SOS Challenge Week 7:

Study Diagnostics

Anti-VEGF and kidney failure

Fan Bu and George Hripcsak

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

'System' required elements:
- Required phenotypes
- Analysis specifications
- Decision thresholds

Distributed data network, standardized to common data model

Network coordination

Data quality evaluation

Research question → Database diagnostics

Fail → STOP

Pass

Phenotype development and evaluation

Cohort definitions → Cohort diagnostics

Fail → STOP

Pass

Analysis reliability evaluation

Analysis design choices → Study diagnostics

Fail → STOP

Pass

Today: evaluating analysis reliability through study diagnostics

Final unblinded results → Interface for exploration

Week 6-8: Analysis execution, study diagnostics, evidence synthesis

Revisiting 2022 Global Symposium Plenary

Plenary: Objective Diagnostics: A pathway to provably reliable evidence

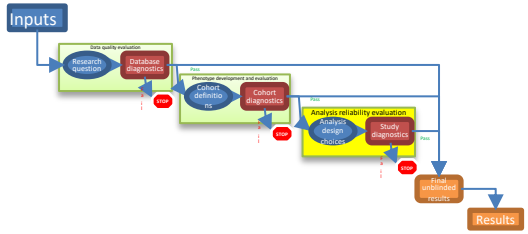
The plenary presentation from the 2022 OHDSI Symposium was led by **Martijn Schuemie (Johnson & Johnson)** and focused on 'Objective Diagnostics: A pathway to provably reliable evidence.' **Patrick Ryan (Johnson & Johnson, Columbia University)** also took part in this session.

This session introduced a series of diagnostics that can be evaluated to determine database, phenotype, and analysis fitness-for-use for generating reliable evidence. The presentation demonstrates the empirical performance of these objective diagnostics across the LEGEND-HTN result set to illustrate how objective diagnostics can be used and how they improve the quality of evidence generated.

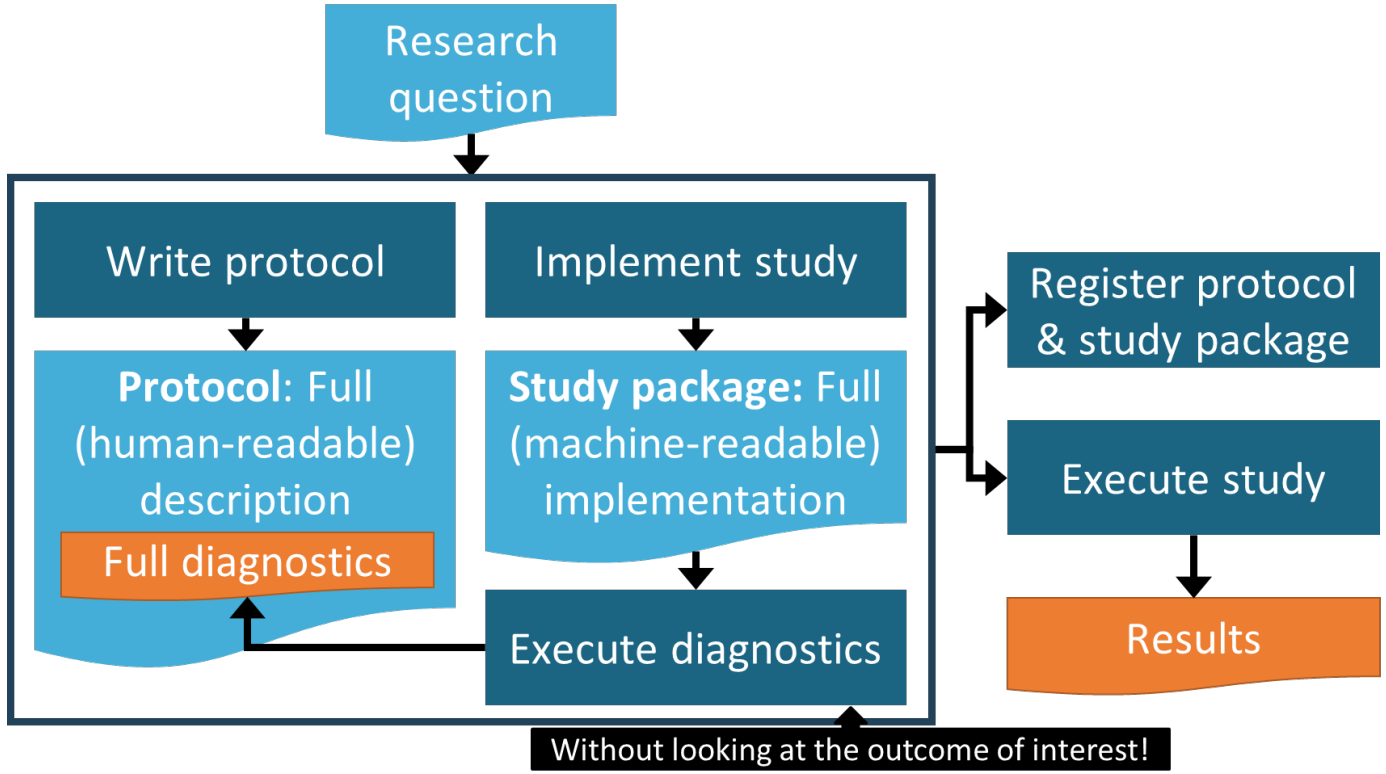
Recording:
https://youtu.be/DJZP5z6r-QE

Key message: To reduce post-hoc investigator bias, we need pre-specified objective diagnostics rules for evaluating the reliability of analyses. Results should be blinded if study fails diagnostics.

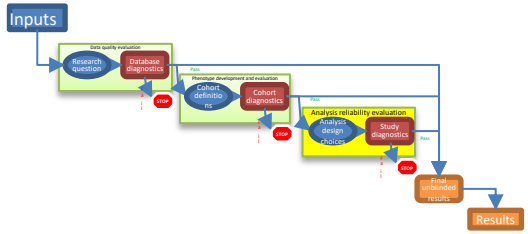
Avoiding investigator bias when interpreting diagnostics



- Diagnostics need to be performed ***before*** looking at study results
- Protocol can contain **diagnostics results**, or
- Protocol can contain **prespecified diagnostics rules** (so long as they are not modified post-hoc)



Study diagnostics

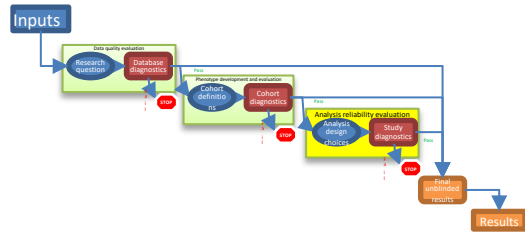


- Characterization
  - Feature summary, incidence, cohort pathways
    - Temporal stability, subpopulation heterogeneity, heterogeneity across data sources
- Population-level Estimation
  - Comparative cohort
    - Statistical power, comparator similarity, between-person confounding, generalizability, residual bias
  - Self-controlled case series
    - Statistical power, time-varying confounding, protopathic bias, residual bias
  - Meta-analysis
    - Statistical power, heterogeneity across data sources
- Patient-level prediction
  - PROBAST criteria (https://doi.org/10.7326/M18-1376) : embedded in *PatientLevelPrediction* package

PROBAST= **P**rediction model **R**isk **O**f **B**ias **AS**sessment **T**ool

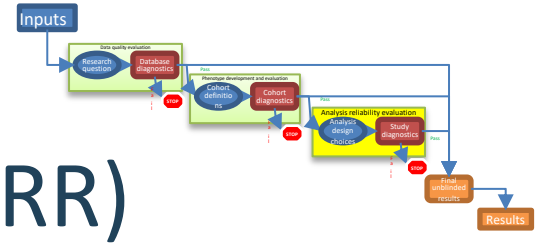
Derived from 2022 Symposium Plenary

Study diagnostics: A (short) checklist



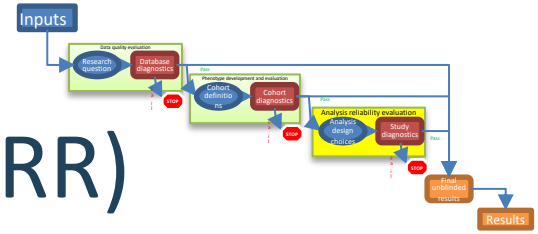
- Statistical power: minimum detectable relative risk
- Target-comparator similarity: empirical equipoise
- Between-person confounding: covariate balance
- Generalizability: attrition fraction
- Residual bias: expected absolute systematic error
- Other design/analysis-specific checks:
  - SCCS: time trends, pre-exposure outcomes, etc.
  - Prediction: PROBAST criteria

Statistical power: Minimum detectable relative risk (MDRR)



- Statistical power = probability of detecting an effect if a true effect exists
  - = 1- Type II error rate
  - Interventional studies: given hypothesized effect size & background incidence, determine sample size needed
  - Non-interventional studies (e.g., OHDSI network studies): sample size already exists, so we ask "**given the available data, what effect size would the analysis be able to detect?**"
- Usually, more data provide greater power
  - Design and analysis choices impact how much data are used to generate estimates
  - But, is less data definitely better than no data (or no results) at all?
- Rationale: to **avoid producing hard-to-interpret, under-powered estimates**
  - E.g., RR = 6.7 (0.5, 37.6)

Statistical power: Minimum detectable relative risk (MDRR) Examples from LEGEND-HTN



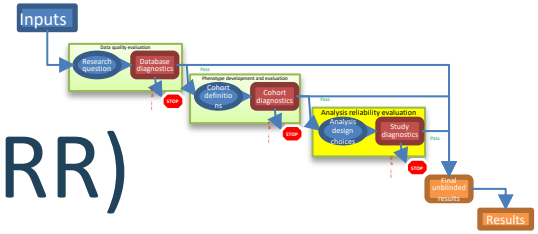
Good:
T = lisinopril
C = hydrochlorothiazide
O = cough

All databases have MDRR < 1.75 (ability to detect 75% increased risk if present), and 5 databases have MDRR < 1.1 (ability to detect 10% increased risk)

Table 1a. Number of subjects, follow-up time (in years), number of outcome events, and event incidence rate (IR) per 1,000 patient years (PY) in the target (*Lisinopril*) and comparator (*Hydrochlorothiazide*) group after stratification, as well as the minimum detectable relative risk (MDRR). Note that the IR does not account for any stratification.

Source	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
CUMC	3,565	3,387	4,563	5,555	284	288	62.23	51.84	1.26
IMSG	2,980	1,443	2,034	683	96	26	47.19	38.04	1.72
MDCD	45,283	24,993	20,591	9,038	3,249	1,206	157.79	133.42	1.09
MDCR	60,853	28,461	48,503	22,586	4,831	1,514	99.60	67.03	1.08
Optum	364,307	154,543	261,838	100,906	25,947	7,631	99.10	75.62	1.03
CCAE	548,859	243,878	380,386	163,469	30,942	9,419	81.34	57.62	1.03
Panther	583,608	189,242	207,470	66,877	21,366	5,369	102.98	80.28	1.04
Summary	1,609,455	645,947	925,388	369,118	86,715	25,453	93.71	68.96	1.02

Statistical power: Minimum detectable relative risk (MDRR) Examples from LEGEND-HTN



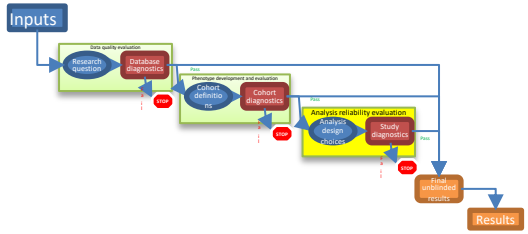
Bad:
T = candesartan
C = chlorthalidone
O = rhabdomyolysis

All databases have MDRR > 6 (underpowered to detect 600% increased risk if present), and two databases have MDRR > 15
<5 cases in target and comparator

Table 1a. Number of subjects, follow-up time (in years), number of outcome events, and event incidence rate (IR) per 1,000 patient years (PY) in the target (*Candesartan*) and comparator (*Chlorthalidone*) group after stratification, as well as the minimum detectable relative risk (MDRR). Note that the IR does not account for any stratification.

Source	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
Optum	4,510	7,682	3,394	5,037	<5	<5	<1.47	<0.99	>6.27
CCAE	4,897	14,092	4,179	8,519	0	<5	0.00	<0.59	>17.53
Panther	3,148	15,105	877	5,626	0	<5	0.00	<0.89	>27.56

Statistical power: MDRR in the Anti-VEGF study

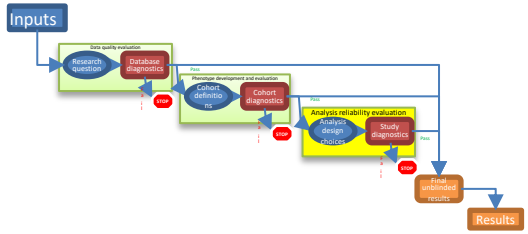


- Results ShinyApp: https://data.ohdsi.org/AntiVegfKidneyFailure/#use

Database	Target	Comparator	Outcome	Max SDM	Shared Max SDM	Equipoise	MDRR	EASE
CCAE	aflibercept	ranibizumab	ESRD	0.065	0.135	0.607	2.05	0.054
CCAE	ranibizumab	bevacizumab	ESRD	0.051	0.097	0.834	1.89	0.054
CCAE	aflibercept	bevacizumab	ESRD	0.055	0.113	0.822	1.82	0.067

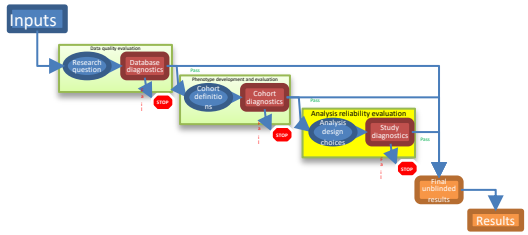
All analyses have MDRR <= 2.05 (ability to detect 105% increased risk if present).
The last two analyses have MDRR <= 1.9 (ability to detect 90% increased risk if present).

Empirical Equipoise: Preference score



- Randomized clinical trials assign treatments to subjects with the same probabilities
  - E.g., 1:1 randomized head-to-head trial: each subject 50%-50% chance to target/comparator group, regardless of patient/provider characteristics
  - Randomization ---> persons assigned to target cohort are exchangeable at baseline with persons assigned to comparator cohort
- Non-interventional studies (OHDSI studies) involve observing treatment choices, which can be influenced by patient or provider characteristics
  - Comparator selection is a pre-analysis design choice
  - **Preference** = probability of patient assigned to target vs. comparator, given baseline features
  - "Preference = 50%" means indifference between treatments for a patient, akin to random assignment
- **Similarity between target & comparator**: equipoise measured by preference scores
  - what proportion of the target population is close to treatment indifference? (**PS between 0.3 and 0.7**)
  - want this proportion to be large (> 0.5, as suggested by literature)

Empirical Equipoise: Preference score Examples from LEGEND-HTN



Good:
T = valsartan
C = olmesartan
DB = CCAE

Even with >40,000 patients on each drug, large-scale propensity score model could not meaningfully discriminate between the two treatments; >90% of persons in 'empirical equipoise' with a preference score between 0.3 and 0.7

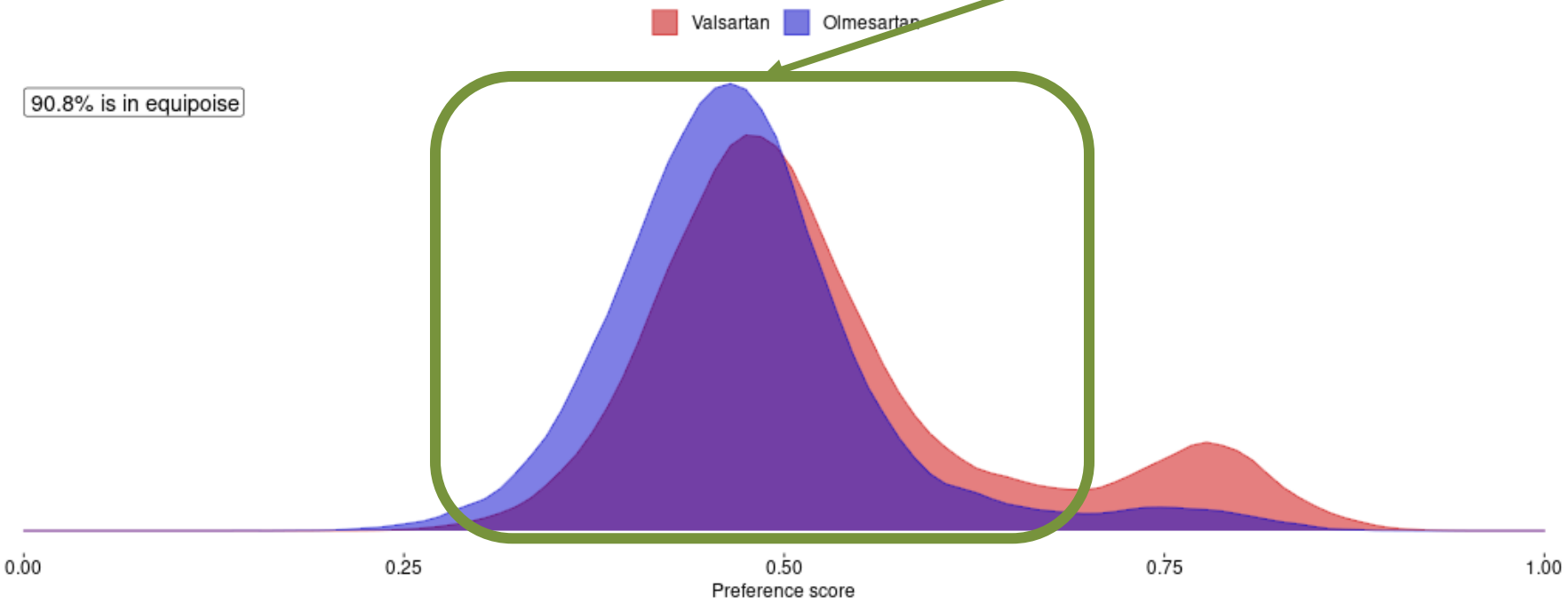
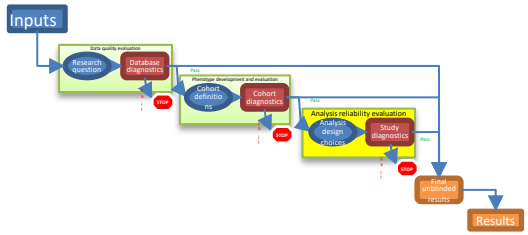


Figure 2. Preference score distribution. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups. A higher overlap indicates subjects in the two groups were more similar in terms of their predicted probability of receiving one treatment over the other.

