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

Data quality evaluation
- Research question
- Database diagnostics
- Phenotype development and evaluation
  - Cohort definitions
  - Cohort diagnostics

Analysis reliability evaluation
- Analysis design choices
- Study diagnostics

Distributed data network, standardized to common data model

Network coordination

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

Today: evaluating analysis reliability through study diagnostics

Final unblinded results

Interface for exploration

Week 6-8: Analysis execution, study diagnostics, evidence synthesis
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
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)
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
Statistical power: MDRR in the Anti-VEGF study

• Results ShinyApp: [https://data.ohdsi.org/AntiVegfKidneyFailure/#use](https://data.ohdsi.org/AntiVegfKidneyFailure/#use)

<table>
<thead>
<tr>
<th>Database</th>
<th>Target</th>
<th>Comparator</th>
<th>Outcome</th>
<th>Max SDM</th>
<th>Shared Max SDM</th>
<th>Equipoise</th>
<th>MDRR</th>
<th>EASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCAE</td>
<td>aflibercept</td>
<td>ranibizumab</td>
<td>ESRD</td>
<td>0.065</td>
<td>0.135</td>
<td>0.607</td>
<td>2.05</td>
<td>0.054</td>
</tr>
<tr>
<td>CCAE</td>
<td>ranibizumab</td>
<td>bevacizumab</td>
<td>ESRD</td>
<td>0.051</td>
<td>0.097</td>
<td>0.834</td>
<td>1.89</td>
<td>0.054</td>
</tr>
<tr>
<td>CCAE</td>
<td>aflibercept</td>
<td>bevacizumab</td>
<td>ESRD</td>
<td>0.055</td>
<td>0.113</td>
<td>0.822</td>
<td>1.82</td>
<td>0.067</td>
</tr>
</tbody>
</table>

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

>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](https://data.ohdsi.org/AntiVegfKidneyFailure/#use)

<table>
<thead>
<tr>
<th>Database</th>
<th>Target</th>
<th>Comparator</th>
<th>Outcome</th>
<th>Max SDM</th>
<th>Shared Max SDM</th>
<th>Equipoise</th>
<th>MDRR</th>
<th>EASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCAE</td>
<td>aflibercept</td>
<td>ranibizumab</td>
<td>ESRD</td>
<td>0.065</td>
<td>0.135</td>
<td>0.607</td>
<td>2.05</td>
<td>0.054</td>
</tr>
<tr>
<td>CCAE</td>
<td>ranibizumab</td>
<td>bevacizumab</td>
<td>ESRD</td>
<td>0.051</td>
<td>0.097</td>
<td>0.834</td>
<td>1.89</td>
<td>0.054</td>
</tr>
<tr>
<td>CCAE</td>
<td>aflibercept</td>
<td>bevacizumab</td>
<td>ESRD</td>
<td>0.055</td>
<td>0.113</td>
<td>0.822</td>
<td>1.82</td>
<td>0.067</td>
</tr>
</tbody>
</table>

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
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
Covariate balance: SMD in the Anti-VEGF study

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 standardized 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: 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

Max abs. SDM = 0.04

>99% of target population remains in analysis cohort

‘Target’ cohort
‘Analysis’ cohort
Generalizability:
Attrition fraction & standardized mean difference

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

Max abs. SDM = 0.40

Only 13% of target population remains in analysis cohort

‘Target’ cohort
‘Analysis’ cohort
Generalizability:
In the Anti-VEGF study

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

>74% of target population remains in analysis cohort

Good!
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)
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)
Residual bias: EASE in the Anti-VEGF study

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

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

<table>
<thead>
<tr>
<th>Diagnostics metric</th>
<th>Literature-derived</th>
<th>Strategus interface</th>
<th>Data-driven (LEGEND-HTN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical power (MDRR)</td>
<td>&lt; 10</td>
<td>&lt; Inf; &lt; 10 (SCCS)</td>
<td></td>
</tr>
<tr>
<td>Equipoise</td>
<td>&gt; 0.50</td>
<td>&gt; 0.20</td>
<td>&gt; 0.50</td>
</tr>
<tr>
<td>Covariate balance (SDM)</td>
<td>&lt; 0.10</td>
<td>&lt; 0.10</td>
<td>&lt; 0.50</td>
</tr>
<tr>
<td>Generalizability (attrition)</td>
<td>-</td>
<td>&lt;= 1</td>
<td></td>
</tr>
<tr>
<td>Systematic error (EASE)</td>
<td>&lt; 0.25</td>
<td>&lt; 0.25</td>
<td></td>
</tr>
</tbody>
</table>

