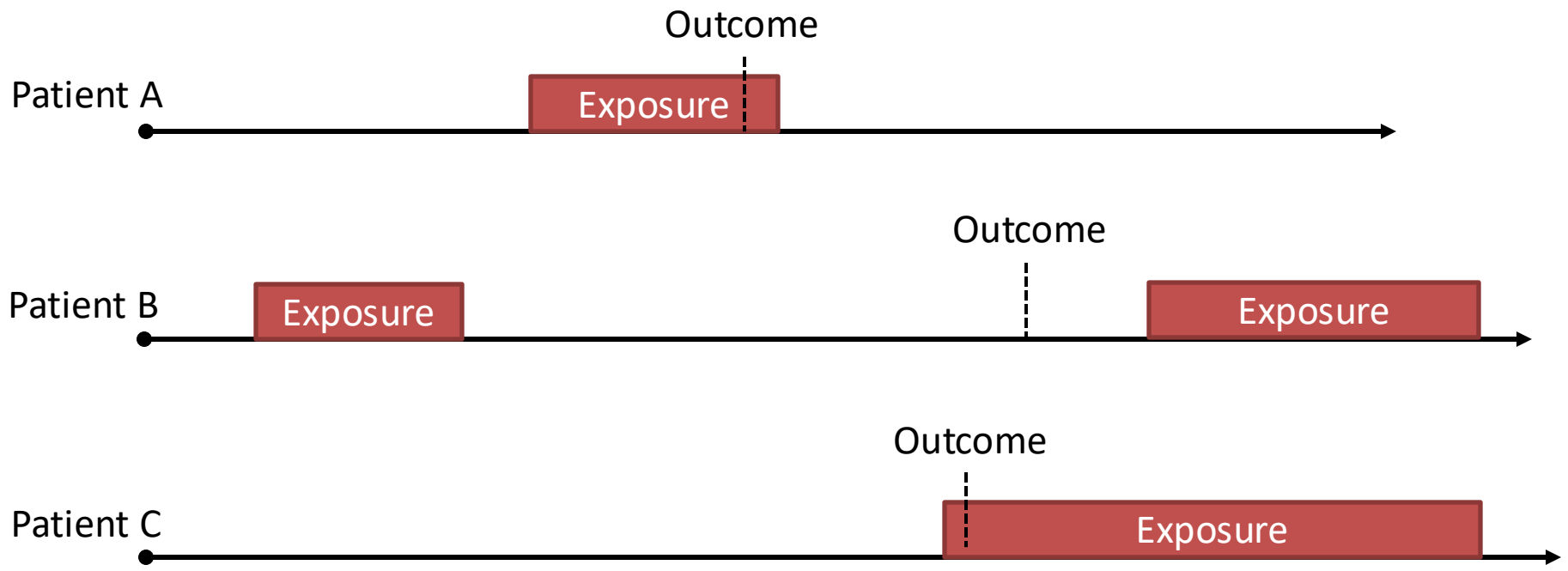
To answer this, we could run:

- Cohort method
- Self-controlled case series (SCCS)
- Meta-analysis of cohort method and SCCS
- Additional characterizations to support causal assessment

Our approach: always include both, let diagnostics decide when a design is appropriate

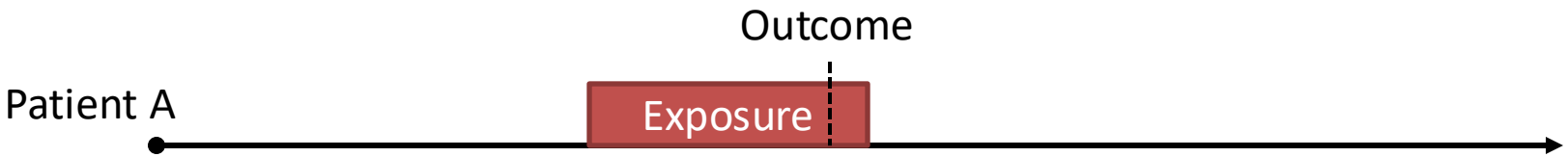
Self-Controlled Case Series

Is the outcome more likely during exposed time compared to non-exposed time?