Empirical Equipoise: Preference score Examples from LEGEND-HTN



Bad:
T = valsartan
C = chlorthalidone
DB = CCAE

Baseline characteristics can clearly discriminate most new users of valsartan vs. chlorthalidone; <30% of persons in 'empirical equipoise' with a preference score between 0.3 and 0.7

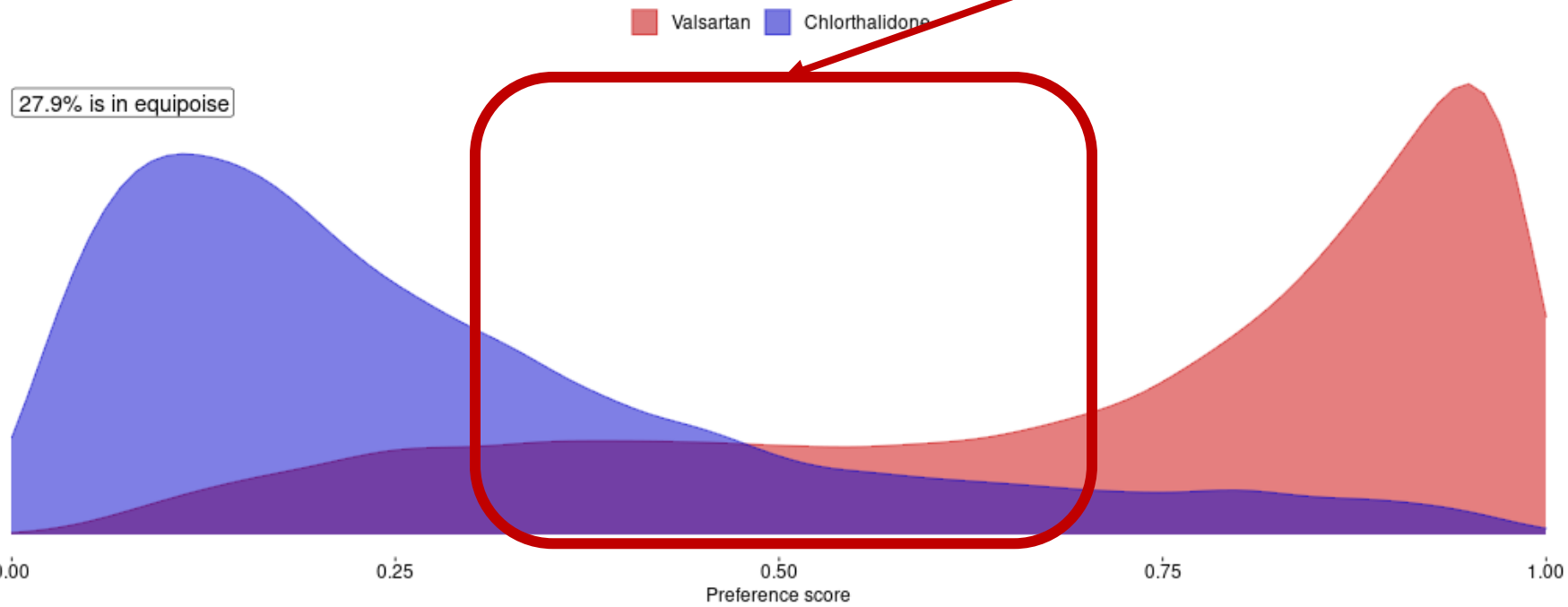
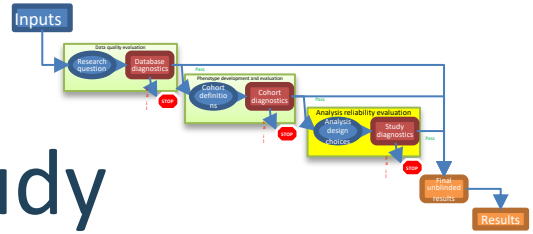


Figure 2. Preference score distribution. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups. A higher overlap indicates subjects in the two groups were more similar in terms of their predicted probability of receiving one treatment over the other.

Empirical Equipoise: Preference scores in the Anti-VEGF study



Good!

>83% of persons in “empirical equipoise” with a preference score between 0.3 and 0.7.

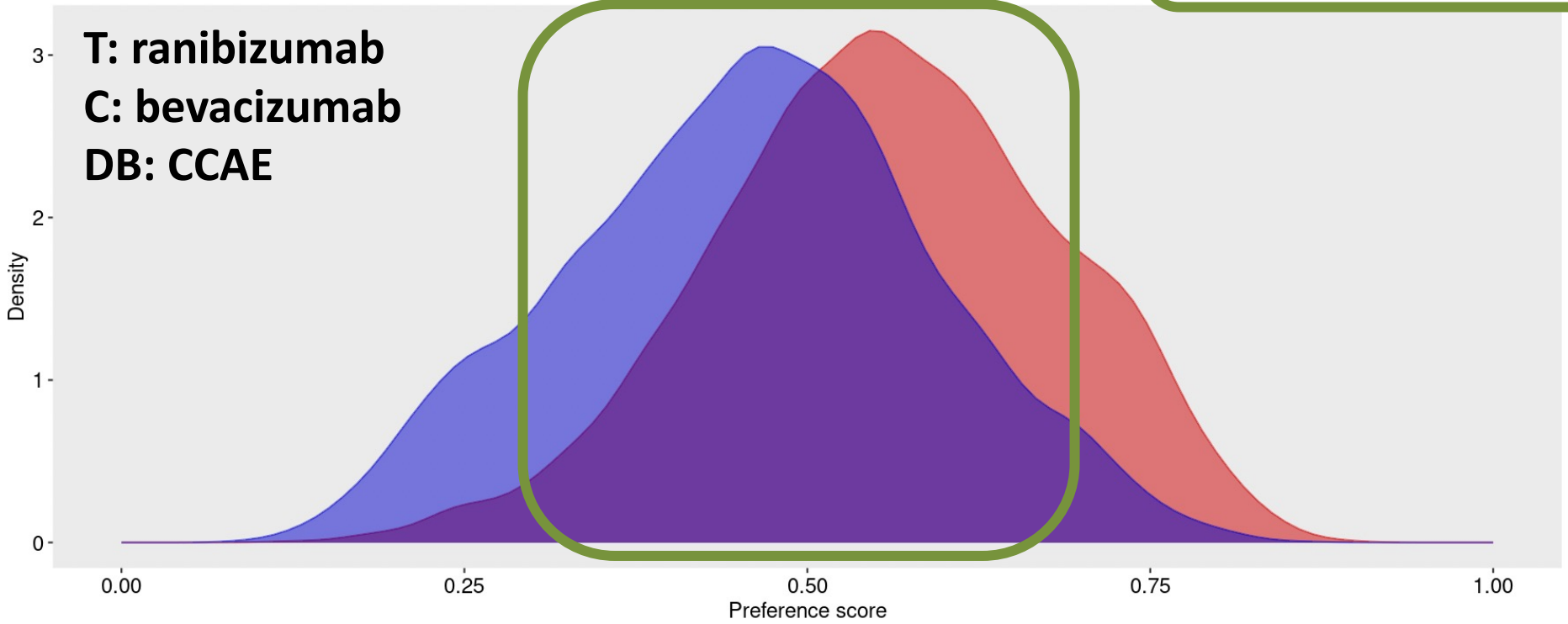
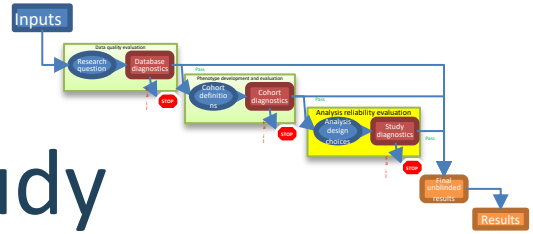


Figure 2. Preference score distribution. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups. A higher overlap indicates subjects in the two groups were more similar in terms of their predicted probability of receiving one treatment over the other.

Empirical Equipoise: Preference scores in the Anti-VEGF study

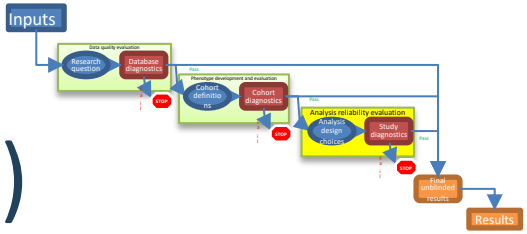


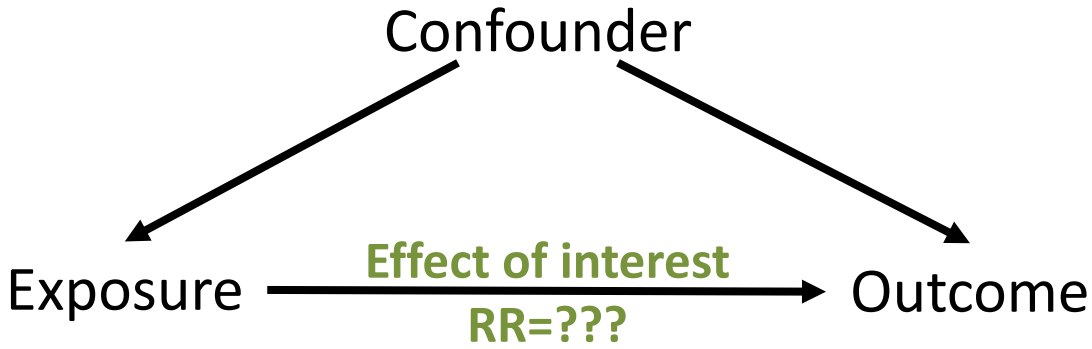
- Results ShinyApp: https://data.ohdsi.org/AntiVegfKidneyFailure/#use

Database	Target	Comparator	Outcome	Max SDM	Shared Max SDM	Equipoise	MDRR	EASE
CCAE	aflibercept	ranibizumab	ESRD	0.065	0.135	0.607	2.05	0.054
CCAE	ranibizumab	bevacizumab	ESRD	0.051	0.097	0.834	1.89	0.054
CCAE	aflibercept	bevacizumab	ESRD	0.055	0.113	0.822	1.82	0.067

All three TC comparisons have at least 50% persons in “empirical equipoise”, which is usually a good sign. The first TC pair has slightly lower proportion of persons in equipoise (we can check out the PS plot).

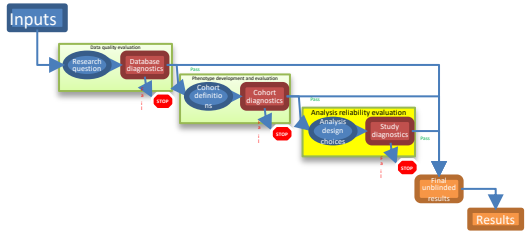
Covariate balance: Standardized mean difference (SMD)





- Confounding variables can bias effect estimates if not properly addressed
- Various design and analysis choices (restriction, matching, propensity score adjustment) offer strategies to reduce the effect of confounding by balancing confounder prevalence in target and comparator cohort
- **Covariate balance**: are all observed baseline characteristics sufficiently similar between target and comparator cohorts?
  - Measured by standardized mean difference (SMD) on each covariate
  - Usually, we want to see **max SMD < 0.1** (rule of thumb)

Covariate balance: Standardized mean difference Examples from LEGEND-HTN



Good:
T = amlodipine
C = atenolol
A = PS matching, on-treatment
DB = CCAE

>45,000 baseline covariates evaluated, many with SMD > 0.1 before matching, but after matching all covariates have SMD <= 0.03

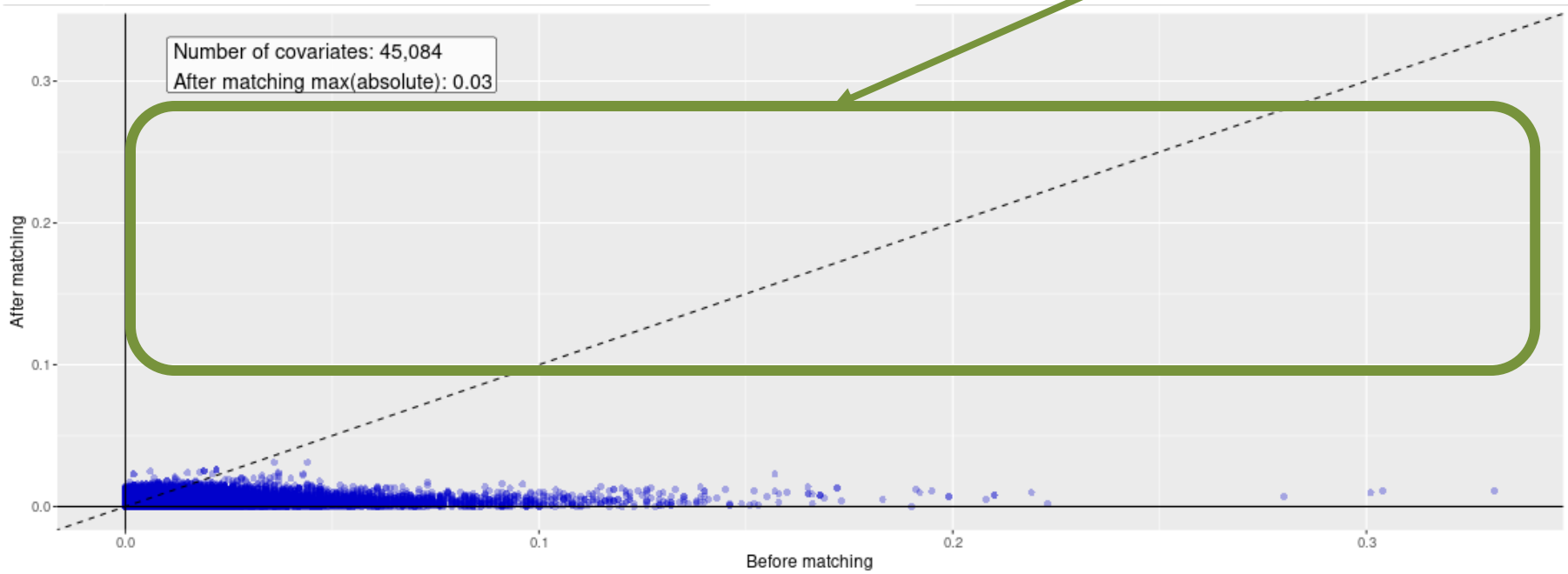
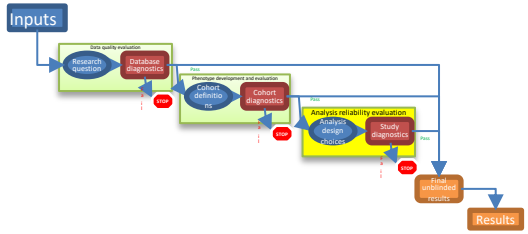


Figure 3. Covariate balance before and after matching. Each dot represents the standardized difference of means for a single covariate before and after matching on the propensity score. Move the mouse arrow over a dot for more details.