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

OHDSI Analysis Viewer

Table of contents
1. Introduction
2. How to use the viewer
3. Analysis types
   1. Characterization
   2. Population-level effect estimation
   3. Patient-level prediction

Introduction
This is an interactive shiny app for exploring standardized outputs for OHDSI analyses including:
- characterization (descriptive studies)
- population-level effect estimation (causal inference)
- patient-level prediction (inference)

Full details of all the analysis tools can be found on the HARES website.

How to use the viewer
Please use the left hand menu to select the type of analysis to explore (click on a button). This show the results that can be interactively explored.

Analysis types
Characterization
The OHDSI community have developed a suite of tools for conducting characterization studies including:
- incidence rate calculation
- baseline characterization
- treatment pathways
- and more

Population-level effect estimation
The OHDSI community have developed several packages that enable users with data in the OMOP common data model to perform causal inference studies.
- CohortMethod
- SelfControlledCaseSeries
- SelfControlledCohort

Patient-level prediction
The OHDSI community have developed several packages that enable users with data in the OMOP common data model to develop and validate patient-level prediction models.
- PatientLevelPrediction
- EnsemblePatientLevelPrediction
- DeepPatientLevelPrediction
## Cohort Diagnostics

### Cohort Level Diagnostics

#### Select Report
- Cohort Counts

#### Database(s)
- IBM Health MarketScan® Commercial Claims and Encounters Database

#### Cohorts

### Cohort Counts

<table>
<thead>
<tr>
<th>Cohort ID</th>
<th>Cohort Name</th>
<th>Persons</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1782164</td>
<td>[SOS] End-stage renal disease</td>
<td>240,258</td>
<td>240,258</td>
</tr>
<tr>
<td>C1782480</td>
<td>[SOS Phenotype Devt] persons with diseases</td>
<td>1,294,100</td>
<td>1,294,100</td>
</tr>
<tr>
<td>C1782481</td>
<td>[SOS Phenotype Devt] rituximab exposures after new use with 3 exposures in 21-70d windows</td>
<td>4,486</td>
<td>6,022</td>
</tr>
<tr>
<td>C1782482</td>
<td>[SOS Phenotype Devt] bevacizumab exposures after new use with 3 exposures in 21-70d windows</td>
<td>15,440</td>
<td>19,874</td>
</tr>
<tr>
<td>C1782483</td>
<td>[SOS Phenotype Devt] afibrocet exposures after new use with 3 exposures in 21-70d windows</td>
<td>4,113</td>
<td>5,192</td>
</tr>
</tbody>
</table>
Characterization: Time-to-event
## Cohort Method

### Cohort Method Evidence Explorer

#### Diagnostics

<table>
<thead>
<tr>
<th>databaseName</th>
<th>analysisDesc</th>
<th>target</th>
<th>comparator</th>
<th>outcome</th>
<th>maxSdm</th>
<th>sharedMaxSdm</th>
<th>equipoise</th>
<th>mdr</th>
<th>attritionFraction</th>
<th>ease</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM CCAE</td>
<td>Cohort method, On treatment</td>
<td>[SOS Phenotype Dev] aflibercept exposures after new use with 3 exposures in 21-70d windows - in cohorts</td>
<td>[SOS Phenotype Dev] ranibizumab exposures after new use with 3 exposures in 21-70d windows - in cohorts</td>
<td>[SOS] End-stage renal disease</td>
<td>0.0549741170 305916</td>
<td>0.134668062 55618</td>
<td>0.6072000986 31488</td>
<td>2.0491241549 8637</td>
<td>0.5364563503 64964</td>
<td>0.0544376653 240157</td>
</tr>
</tbody>
</table>
Estimation: Cohort method diagnostics
pass/fail based on a priori decision thresholds