Self-Controlled Case Series

Given that a patient has the outcome, is the outcome more likely during exposed time compared to non-exposed time?



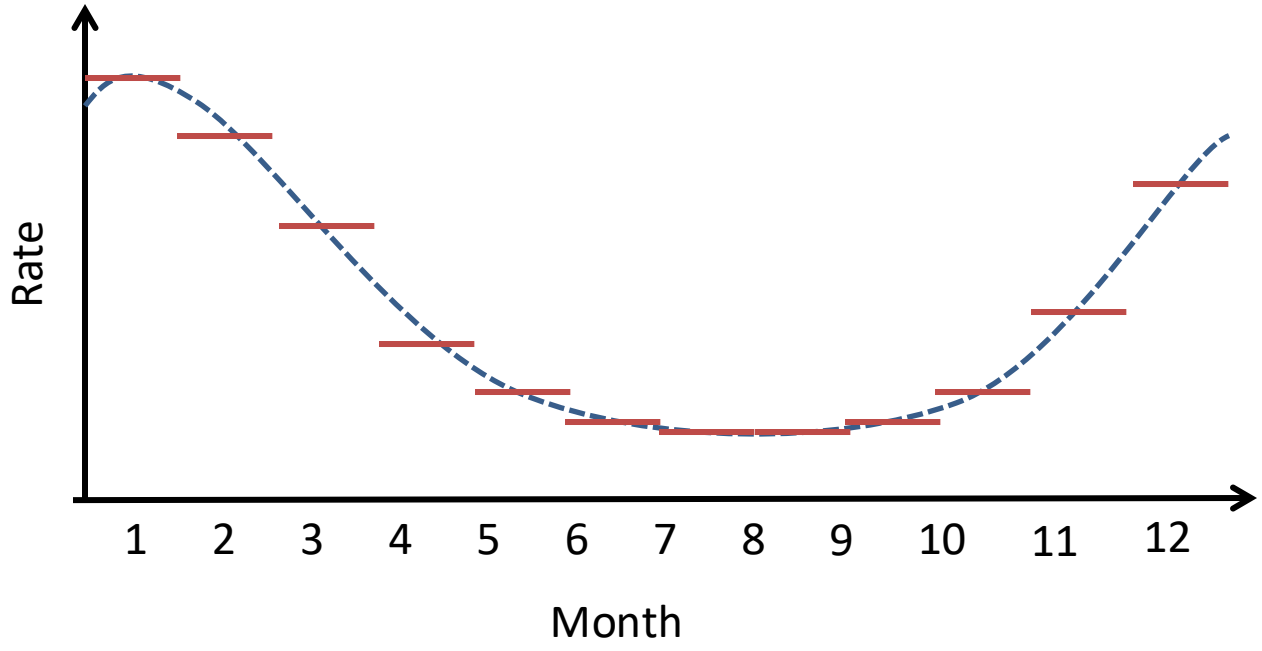
Conditioning on the outcome helps:

- Insensitive to differences between subjects that are **constant over time**
- Only require data on subjects with the outcome