Covariate balance: Standardized mean difference Examples from LEGEND-HTN



Bad:
T = candesartan
C = atenolol
A = PS stratification, on-treatment
DB = CCAE

>50,000 baseline covariates evaluated, many with SMD > 0.1 before stratification. After stratification, many covariates have higher SMD than pre-stratification, many covariates with SMD > 0.1

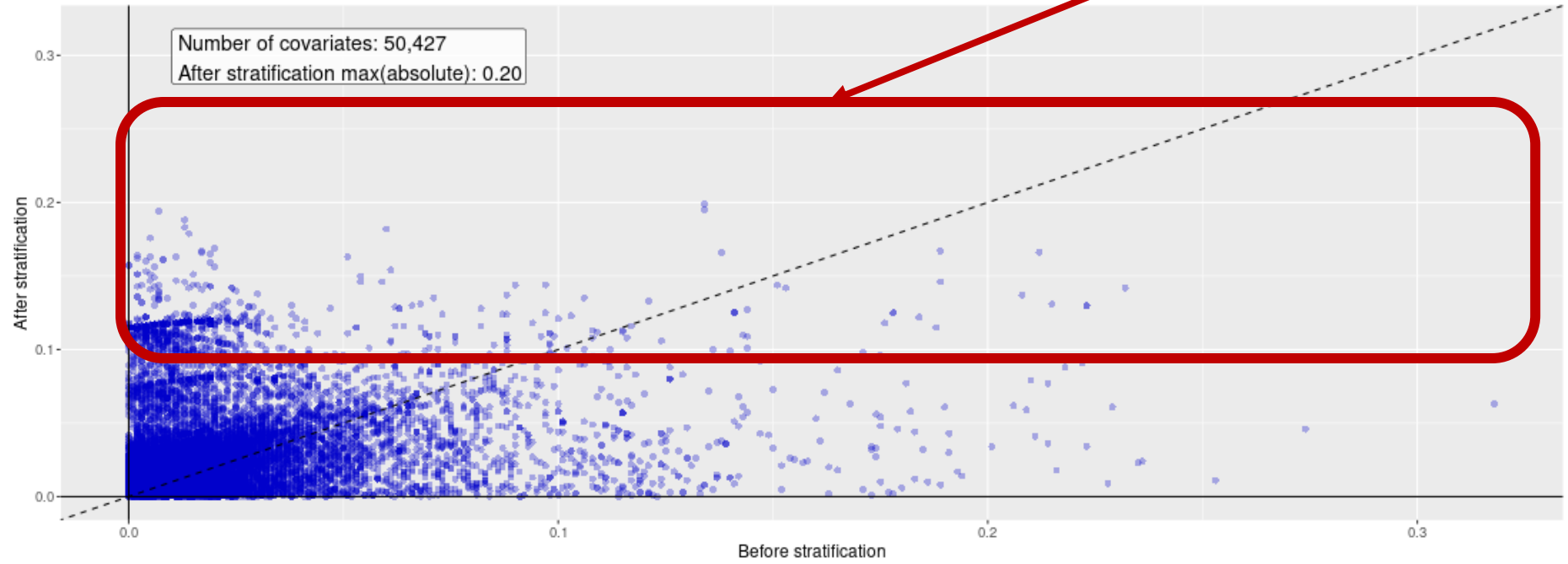


Figure 3. Covariate balance before and after stratification. Each dot represents the standardized difference of means for a single covariate before and after stratification on the propensity score. Move the mouse arrow over a dot for more details.

Covariate balance: SMD in the Anti-VEGF study

Good!

T: ranibizumab
C: bevacizumab
A: PS matching, on treatment
DB: CCAE

Many covariates with SMD > 0.1 before propensity score matching, but all SMD < 0.1 after matching.

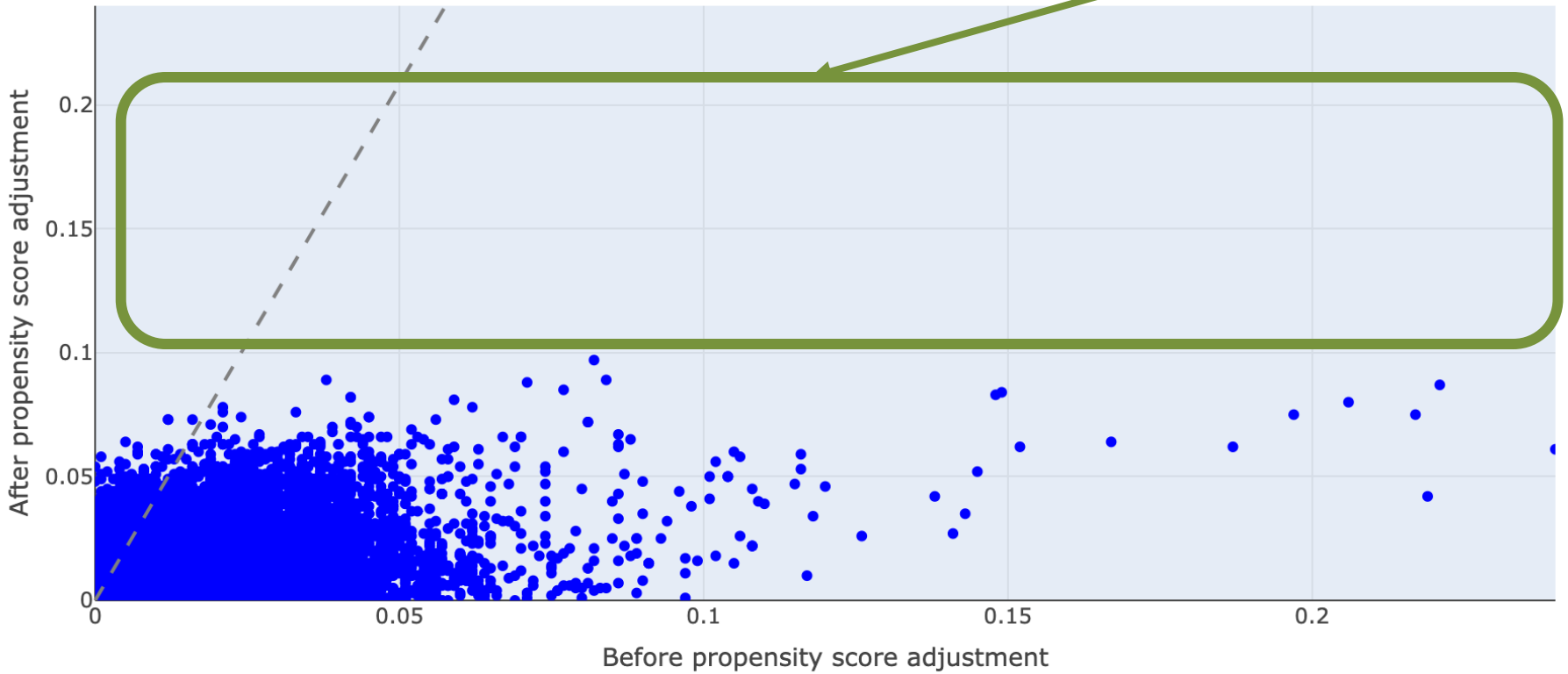
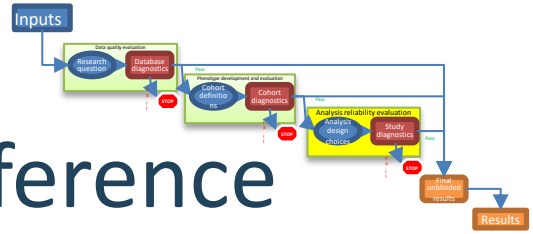


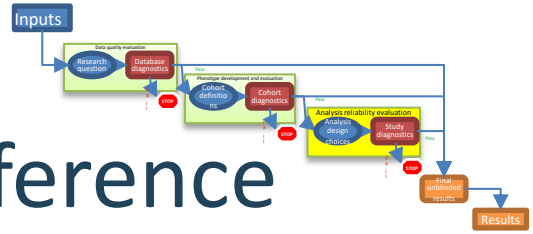
Figure 3. Covariate balance before and after propensity score adjustment. Each dot represents the standardizes difference of means for a single covariate before and after propensity score adjustment on the propensity score. Move the mouse arrow over a dot for more details.

Generalizability: Attrition fraction & standardized mean difference

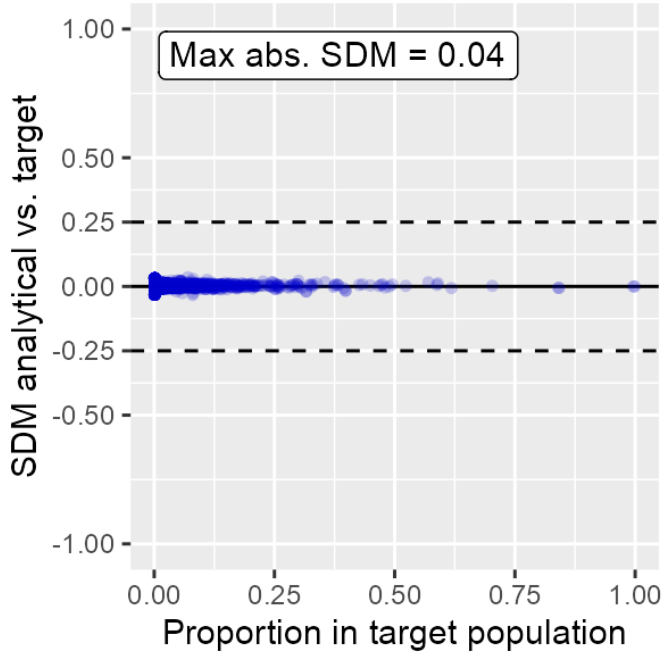


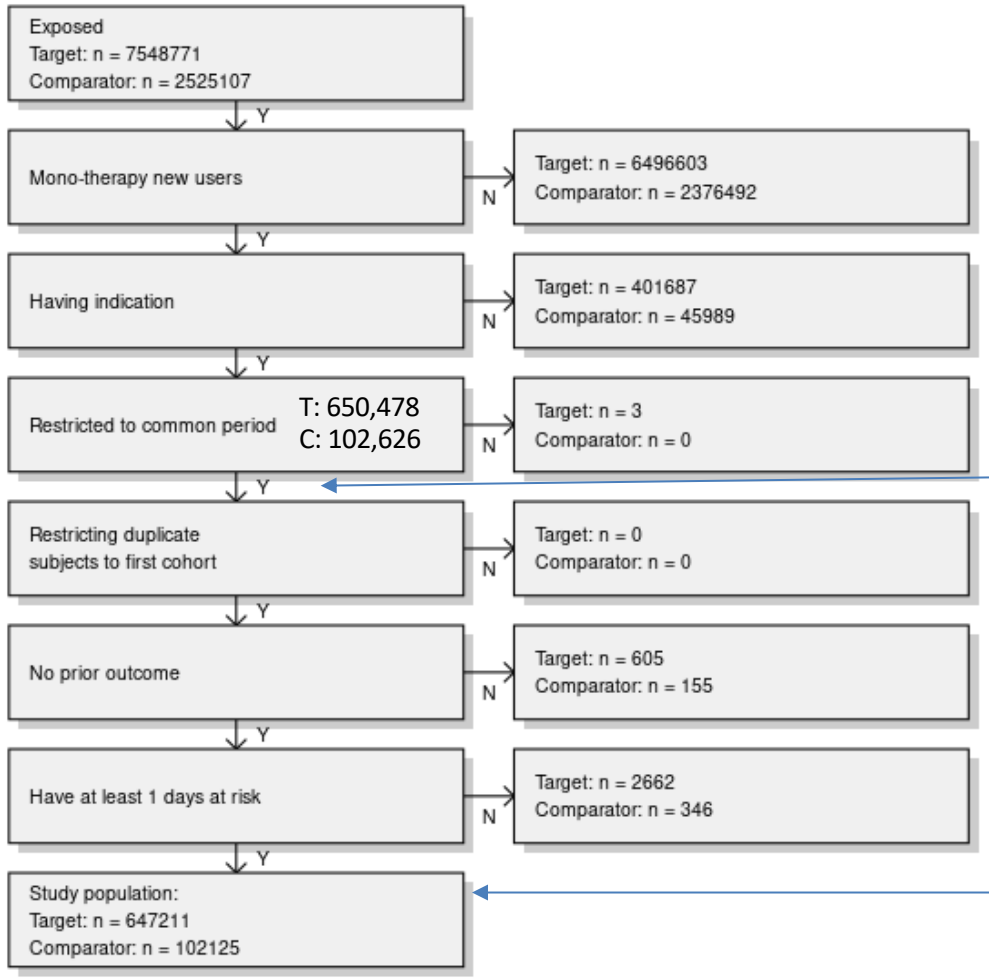
- Generalizability: **to what extent can a study result be applied to a target population of interest?**
- The same design and analytic strategies employed to improve internal validity by reducing confounding can potentially **decrease external validity** by shifting the composition of the analytic cohort away from the original target population
- **Similarity between target population and analytic cohort**:
  - does a substantial fraction of the initial target cohort remain in the analytic target cohort? (**attrition fraction**)
  - are all observed baseline characteristics sufficiently similar between the pre-adjustment target and post-adjustment analytic cohorts? (**SMD**)

Generalizability:
Attrition fraction & standardized mean difference



Good:
T = lisinopril
C = losartan
O = angioedema
A = PS stratification, on-treatment
DB = CCAE



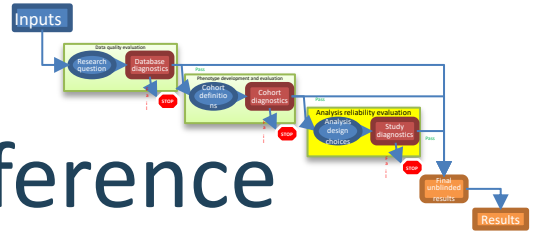


'Target' cohort

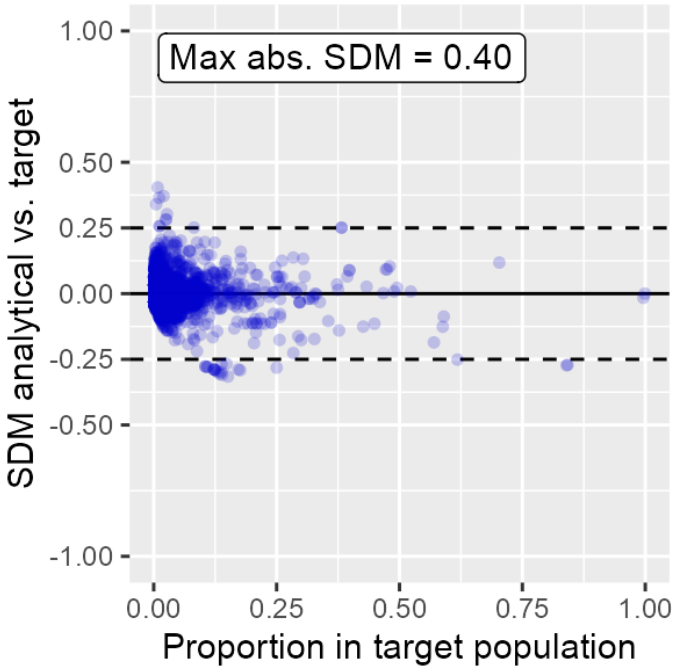
>99% of target population remains in analysis cohort

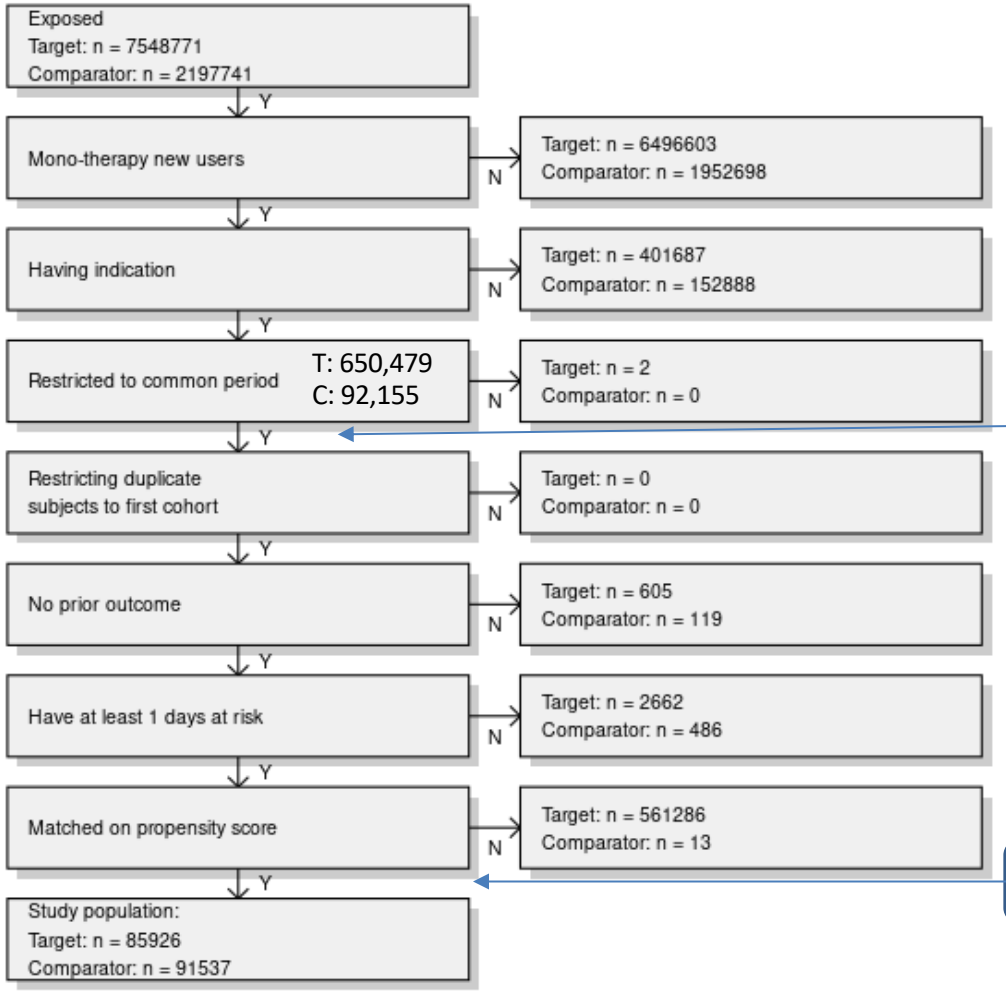
'Analysis' cohort

Generalizability:
Attrition fraction & standardized mean difference



Bad:
T = lisinopril
C = atenolol
O = angioedema
A = PS matching, on-treatment
DB = CCAE