<table>
<thead>
<tr>
<th>target</th>
<th>comparator</th>
<th>outcome</th>
<th>maxSdm</th>
<th>sharedMaxSdm</th>
<th>equiPOISE</th>
<th>mdr</th>
<th>attritionFraction</th>
<th>ease</th>
<th>balanceDiagnostic</th>
<th>sharedBalancedDiagnostic</th>
<th>equipoiseDiagnostic</th>
<th>mdrDiagnostic</th>
<th>attritionDiagnostic</th>
<th>easeDiagnostic</th>
<th>unblind</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOS Phenotype Dev</td>
<td>[SOS Phenotype Dev]</td>
<td>[SOS] End-stage renal disease</td>
<td>0.0549741170</td>
<td>305916</td>
<td>0.116452246</td>
<td>59818</td>
<td>0.807200005966</td>
<td>31488</td>
<td>2.049312549</td>
<td>9837</td>
<td>0.536449342503</td>
<td>84584</td>
<td>0.0544378853</td>
<td>240157</td>
<td>PASS</td>
</tr>
</tbody>
</table>

\[
<0.1 \quad NA \quad >0.2 \quad NA \quad <1 \quad <0.25
\]
Estimation: Cohort method diagnostic drilldown: Minimum Detectable Relative Risk (MDRR)
Estimation: Cohort method diagnostic drilldown: Attrition Fraction
Estimation: Cohort method diagnostic drilldown: Equipoise

**Cohort Method**

**Cohort Method Evidence Explorer**

**Target**

[SOS Phenotype Dev] afibriccept exposures after new use with 3 exposures in 21-70d windows - in cohorts: (3782460) starts within D: 99999 - D: 99999 of cohort start and ends D: 99999 - D: 99999 of cohort start, first ever occurrence with at least 365 days prior observation and 1 days follow up observation, males, females aged 18+

**Comparator**

[SOS Phenotype Dev] rituximab exposures after new use with 3 exposures in 21-70d windows - in cohorts: (3782460) starts within D: 99999 - D: 99999 of cohort start and ends D: 99999 - D: 99999 of cohort start, first ever occurrence 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 CCACE

**Analysis**

Cohort method, On treatment

---

**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.
Estimation: Cohort method diagnostic drilldown: Covariate balance maxSDM
Estimation: Cohort method diagnostic drilldown: Expected Average Systematic Error (EASE)
Estimation: Self-controlled case series (SCCS) pass/fail based on a priori decision thresholds

Threshold
- <10
- >0.05
- >0.05
- <0.25
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.
Estimation: SCCS diagnostic drilldown: Time to event

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

### Model Designs Summary

<table>
<thead>
<tr>
<th>Design ID</th>
<th>Model Type</th>
<th>Target Pop</th>
<th>Outcome</th>
<th>TAR</th>
<th>min AUROC</th>
<th>mean AUROC</th>
<th>max AUROC</th>
<th>Num. Diagnostic Dbs</th>
<th>Num. Development Dbs</th>
<th>Num. Validation Dbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>logistic</td>
<td>Cohort: 1782483001</td>
<td>[SOS] End-stage renal disease</td>
<td>(cohort start + 1) - (cohort start + 365)</td>
<td>0.929</td>
<td>0.929</td>
<td>0.929</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>logistic</td>
<td>Cohort: 1782481001</td>
<td>[SOS] End-stage renal disease</td>
<td>(cohort start + 1) - (cohort start + 365)</td>
<td>0.887</td>
<td>0.887</td>
<td>0.887</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>logistic</td>
<td>Cohort: 1782482001</td>
<td>[SOS] End-stage renal disease</td>
<td>(cohort start + 1) - (cohort start + 365)</td>
<td>0.929</td>
<td>0.929</td>
<td>0.929</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Patient-level prediction: PROBAST criteria

- **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
PLP model results are diagnostics: calibration
PLP model diagnostics: pulldowns
Session 2: fluoroquinolone and aortic aneurysms

Week 7, session 2: Study diagnostics
Next week: synthesizing evidence across databases

‘System’ required elements:
- Required phenotypes
- Analysis specifications
- Decision thresholds

Distributed data network, standardized to common data model

Network coordination

Week 8: Evidence synthesis

Data quality evaluation
- Research question
- Database diagnostics
  - Pass
  - Fail
  - STOP

Phenotype development and evaluation
- Cohort definitions
- Cohort diagnostics
  - Pass
  - Fail
  - STOP

Analysis reliability evaluation
- Analysis design choices
- Study diagnostics
  - Pass
  - Fail
  - STOP

Final unblinded results

Interface for exploration

Week 8: Evidence synthesis