Correcting for age, season and calendar

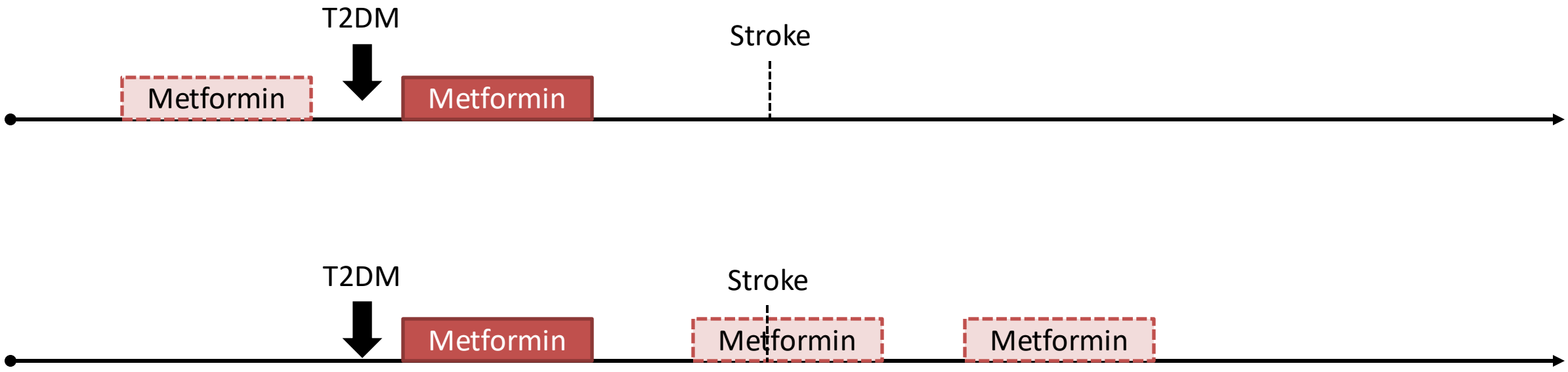
Problem: Time-varying confounding by, e.g., changing prevalence of exposure and outcome over age or calendar-time

Solution: SCCS package uniquely uses *objective diagnostics* and *spline adjustment* for time-varying covariates / confounders (assume effect constant within calendar month, age, etc.)



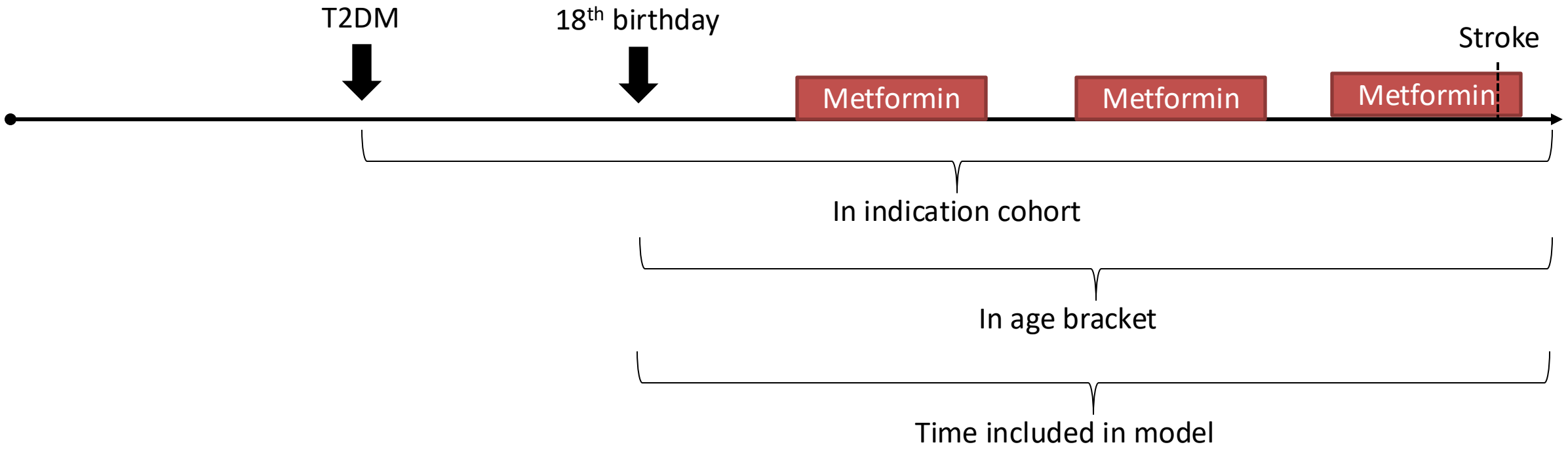
Take care in defining cohorts

- Don’t always use the same exposure cohort in CohortMethod and SCCS
  - E.g. First use of metformin, requiring a prior diagnosis of type 2 DM