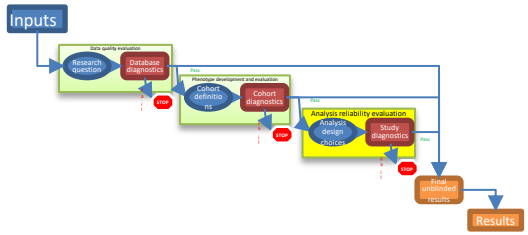


'Target' cohort

Only 13% of target population remains in analysis cohort

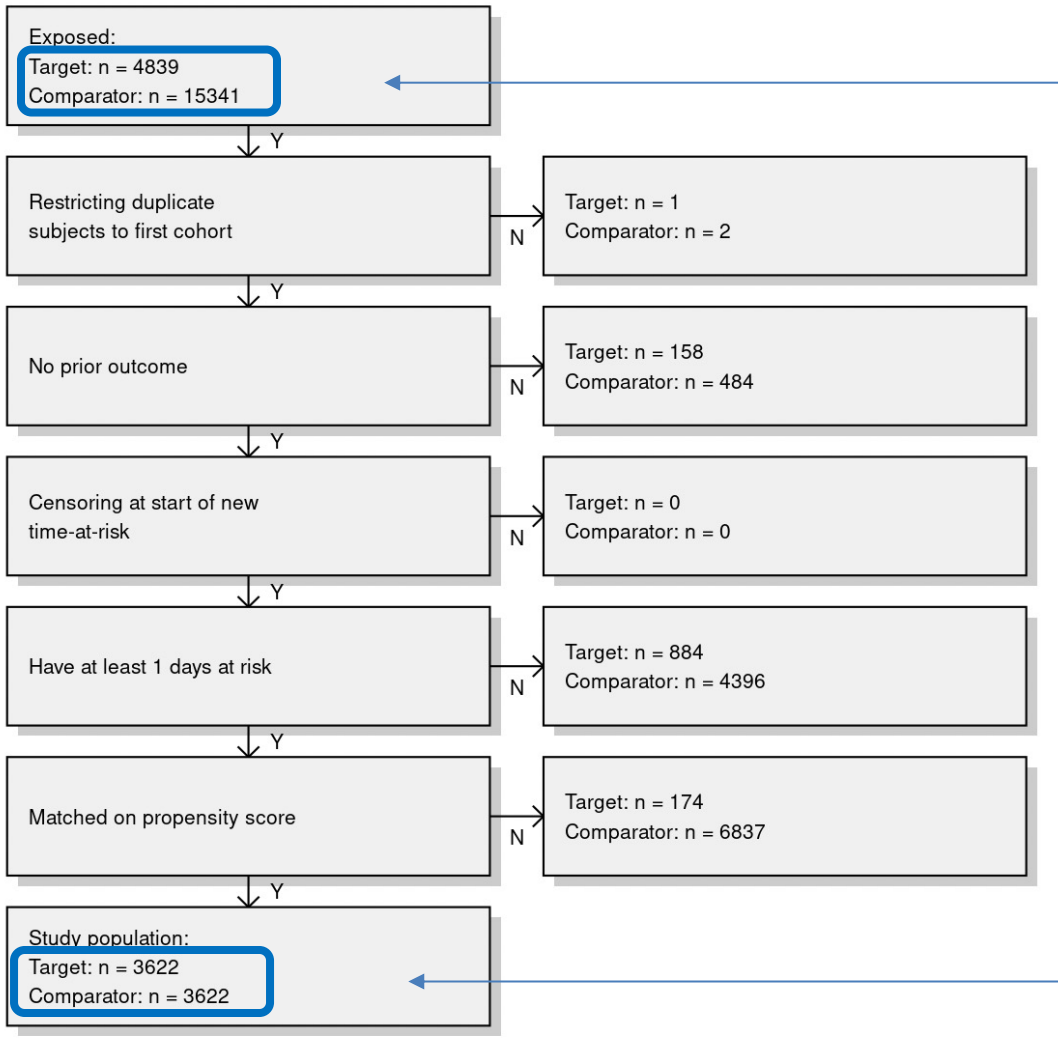
'Analysis' cohort

Generalizability:
In the Anti-VEGF study



T: ranibizumab
C: bevacizumab
O: ESRD
A: PS matching, on treatment
DB: CCAE

Good!

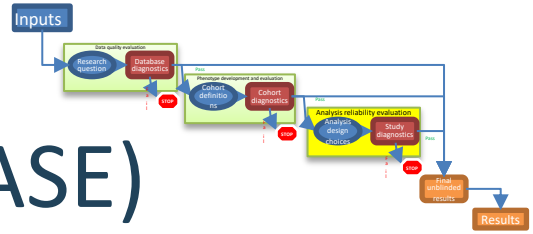


'Target' cohort
T = 4839
C = 15341

>74% of target population remains in analysis cohort

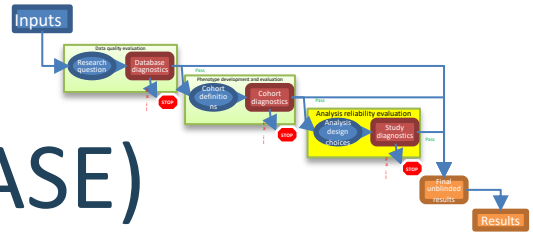
'Analysis' cohort
T = 3622
C = 3622

Residual bias: Expected Absolute Systematic Error (EASE)

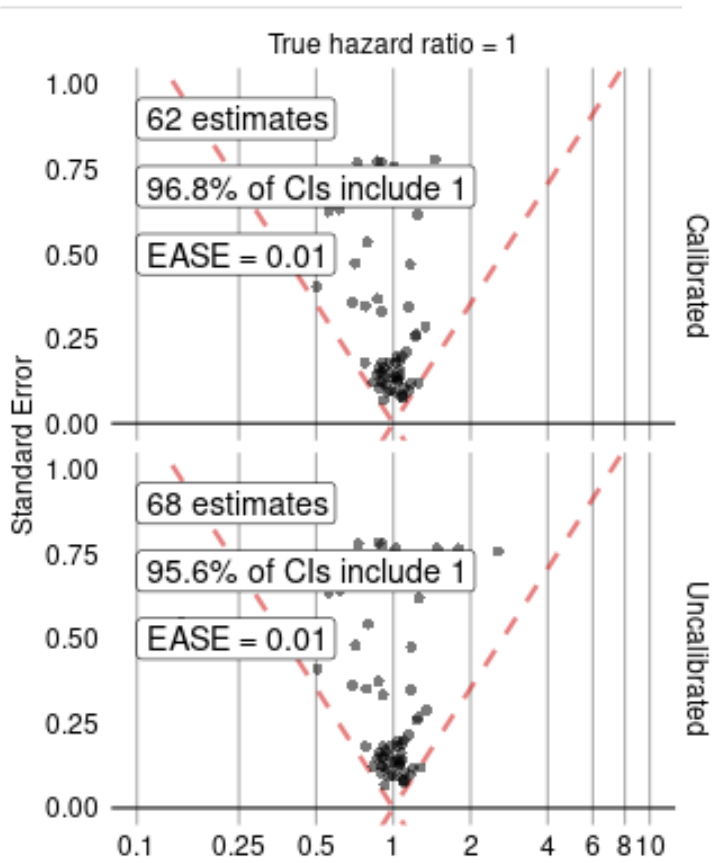


- Residual systematic error can exist due to model misspecification inherent to analysis or data
- Measure bias by expected absolute systematic error (EASE)
  - average of abs(log(estimated RR) – log(true RR)) across negative control outcomes
- **Residual bias**: is the estimated residual bias (EASE) small enough to accept that calibrated effect estimates can be trusted as unbiased?
  - we advocate for empirical calibration, but calibrated results are harder to trust if there is huge bias

Residual bias: Expected Absolute Systematic Error (EASE)



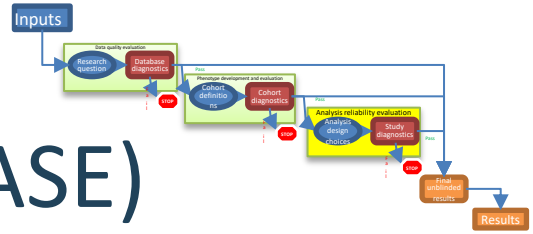
Good:
T = hydrochlorothiazide
C = chlorthalidone
O = acute myocardial infarction
A = PS stratification, on-treatment
DB = CCAE



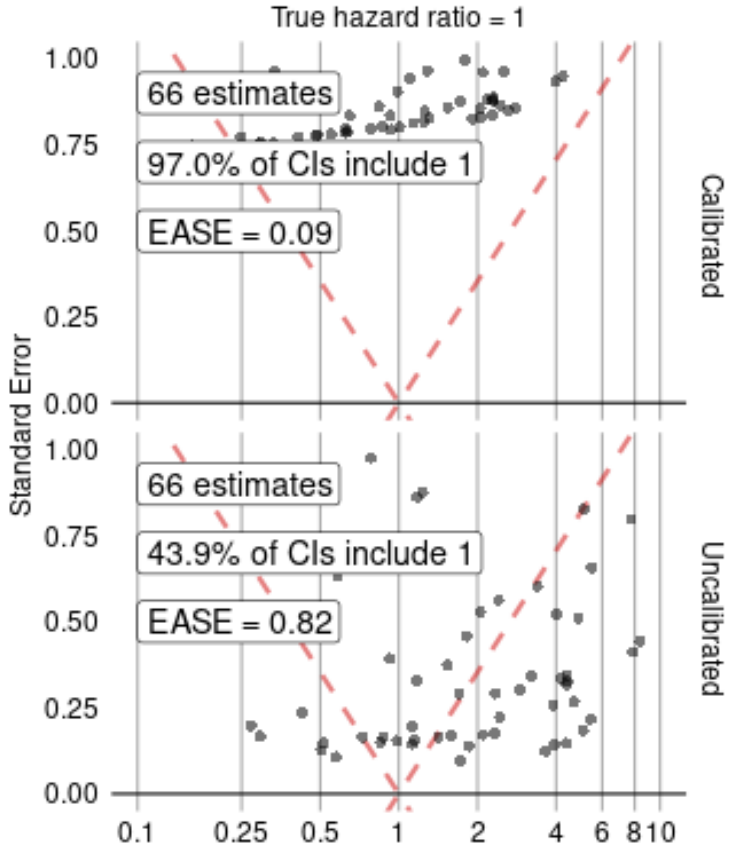
Little residual bias observed (EASE=0.01), so calibration has very little impact on effect estimate (HR=1.54 → HR=1.51)

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
PS stratification, on-treatment	CCAE	1.54	0.88	3.00	0.17	1.51	0.82	2.79	0.18

Residual bias: Expected Absolute Systematic Error (EASE)



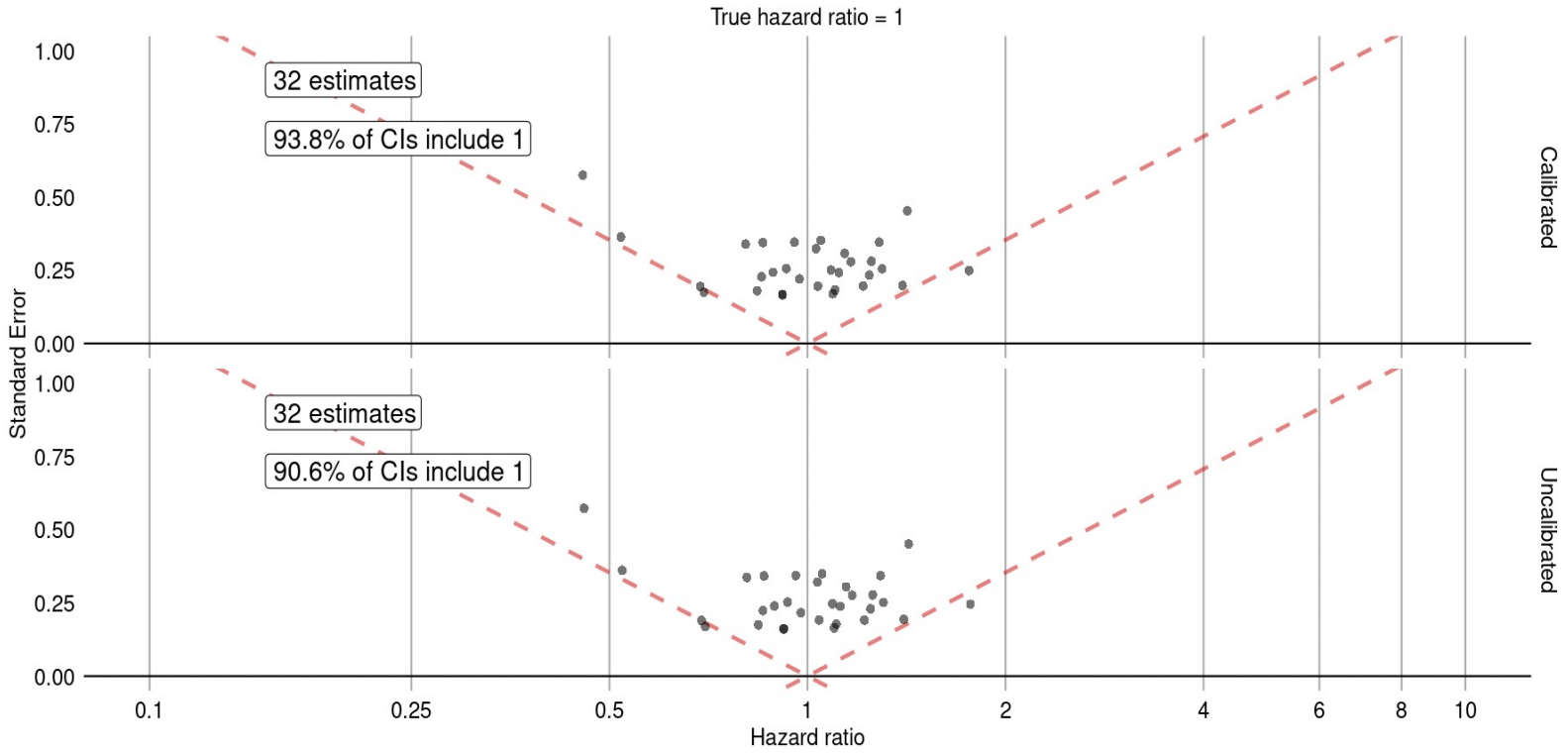
Bad:
T = furosemide
C = labetalol
O = acute myocardial infarction
A = PS stratification, on-treatment
DB = CCAE



Substantial positive bias and variance observed (EASE=0.82), so calibration has substantial impact on effect estimate (HR=5.55, p<0.01 → HR=2.86, p<0.20)

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
PS stratification, on-treatment	CCAE	5.55	3.50	9.25	0.00	2.86	0.59	17.78	0.20

Residual bias:
EASE in the Anti-VEGF study



T: ranibizumab
C: bevacizumab
O: ESRD
A: PS matching, on treatment
DB: CCAE

Little residual bias observed (EASE=0.054). Calibration has little impact on effect estimate (HR=0.79 unchanged).

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
Cohort method, On treatment	IBM CCAE	0.79	0.50	1.24	0.30	0.79	0.50	1.24	0.31

Good!