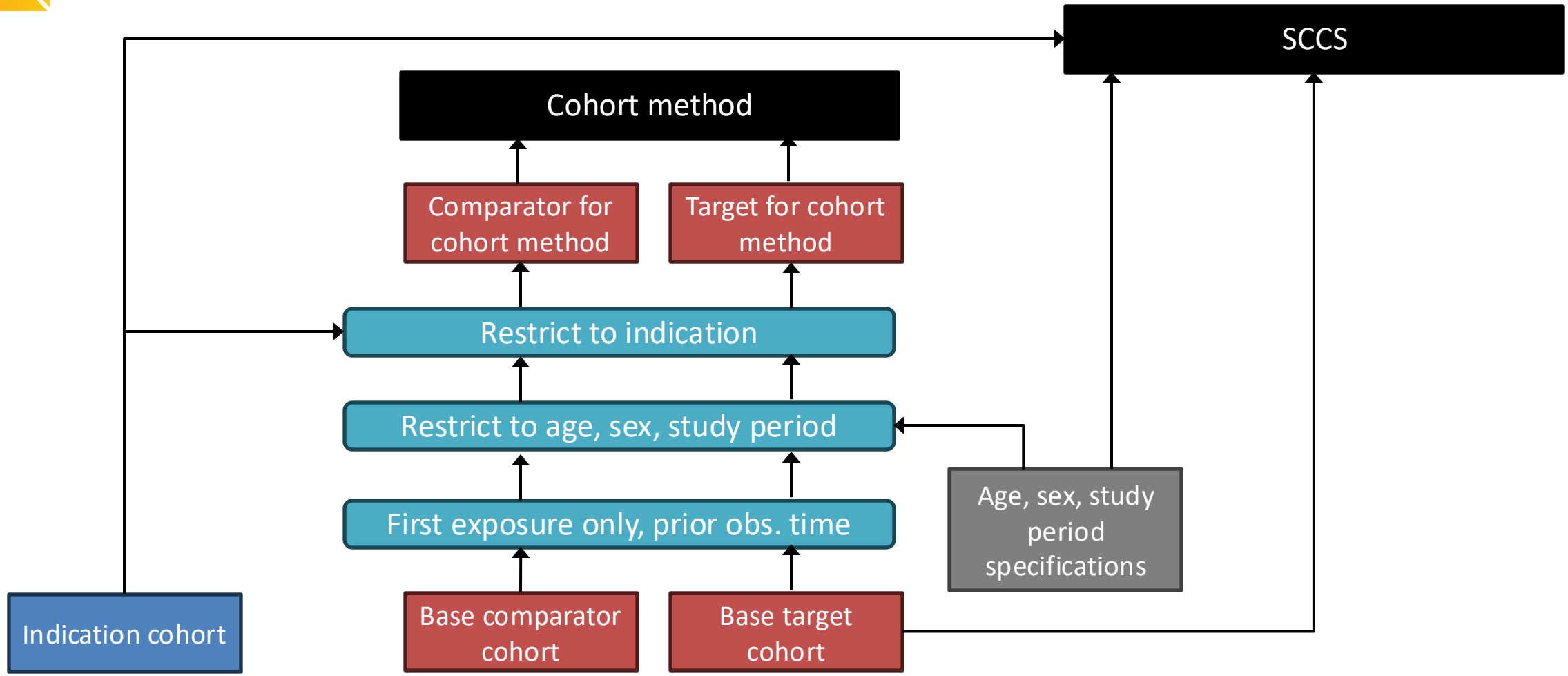


Use the indication (nesting) cohort

- Do apply all inclusion criteria to the observation period instead
  - E.g. Metformin in >= 18 year olds with a diagnosis of T2DM