Remarks: diagnostics thresholds are rules of thumb
We can pre-specify thresholds given empirical objectives

Diagnostics metric	Literature-derived	Strategus interface	Data-driven (LEGEND-HTN)
Statistical power (MDRR)	< 10	< Inf; < 10 (SCCS)	-
Equipoise	> 0.50	> 0.20	> 0.50
Covariate balance (SDM)	< 0.10	< 0.10	< 0.50
Generalizability (attrition)	-	<= 1	-
Systematic error (EASE)	< 0.25	< 0.25	-

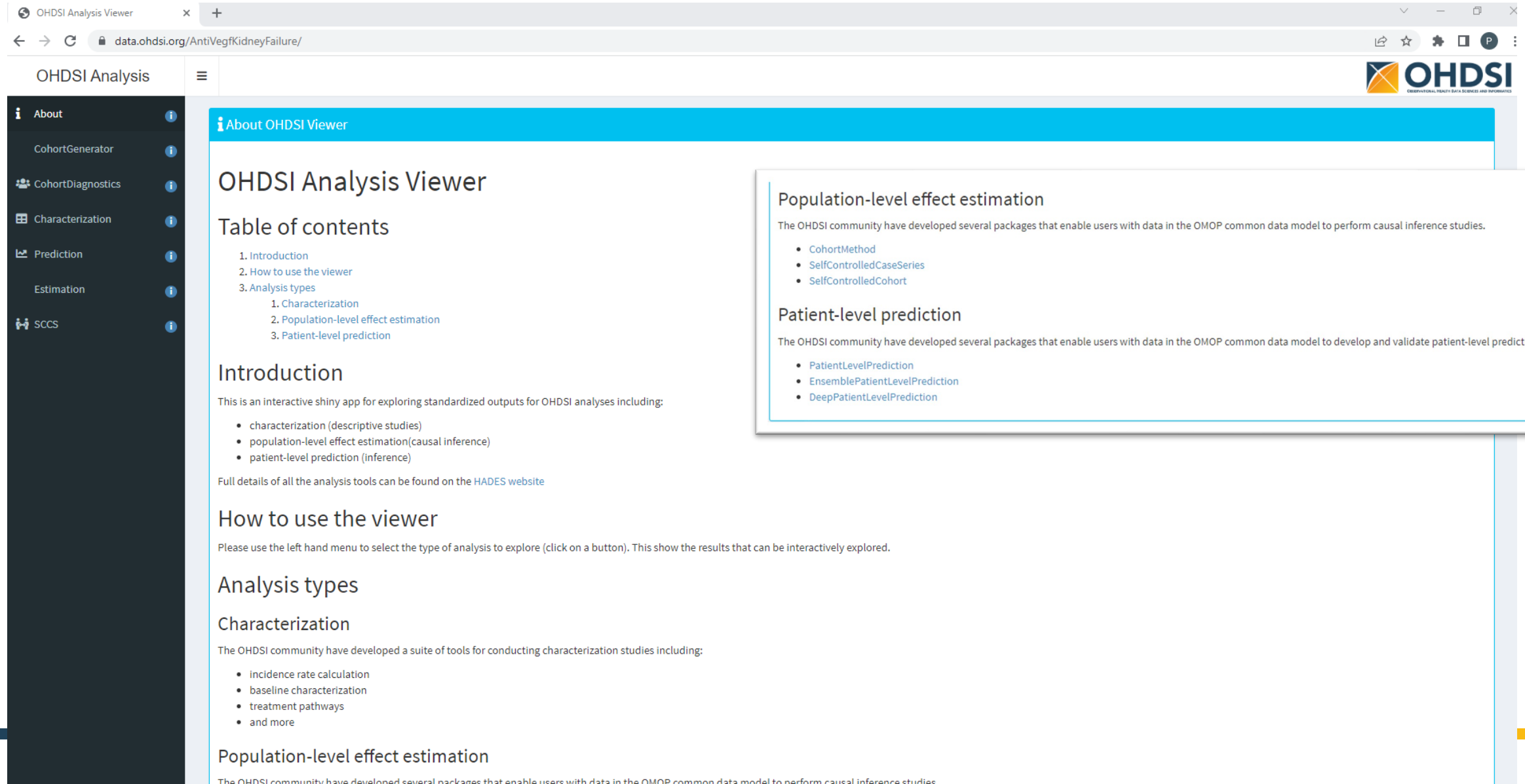
Interpretability: MDRR
Internal validity: equipoise, SDM , EASE
External validity: attrition (or SDM)

We need to pre-specify thresholds and run diagnostics before seeing results!

Let’s see the Strategus user interface

(Already seen pieces of cohort diagnostics)

Strategus standard user interface



Cohort Diagnostics

OHDSI Analysis

- About
- CohortGenerator
- CohortDiagnostics
- Characterization
- Prediction
- Estimation
- SCCS

Cohort Level Diagnostics

Select Report

Cohort Counts

Database(s)

IBM Health MarketScan® Commercial Claims and Encounters Database

Cohorts

C1782164: [SOS] End-stage renal disease, C1782480: [SOS Phenotype Devt] persons with blinding diseases, C1782481: [SOS Phenotype Devt] ranibizumab exposures after new use with 3 exposures in 21-70d windows, C1782482: [SOS Phenotype Devt] bevacizumab expo

Cohort Counts

C1782164: [SOS] End-stage renal disease
C1782480: [SOS Phenotype Devt] persons with blinding diseases
C1782481: [SOS Phenotype Devt] ranibizumab exposures after new use with 3 exposures in 21-70d windows
C1782482: [SOS Phenotype Devt] bevacizumab exposures after new use with 3 exposures in 21-70d windows
C1782483: [SOS Phenotype Devt] aflibercept exposures after new use with 3 exposures in 21-70d windows

Display

(•) Both ( ) Persons ( ) Records

Search

			IBM Health MarketScan® Commercial Claims and Encounters D...	
	Cohort Id	Cohort Name	Persons	Records
( )	1782164	[SOS] End-stage renal disease	249,258	249,258
( )	1782480	[SOS Phenotype Devt] persons with blinding diseases	1,294,165	1,294,165
( )	1782481	[SOS Phenotype Devt] ranibizumab exposures after new use with 3 exposures in 21-70d windows	4,846	6,022
( )	1782482	[SOS Phenotype Devt] bevacizumab exposures after new use with 3 exposures in 21-70d windows	15,440	19,874
( )	1782483	[SOS Phenotype Devt] aflibercept exposures after new use with 3 exposures in 21-70d windows	4,115	5,192

Characterization: Time-to-event

OHDSI Analysis

- About
- CohortGenerator
- CohortDiagnostics
- Characterization
- Prediction
- Estimation
- SCCS

Characterization Viewer

Target Viewer | Outcome Stratified | Incidence Rate | **Time To Event** | Dechallenge Rechallenge

Time-to-events

Options

Target id: [SOS Phenotype Devt] aflibercept exposures after new use with 3 exposures in 21-70d windows

Outcome id: [SOS] End-stage renal disease

Generate Report

Selected:

Target: [SOS Phenotype Devt] aflibercept exposures after new use with 3 exposures in 21-70d windows

Outcome: [SOS] End-stage renal disease

Results

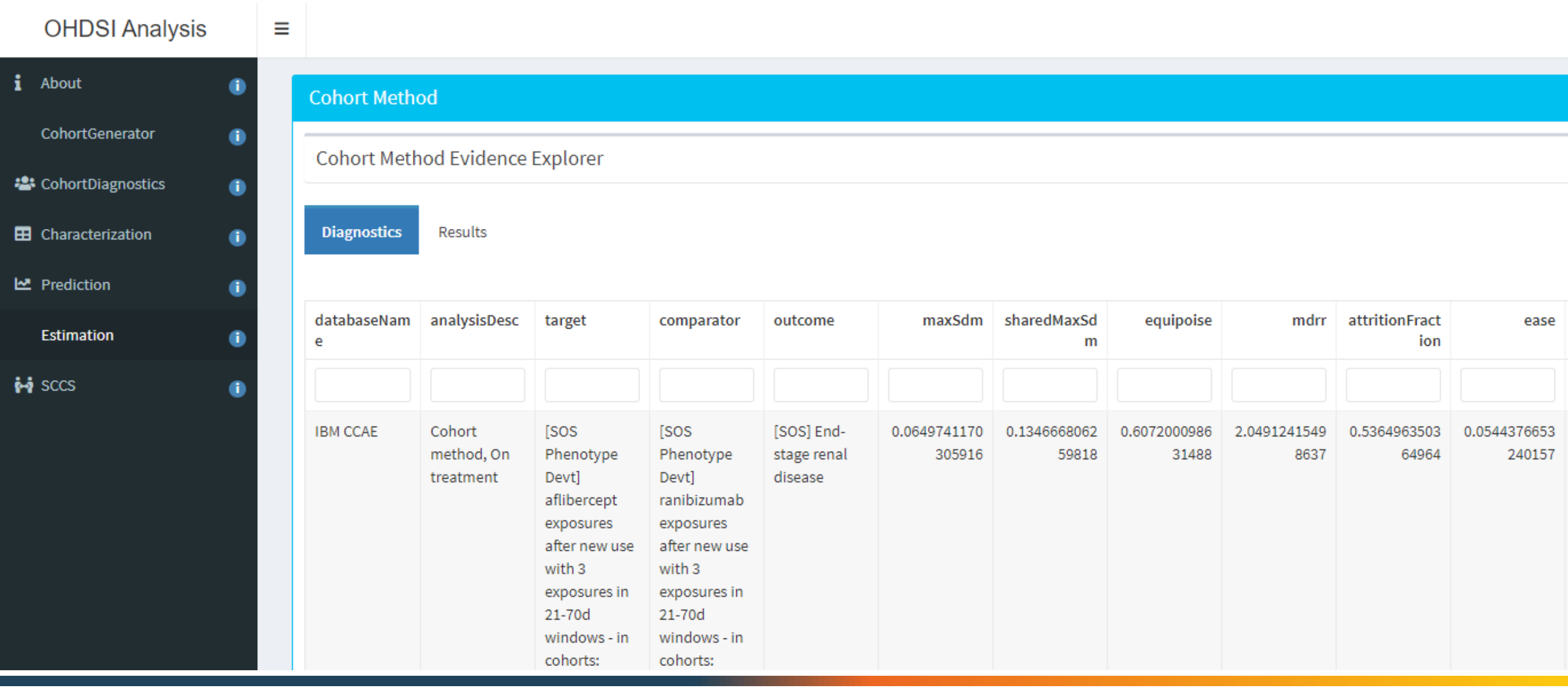
Databases:

- [x] IBM CCAE

Timespan:

- [x] per 30-day
- [x] per 365-day
- [x] per 1-day

Estimation: Cohort method diagnostics

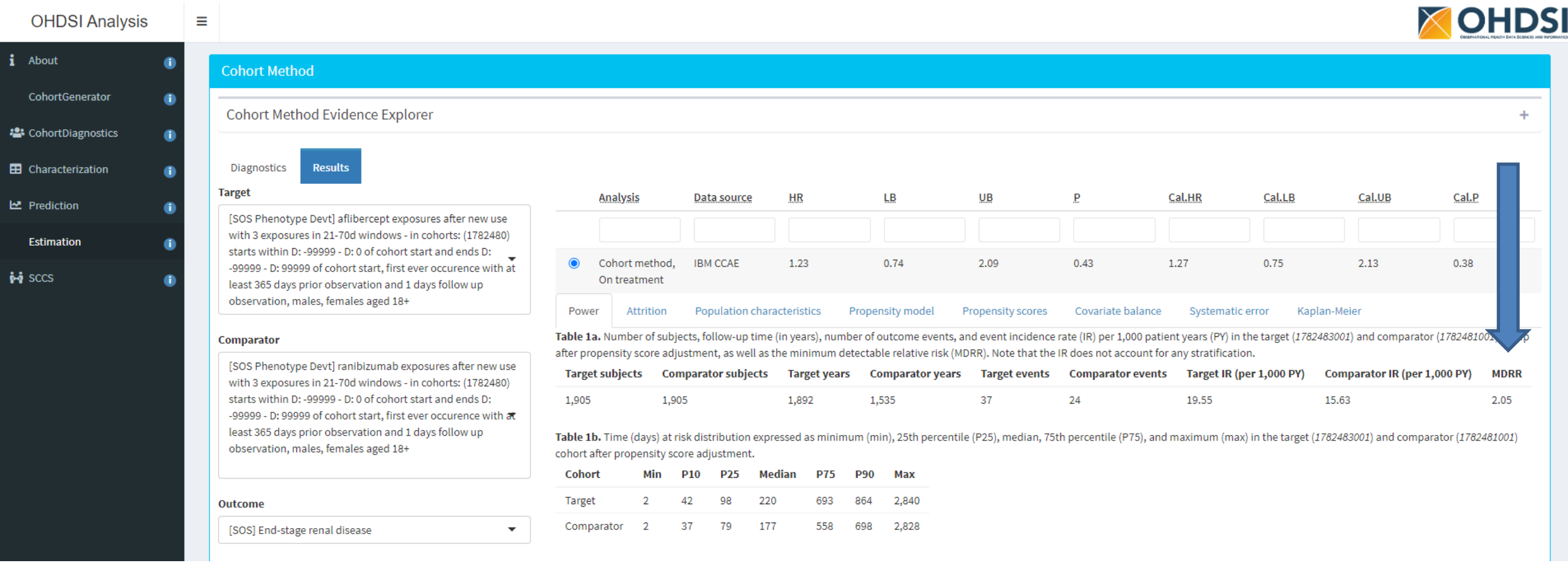


Estimation: Cohort method diagnostics pass/fail based on a priori decision thresholds

target	comparator	outcome	maxSdm	sharedMaxSdm	equipoise	mdrr	attritionFraction	ease	balanceDiagnostic	sharedBalanceDiagnostic	equipoiseDiagnostic	mdrrDiagnostic	attritionDiagnostic	easeDiagnostic	unblind
[SOS Phenotype Devt] aflibercept exposures after new use with 3 exposures in 21-70d windows - in cohorts:	[SOS Phenotype Devt] ranibizumab exposures after new use with 3 exposures in 21-70d windows - in cohorts:	[SOS] End-stage renal disease	0.0649741170305916	0.134666806259818	0.607200098631488	2.049124154981637	0.536496350364964	0.0544376653240157	PASS	FAIL	PASS	PASS	PASS	PASS	1
			<0.1	**NA**	**>0.2**	**NA**	**<1**	**<0.25**							


```
{
  "sharedResources": [
  "moduleSpecifications": [
    {
    {
    {
    {
    {
      "module": "CohortMethodModule",
      "version": "0.1.0",
      "remoteRepo": "github.com",
      "remoteUsername": "ohdsi",
      "settings": {
        "cmAnalysisList": [
        "targetComparatorOutcomesList": [
        "refitPsForEveryOutcome": false,
        "refitPsForEveryStudyPopulation": false,
        "cmDiagnosticThresholds": {
          "mdrrThreshold": "Inf",
          "easeThreshold": 0.25,
          "sdmThreshold": 0.1,
          "equipoiseThreshold": 0.2,
          "attritionFractionThreshold": 1,
          "attr_class": "CmDiagnosticThresholds"
        }
      },
```

Estimation: Cohort method diagnostic drilldown: Minimum Detectable Relative Risk (MDRR)



Estimation: Cohort method diagnostic drilldown: Attrition Fraction

OHDSI Analysis

- About
- CohortGenerator
- CohortDiagnostics
- Characterization
- Prediction
- Estimation
- SCCS

Cohort Method

Cohort Method Evidence Explorer

Diagnostics | **Results**

Target

[SOS Phenotype Devt] aflibercept exposures after new use with 3 exposures in 21-70d windows - in cohorts: (1782480) starts within D: -99999 - D: 0 of cohort start and ends D: -99999 - D: 99999 of cohort start, first ever occurence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+

Comparator

[SOS Phenotype Devt] ranibizumab exposures after new use with 3 exposures in 21-70d windows - in cohorts: (1782480) starts within D: -99999 - D: 0 of cohort start and ends D: -99999 - D: 99999 of cohort start, first ever occurence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+

Outcome

[SOS] End-stage renal disease

Data source

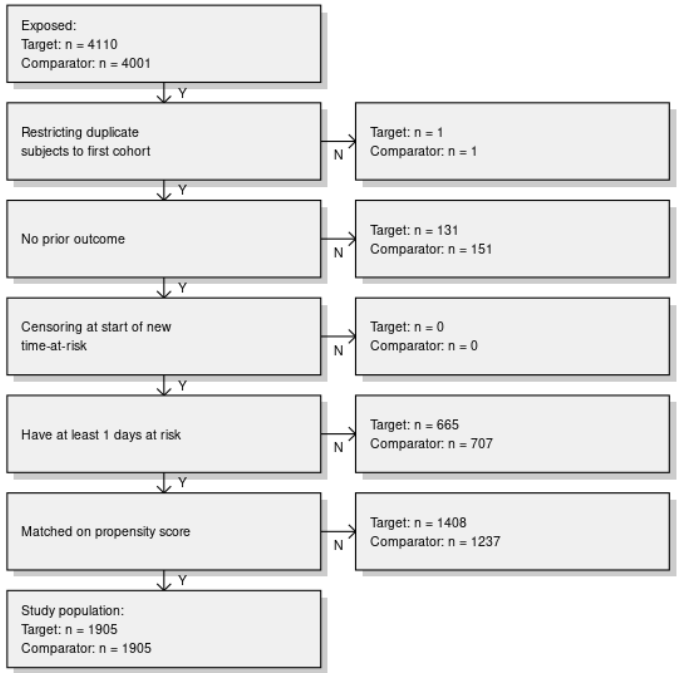
☑ IBM CCAE

Analysis

☑ Cohort method, On treatment

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
◉ Cohort method, On treatment	IBM CCAE	1.23	0.74	2.09	0.43	1.27	0.75	2.13	0.38

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



Estimation: Cohort method diagnostic drilldown: Equipoise

OHDSI Analysis

- About
- CohortGenerator
- CohortDiagnostics
- Characterization
- Prediction
- Estimation
- SCCS

Cohort Method

Cohort Method Evidence Explorer

Diagnostics | **Results**

Target

[SOS Phenotype Devt] aflibercept exposures after new use with 3 exposures in 21-70d windows - in cohorts: (1782480) starts within D: -99999 - D: 0 of cohort start and ends D: -99999 - D: 99999 of cohort start, first ever occurence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+

Comparator

[SOS Phenotype Devt] ranibizumab exposures after new use with 3 exposures in 21-70d windows - in cohorts: (1782480) starts within D: -99999 - D: 0 of cohort start and ends D: -99999 - D: 99999 of cohort start, first ever occurence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+