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

Typical design

- Restrict to first outcome only
- 365-day naïve period (or 180)
- Restrict to indication
- Pre-exposure covariate: -30 to -1 days relative to exposure start
- Exposure covariate : On-treatment, except day 0
- Include season and calendar time splines
- **Include negative control outcomes** (same as CohortMethod)

The above-the-line below-the-line script

```r
#######################################################
# Above the line - MODIFY ---------------------------
#######################################################

# Get the list of cohorts - NOTE: you should modify this for your
# study to retrieve the cohorts you downloaded as part of
# DownloadCohorts.R
cohortDefinitionSet <- CohortGenerator::getCohortDefinitionSet(
  settingsFileName = "inst/sampleStudy/Cohorts.csv",
  jsonFolder = "inst/sampleStudy/cohorts",
  sqlFolder = "inst/sampleStudy/sql/sql_server"
)

tcis <- list(
  #standard analyses that would be performed during routine signal detection
  list(
    targetId = 20126, # Ace inhibitor
    comparatorId = 20127, # Diuretic
    indicationId = 20128, # Hypertensive disorder
    genderConceptIds = c(8507, 8532), # use valid genders (remove unknown)
    minAge = NULL, # All ages In years. Can be NULL
    maxAge = NULL, # All ages In years. Can be NULL
    excludedCovariateConceptIds = c(
      21601783,
      21601461
    )
  )
)
```

```r
  # Try to avoid intent-to-treat TARs for SCCS, or then at least disable calendar time sp
  sccsTimeAtRisks <- tibble(
    label = c("On treatment", "On treatment"),
    riskWindowStart  = c(1, 1),
    startAnchor = c("cohort start", "cohort start"),
    riskWindowEnd  = c(0, 0),
    endAnchor = c("cohort end", "cohort end")
  )
```

```r
  # Consider these settings for estimation  ---------------------------------------

  useCleanWindowForPriorOutcomeLookback <- FALSE # If FALSE, lookback window is all time prior, i.e.
  psMatchMaxRatio <- 1 # If bigger than 1, the outcome model will be conditioned on the matched set
  maxCohortSizeForFitting <- 250000 # Downsampled example study to 10000
  maxCohortSize <- maxCohortSizeForFitting
  maxCasesPerOutcome <- 1000000 # Downsampled example study to 10000

  # Consider these settings for patient-level prediction  -------------------------------------
  plpMaxSampleSize <- 1000000 # Downsampled example study to 20000

  #######################################################
  # Below the line - DO NOT MODIFY ---------------------
  #######################################################
```

https://github.com/ohdsi-studies/StrategusStudyRepoTemplate/blob/main/CreateStrategusAnalysisSpecificationTcis.R

ShinyApp diagnostics and results viewer

- Auto-magical byproduct of Strategus
- https://results.ohdsi.org/app/28_OhdsiExampleStudyApp

Database	Target	Outcome	Diagnostic	mdrr	ease	timeTrendP	preExposureP
IBM MDCD	ACE inhibitor	Angioedema	Pass	1.1649	0.179064803046526	1	1