Outcome

[SOS] End-stage renal disease

Data source

☑ IBM CCAE

Analysis

☑ Cohort method, On treatment

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
◉ Cohort method, On treatment	IBM CCAE	1.23	0.74	2.09	0.43	1.27	0.75	2.13	0.38

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

ccurence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+ [SOS Phenotype Devt] ranibizumab exposures after new use with 3 exposures in 21-70d windows - in

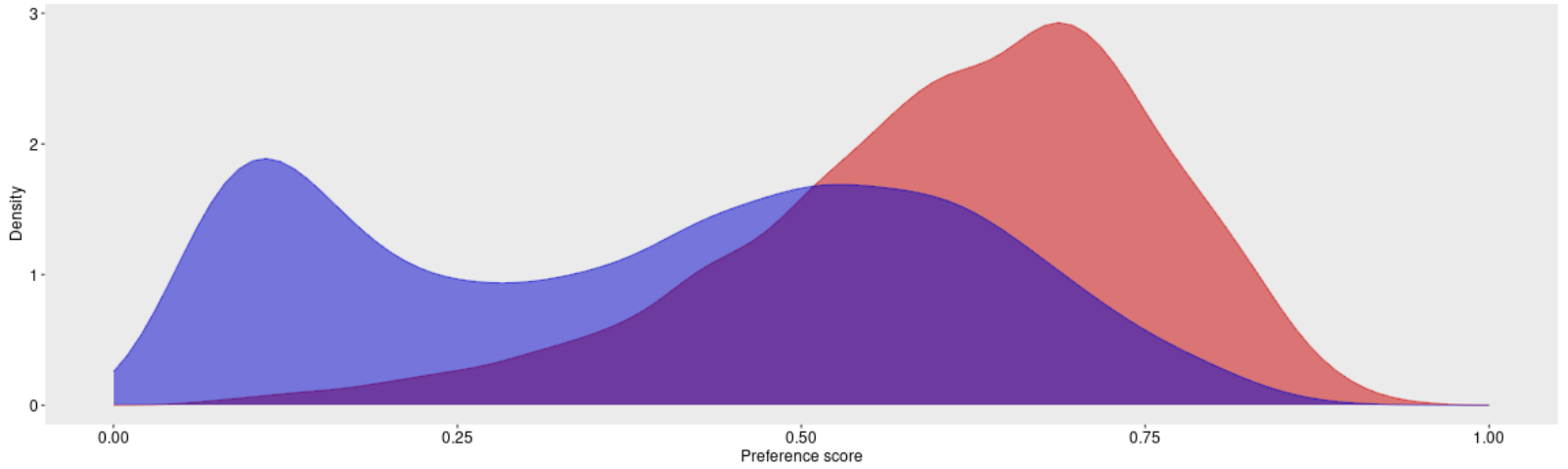


Figure 2. Preference score distribution. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups. A higher overlap indicates subjects in the two groups were more similar in terms of their predicted probability of receiving one treatment over the other.

Download plot as PNG | Download plot as PDF

Estimation: Cohort method diagnostic drilldown: Covariate balance maxSDM

OHDSI Analysis

- About
- CohortGenerator
- CohortDiagnostics
- Characterization
- Prediction
- Estimation
- SCCS

Cohort Method

Cohort Method Evidence Explorer

Diagnostics | **Results**

Target

[SOS Phenotype Devt] aflibercept exposures after new use with 3 exposures in 21-70d windows - in cohorts: (1782480) starts within D: -99999 - D: 0 of cohort start and ends D: -99999 - D: 99999 of cohort start, first ever occurence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+

Comparator

[SOS Phenotype Devt] ranibizumab exposures after new use with 3 exposures in 21-70d windows - in cohorts: (1782480) starts within D: -99999 - D: 0 of cohort start and ends D: -99999 - D: 99999 of cohort start, first ever occurence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+

Outcome

[SOS] End-stage renal disease

Data source

☑ IBM CCAE

Analysis

☑ Cohort method, On treatment

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
Cohort method, On treatment	IBM CCAE	1.23	0.74	2.09	0.43	1.27	0.75	2.13	0.38

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

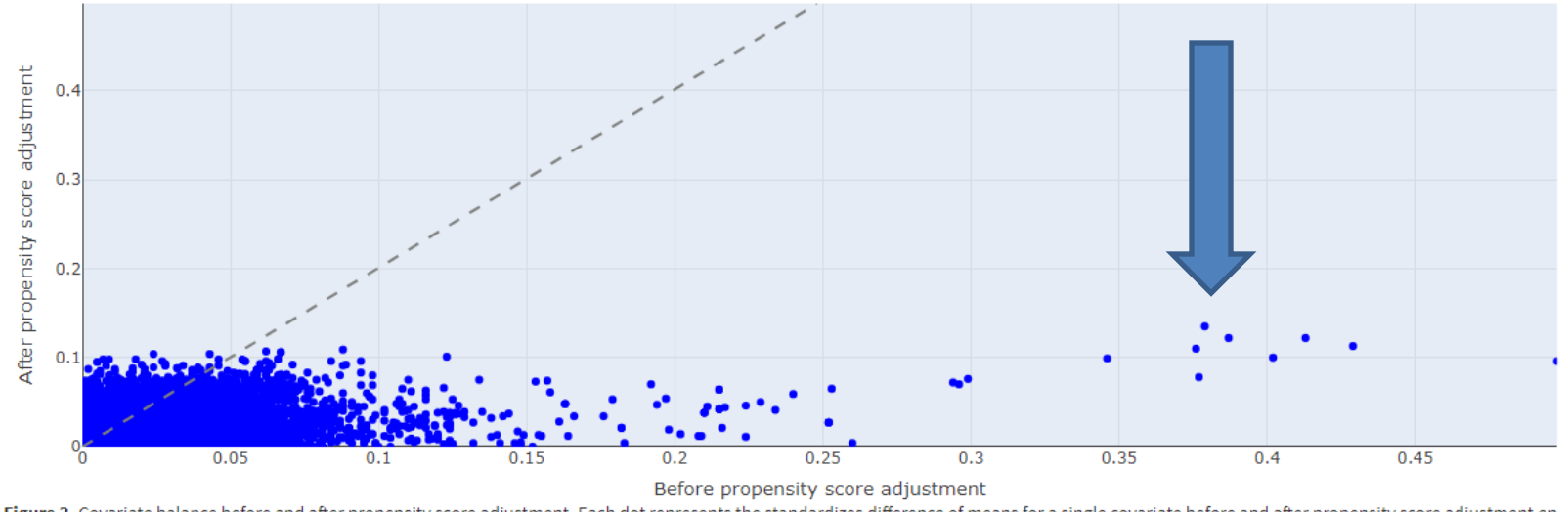


Figure 3. Covariate balance before and after propensity score adjustment. Each dot represents the standardizes difference of means for a single covariate before and after propensity score adjustment on the propensity score. Move the mouse arrow over a dot for more details.

Download

Estimation: Cohort method diagnostic drilldown: Expected Average Systematic Error (EASE)

OHDSI Analysis

- About
- CohortGenerator
- CohortDiagnostics
- Characterization
- Prediction
- Estimation
- SCCS

Cohort Method

Cohort Method Evidence Explorer

Diagnostics | **Results**

Target

[SOS Phenotype Devt] aflibercept exposures after new use with 3 exposures in 21-70d windows - in cohorts: (1782480) starts within D: -99999 - D: 0 of cohort start and ends D: -99999 - D: 99999 of cohort start, first ever occurence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+

Comparator

[SOS Phenotype Devt] ranibizumab exposures after new use with 3 exposures in 21-70d windows - in cohorts: (1782480) starts within D: -99999 - D: 0 of cohort start and ends D: -99999 - D: 99999 of cohort start, first ever occurence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+

Outcome

[SOS] End-stage renal disease

Data source

☑ IBM CCAE

Analysis

☑ Cohort method, On treatment

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
Cohort method, On treatment	IBM CCAE	1.23	0.74	2.09	0.43	1.27	0.75	2.13	0.38

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

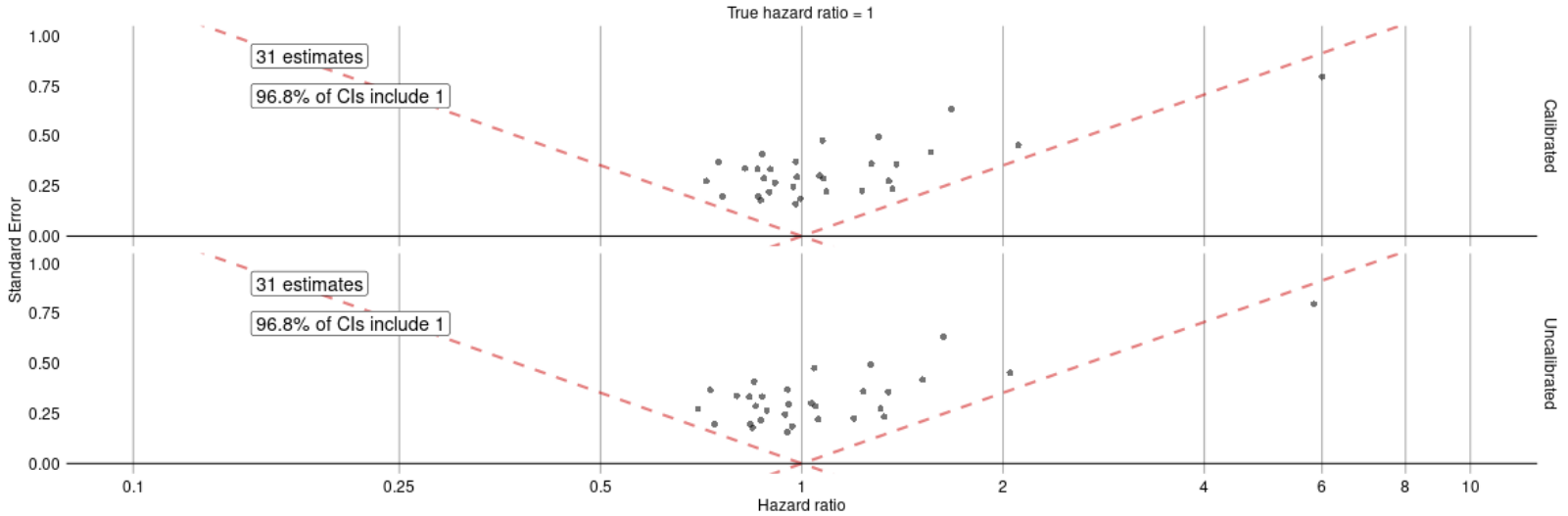
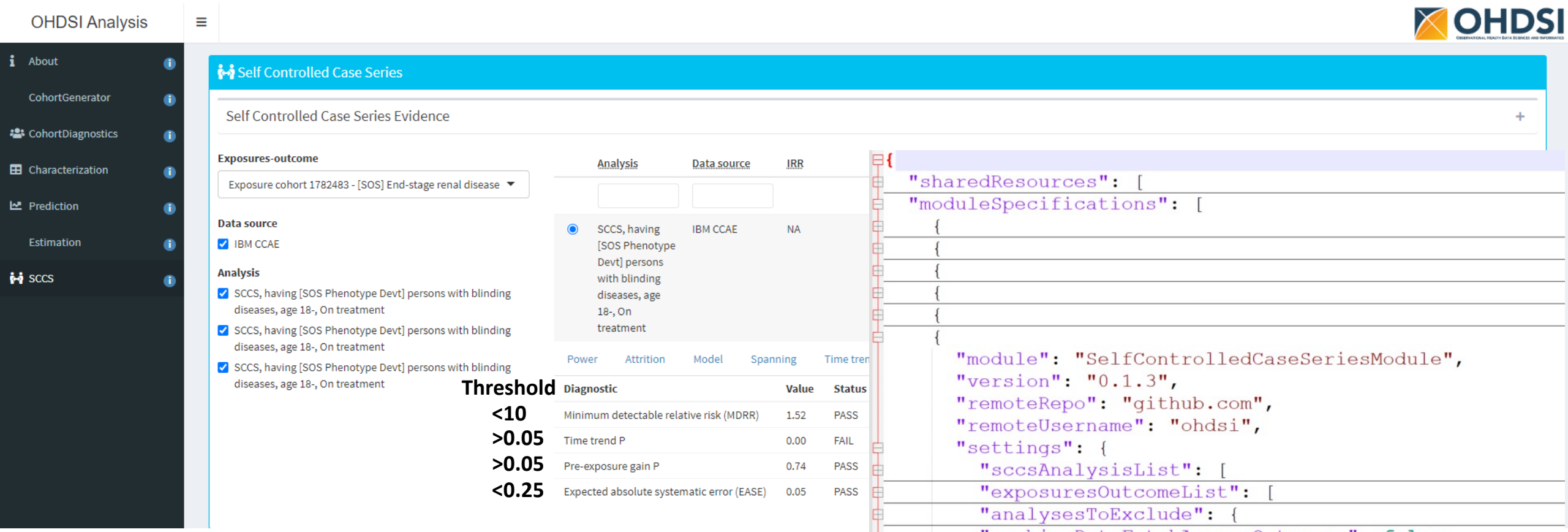


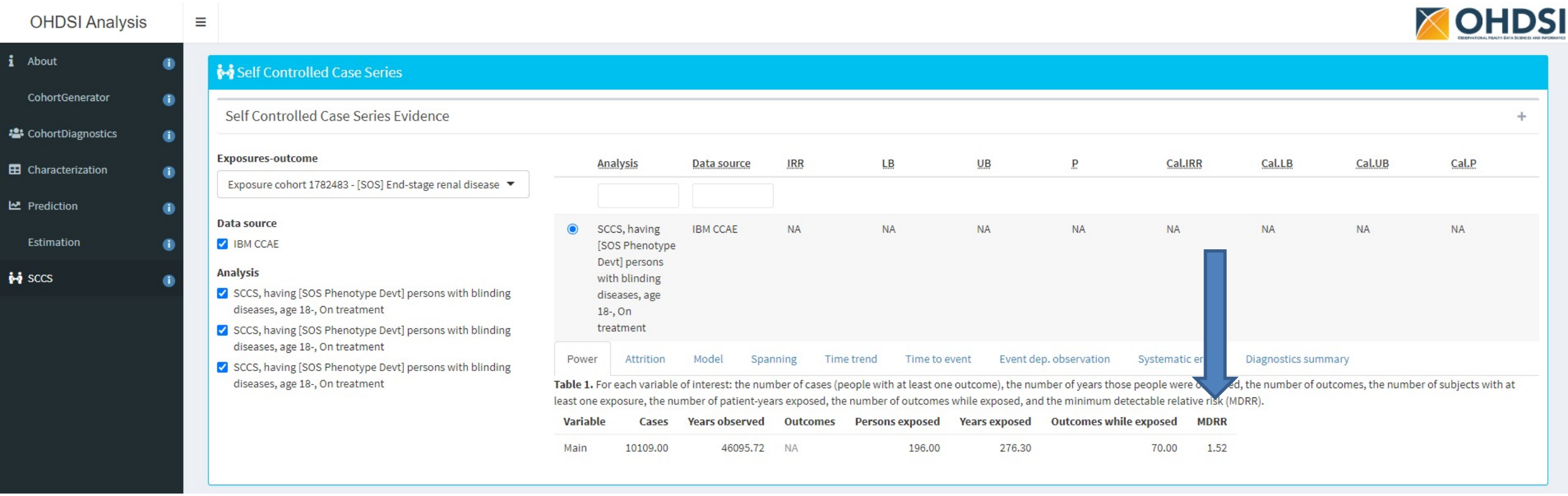
Figure 4. Systematic error. Effect size estimates for the negative controls (true hazard ratio = 1) and positive controls (true hazard ratio > 1), before and after calibration. Estimates below the diagonal dashed lines are statistically significant (alpha = 0.05) different from the true effect size. A well-calibrated estimator should have the true effect size within the 95 percent confidence interval 95 percent of times.

Download plot as PNG | Download plot as PDF

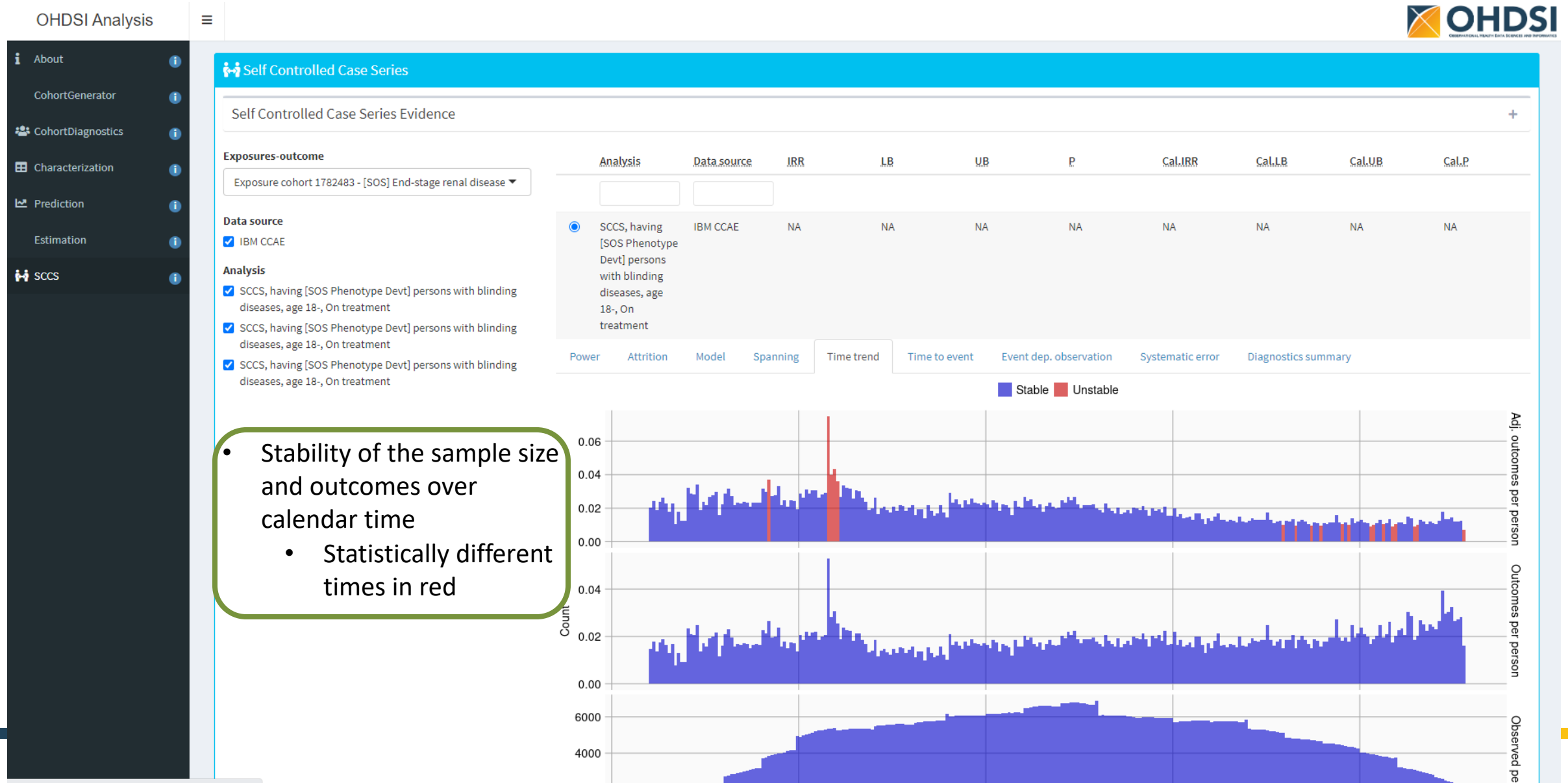
Estimation: Self-controlled case series (SCCS) pass/fail based on a priori decision thresholds



Estimation: SCCS diagnostic drilldown: Minimum detectable relative risk (MDRR)



Estimation: SCCS diagnostic drilldown: Time trend



- Stability of the sample size and outcomes over calendar time
  - Statistically different times in red

Estimation: SCCS diagnostic drilldown: Bad time trend

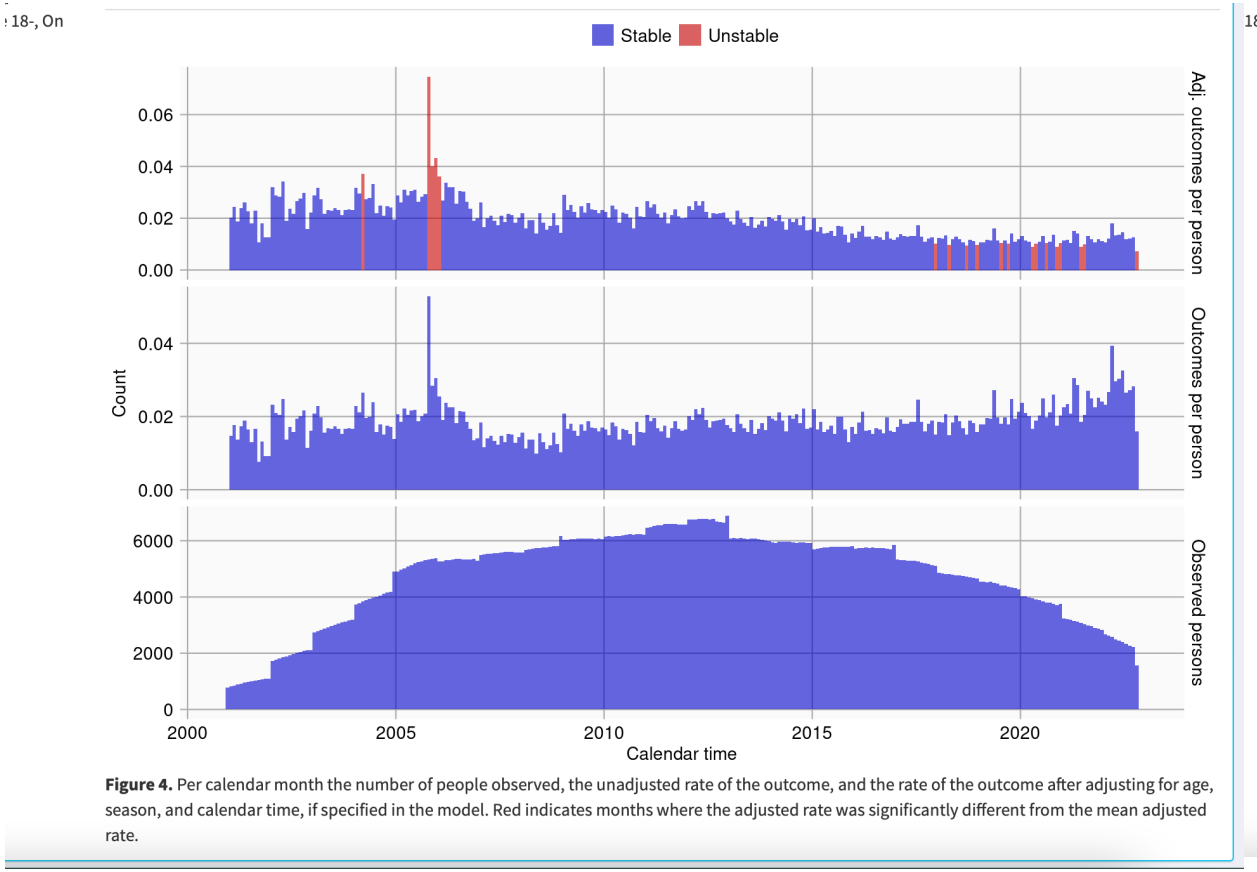


Figure 4. Per calendar month the number of people observed, the unadjusted rate of the outcome, and the rate of the outcome after adjusting for age, season, and calendar time, if specified in the model. Red indicates months where the adjusted rate was significantly different from the mean adjusted rate.

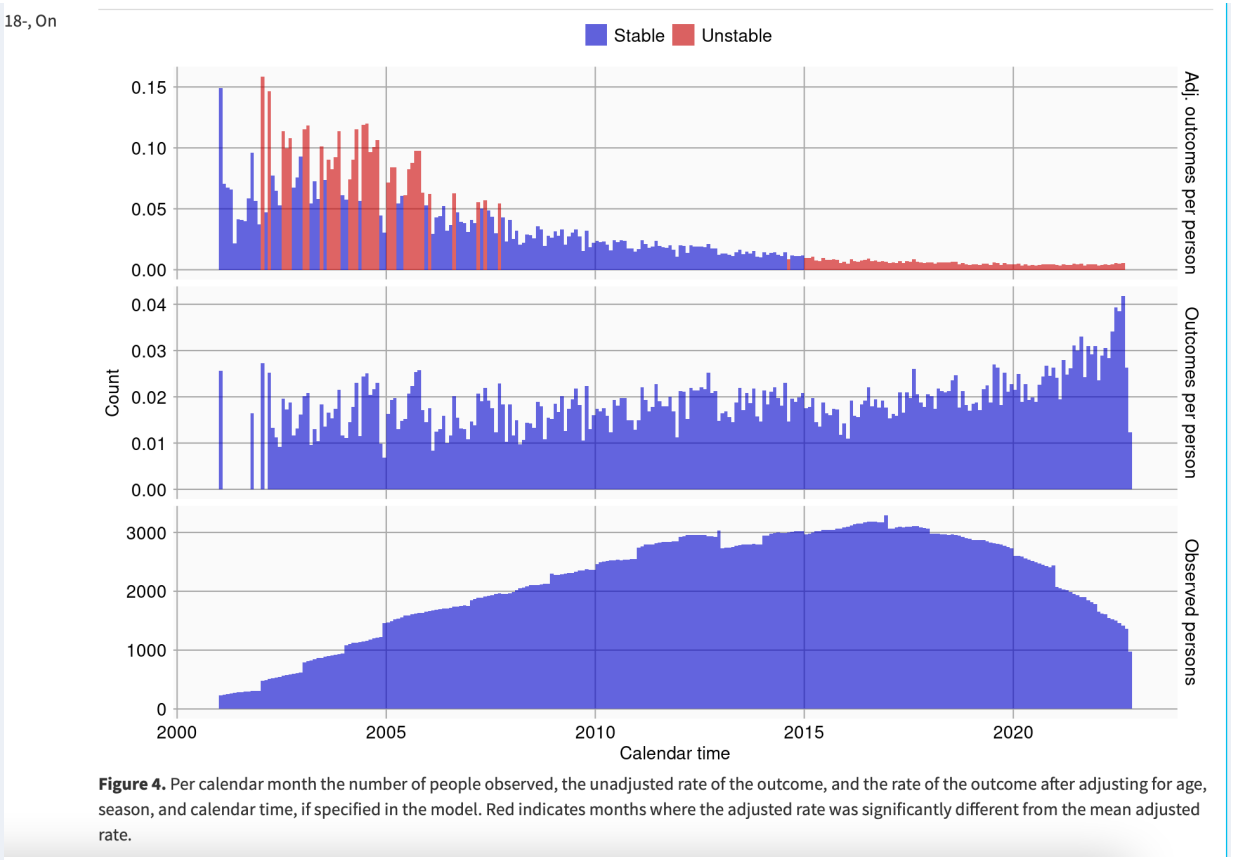


Figure 4. Per calendar month the number of people observed, the unadjusted rate of the outcome, and the rate of the outcome after adjusting for age, season, and calendar time, if specified in the model. Red indicates months where the adjusted rate was significantly different from the mean adjusted rate.

Estimation: SCCS diagnostic drilldown: Time to event

- Distribution of outcomes over time w.r.t. index event
  - Are there more outcomes just before exposure?

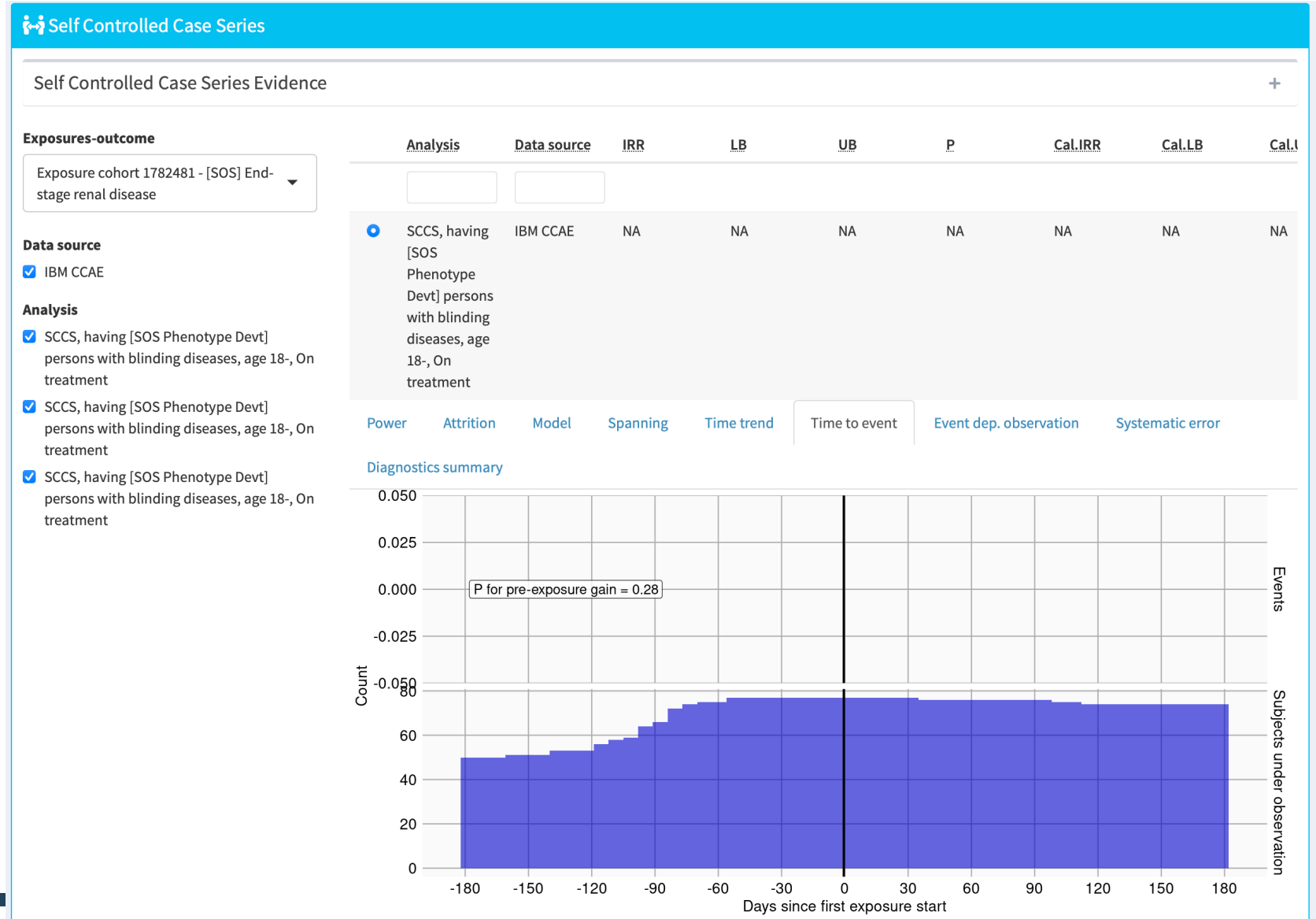


Figure 5. The number of events and subjects observed per week relative to the start of the first exposure (indicated by the thick vertical line).

Patient-level prediction (PLP): View diagnostics

Prediction Viewer

Model Designs Summary

Design ID	Model Type	Target Pop	Outcome	TAR	min AUROC	mean AUROC	max AUROC	Num. Diagnostic Dbs	Num. Development Dbs	Num. Validation Dbs			
1	logistic	Cohort: 1782483001	[SOS] End-stage renal disease	(cohort start + 1) - (cohort start + 365)	0.929	0.929	0.929	1	1	1	View Diagnostics	View Results	View Report
2	logistic	Cohort: 1782481001	[SOS] End-stage renal disease	(cohort start + 1) - (cohort start + 365)	0.887	0.887	0.887	1	1	1	View Diagnostics	View Results	View Report
3	logistic	Cohort: 1782482001	[SOS] End-stage renal disease	(cohort start + 1) - (cohort start + 365)	0.929	0.929	0.929	1	1	1	View Diagnostics	View Results	View Report

Patient-level prediction: PROBAST criteria

Diagnostic

diagnosticId	databaseName	targetName	outcomeName	1.1	1.2	2.1	2.2	2.3	3.4	3.6	4.1
1	IBM CCAE	Cohort: 1782483001	[SOS] End-stage renal disease	✓ Pass	✓ Pass	✓ Pass	✓ Pass	? Unkown	✓ Pass	✓ Pass	✗ Fail

- 1.1 Appropriate data sources
- 1.2 Appropriate inclusions/exclusions
- 2.1 Predictors defined similarly for all
- 2.2 Predictor assessed without outcome knowledge
- 2.3 Predictor available when model is to be used
- 3.4 Outcomes defined similarly for all
- 3.6 Time interval from predictor to outcome is okay
- 4.1 Are there enough outcomes (200)

PLP model results are diagnostics: discrimination

OHDSI Analysis

- About
- CohortGenerator
- CohortDiagnostics
- Characterization
- Prediction
- Estimation
- SCCS

Prediction Viewer

Back To Models Summary

Full Result Explorer

modelDesignId : 1
developmentDb : IBM CCAE
modelType : logistic
validationDb : IBM CCAE
Target : Cohort: 1782483001
outcome : [SOS] End-stage renal disease

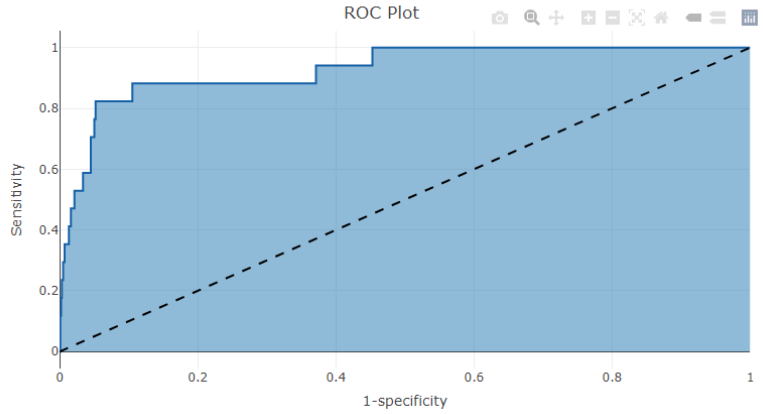
Design Settings | Model | Threshold Dependant | **Discrimination** | Calibration | Net Benefit | Validation