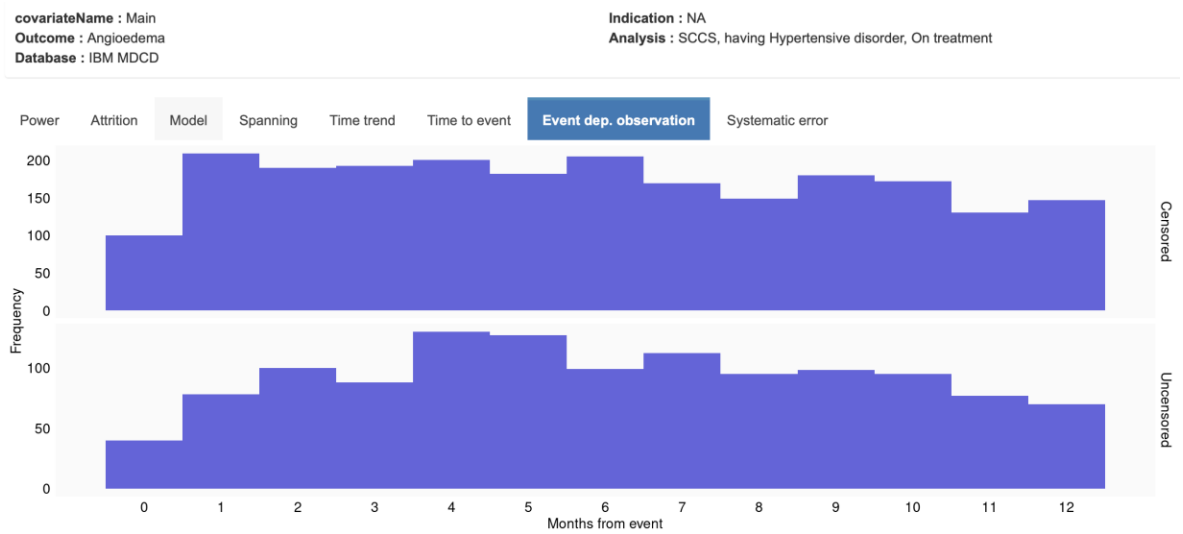


Figure 6. Histograms for the number of months between the first occurrence of the outcome and the end of observation, stratified by whether the end of observation was censored (inferred as not being equal to the end of database time), or uncensored (inferred as having the subject still be observed at the end of database time).

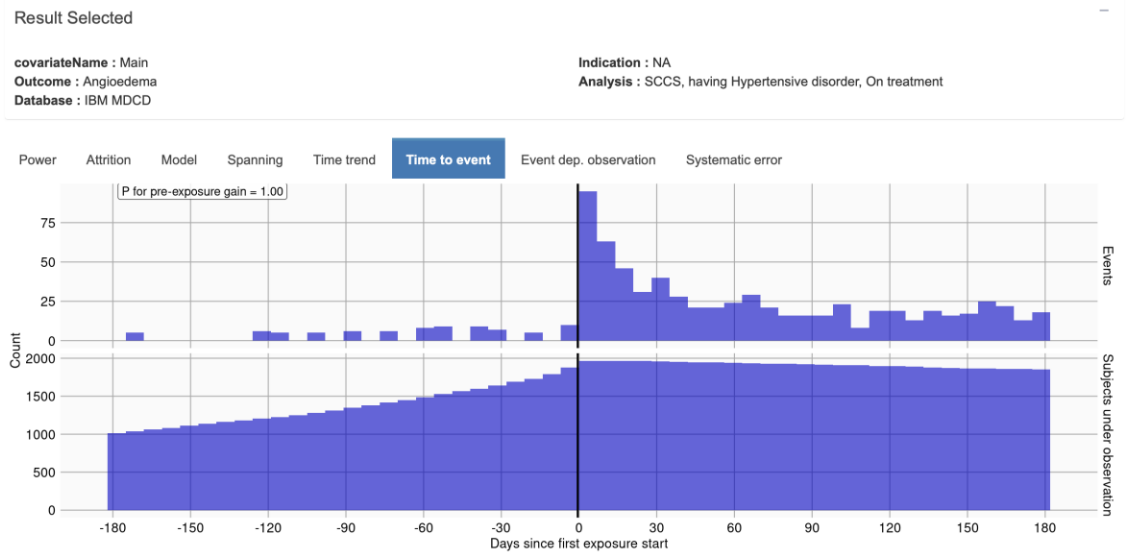


Figure 5. The number of events and subjects observed per week relative to the start of the first exposure (indicated by the thick vertical line).