Summary

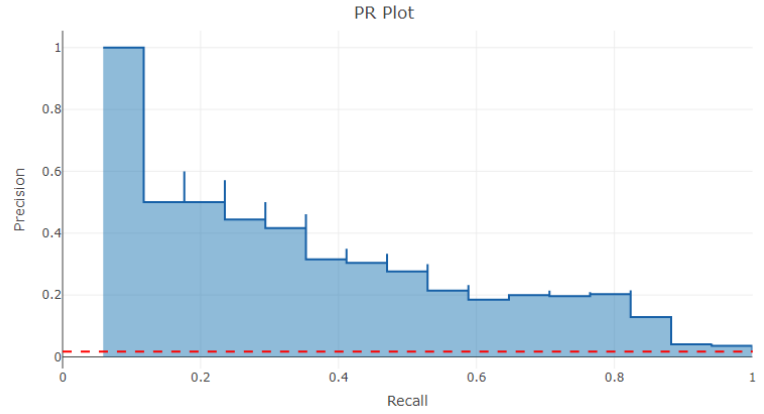
Click view to see the corresponding plots:

	performanceId	evaluation	AUROC	95% lower AUROC	95% upper AUROC	AUPRC
View	1	Test	0.9287261	0.8663226	0.9911296	0.3638089
View	1	Train	0.9549709	0.9337148	0.9762270	0.4956892
View						6320

ROC Plot



Precision recall plot



PLP model results are diagnostics: calibration

OHDSI Analysis

- About
- CohortGenerator
- CohortDiagnostics
- Characterization
- Prediction
- Estimation
- SCCS

Prediction Viewer

Back To Models Summary

Full Result Explorer

modelDesignId : 1
developmentDb : IBM CCAE
modelType : logistic
validationDb : IBM CCAE
Target : Cohort: 1782483001
outcome : [SOS] End-stage renal disease

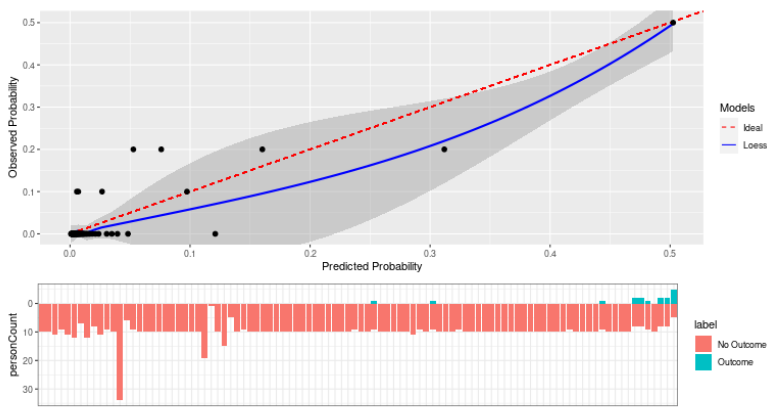
Design Settings | Model | Threshold Dependant | Discrimination | **Calibration** | Net Benefit | Validation

Summary

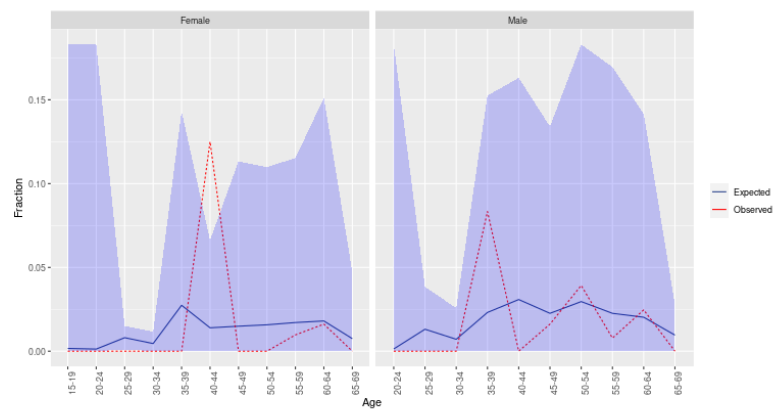
Click view to see corresponding plots:

	performanceId	evaluation	E90	Emax	calibrationInLarge mean prediction	calibrationInLarge observed risk	calibrationInLarge intercept	weak calibration intercept	weak calibration gradient
View	1	Test	0.0126729	0.0732649	0.0201014	0.0170854	-0.2045637	-0.0010278	1.0852866
View	1	Train							
View	1	CV							

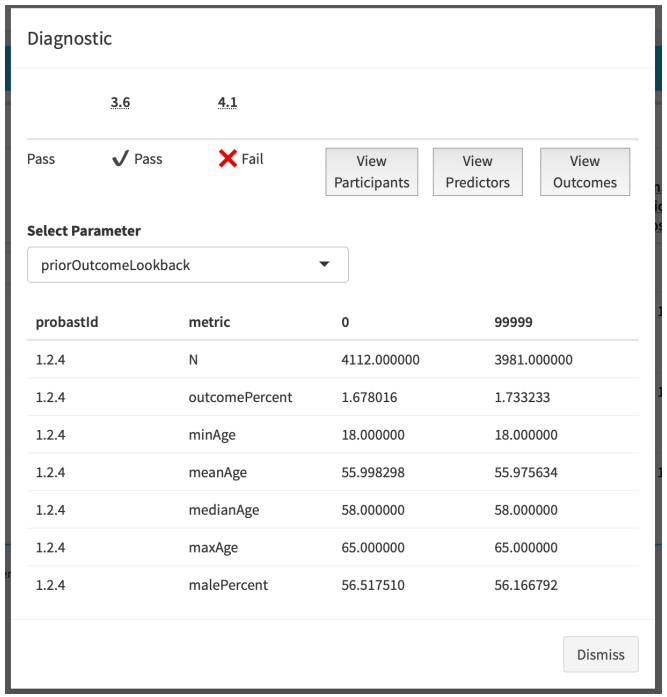
Calibration Plot

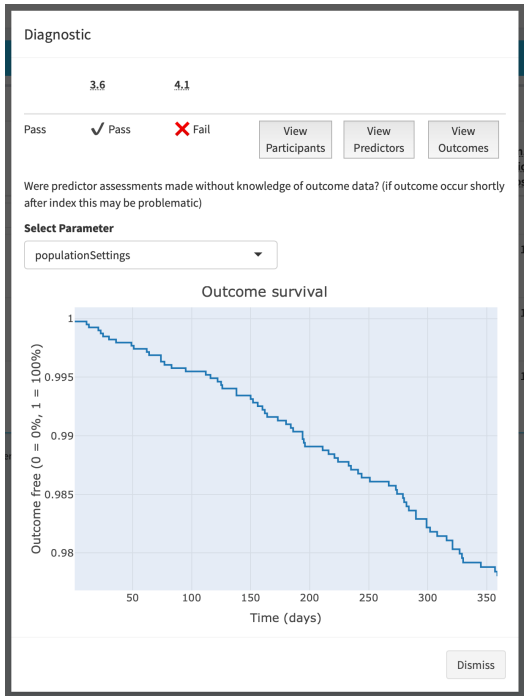


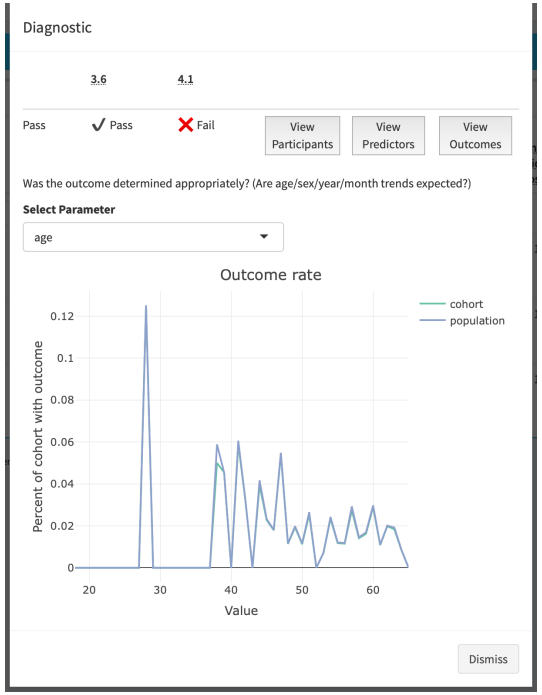
Demographic Plot

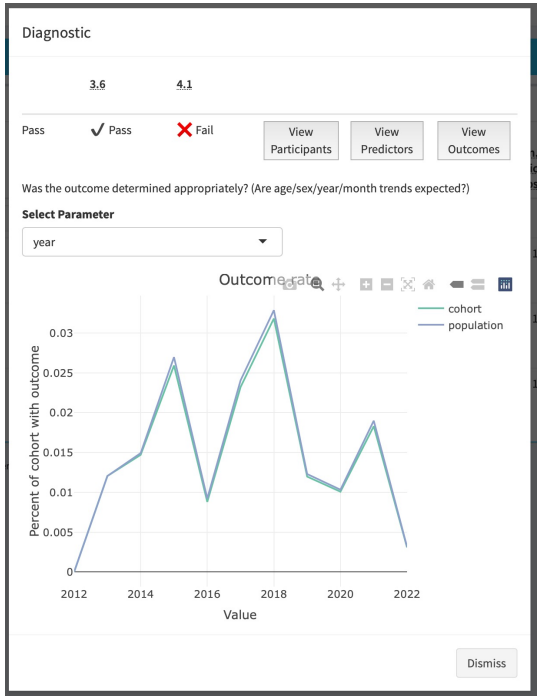


PLP model diagnostics: pulldowns

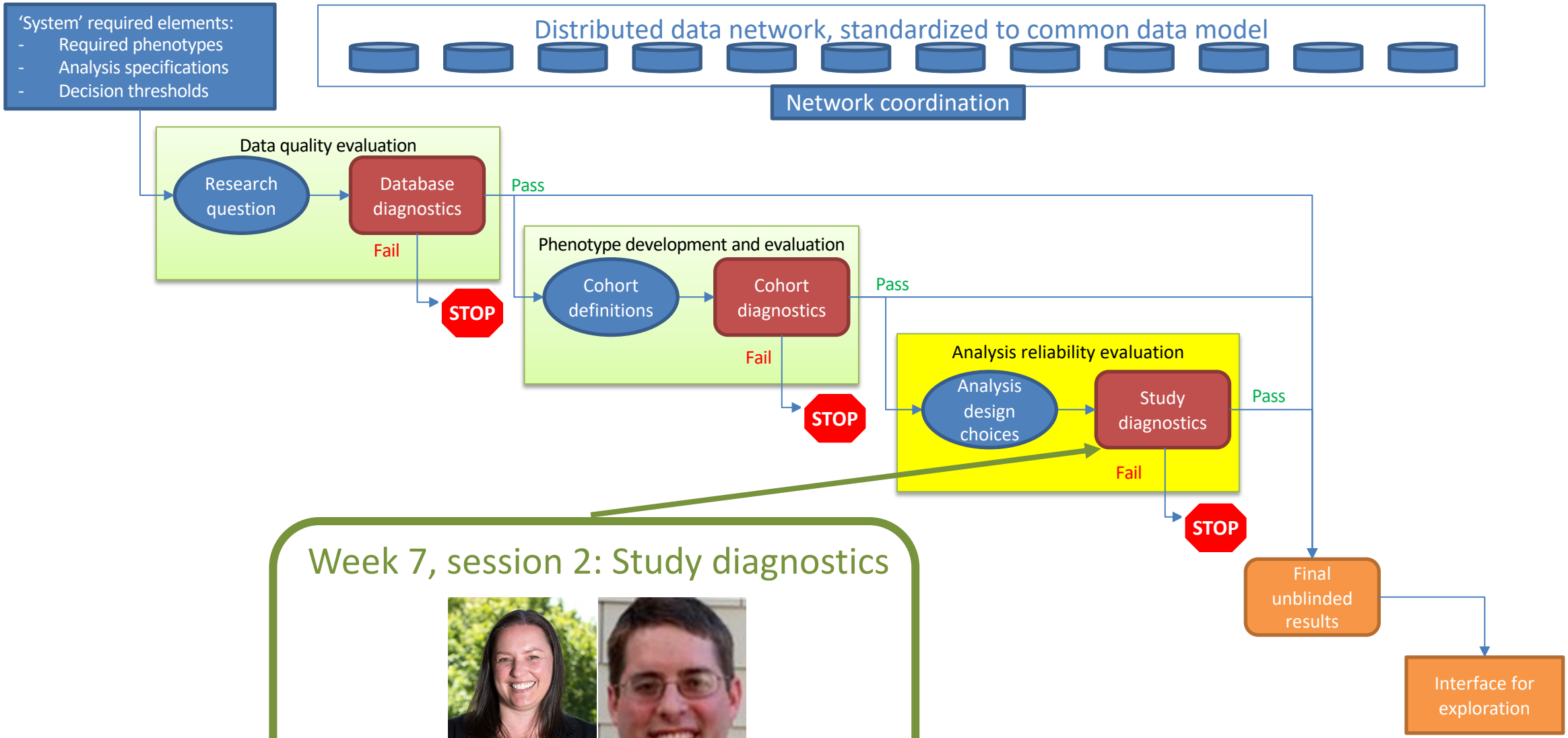






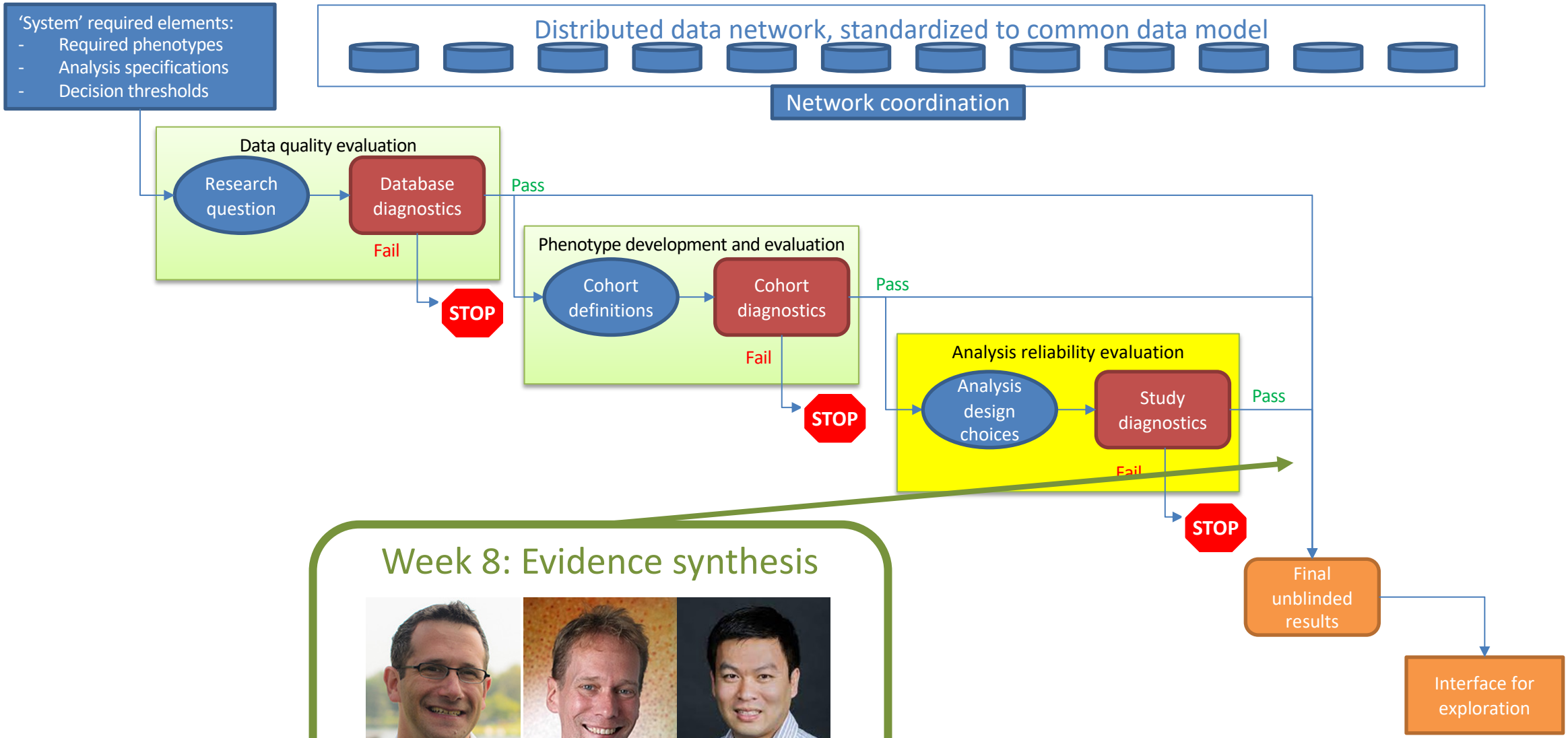


Session 2: fluoroquinolone and aortic aneurysms



Week 7, session 2: Study diagnostics

Next week: synthesizing evidence across databases



Week 8: Evidence synthesis