Running multiple analyses at once using the SelfControlledCaseSeries package

Martijn J. Schuemie

2022-11-25

Contents

1 Introduction

In this vignette we focus on running several different analyses on several exposure-outcome pairs This can be useful when we want to explore the sensitivity to analyses choices, include controls, or run an experiment similar to the OMOP experiment to empirically identify the optimal analysis choices for a particular research question.

This vignette assumes you are already familiar with the `SelfControlledCaseSeries` package and are able to perform single studies. We will walk through all the steps needed to perform an exemplar set of analyses, and we have selected the well-studied topic of the effect of nonsteroidal anti-inflammatory drugs (NSAIDs) on gastrointestinal (GI) bleeding-related hospitalization. For simplicity, we focus on one NSAID: diclofenac. We will execute various variations of an analysis for the primary exposure pair and a large set of negative control exposures.

2 General approach

The general approach to running a set of analyses is that you specify all the function arguments of the functions you would normally call, and create sets of these function arguments. The final outcome models as well as intermediate data objects will all be saved to disk for later extraction.

An analysis will be executed by calling these functions in sequence:

1. `getDbSccsData()`
2. `createStudyPopulation()`
3. `createSccsIntervalData()`
4. `fitSccsModel()`

When you provide several analyses to the `SelfControlledCaseSeries` package, it will determine whether any of the analyses and exposure-outcome pairs have anything in common, and will take advantage of this fact. For example, if we specify several exposure-outcome pairs with the same outcome, the data for the outcome will be extracted only once.

The function arguments you need to define have been divided into four groups:

1. **Hypothesis of interest**: arguments that are specific to a hypothesis of interest, in the case of the self-controlled case series this is a combination of exposure and outcome.
2. **Analyses**: arguments that are not directly specific to a hypothesis of interest, such as the washout window, whether to adjust for age and seasonality, etc.
3. Arguments that are the output of a previous function in the `SelfControlledCaseSeries` package, such as the `SccsIntervalData` argument of the `createSccsIntervalData` function. These cannot be specified by the user.
4. Arguments that are specific to an environment, such as the connection details for connecting to the server, and the name of the schema holding the CDM data.

3 Preparation for the example

We need to tell R how to connect to the server where the data are. `SelfControlledCaseSeries` uses the `DatabaseConnector` package, which provides the `createConnectionDetails` function. Type `?createConnectionDetails` for the specific settings required for the various database management systems (DBMS). For example, one might connect to a PostgreSQL database using this code:

```
connectionDetails <- createConnectionDetails(dbms = "postgresql",
                                             server = "localhost/ohdsi",
                                             user = "joe",
                                             password = "supersecret")

outputFolder <- "s:/temp/sccsVignette2"

cdmDatabaseSchema <- "my_cdm_data"
cohortDatabaseSchema <- "my_cohorts"
options(sqlRenderTempEmulationSchema = NULL)
cdmVersion <- "5"
```

The last three lines define the `cdmDatabaseSchema` and `cohortDatabaseSchema` variables, as well as the CDM version. We'll use these later to tell R where the data in CDM format live, where we want to store the (outcome) cohorts, and what version CDM is used. Note that for Microsoft SQL Server, databaseschemas need to specify both the database and the schema, so for example `cdmDatabaseSchema <- "my_cdm_data.dbo"`.

We also need to prepare our exposures and outcomes of interest. The drug_era table in the OMOP Common Data Model already contains pre-specified cohorts of users at the ingredient level, so we will use that for the exposures. For the outcomes, we want to restrict our analysis only to those events that are recorded in an inpatient setting, so we will need to create a custom cohort table. For this example, we are only interested in GI bleed (concept ID 192671) .

We create a text file called *vignette.sql* with the following content:

```
/**********************************
File vignette.sql
**********************************/

IF OBJECT_ID('@cohortDatabaseSchema.@outcomeTable', 'U') IS NOT NULL
DROP TABLE @cohortDatabaseSchema.@outcomeTable;

SELECT 1 AS cohort_definition_id,
condition_start_date AS cohort_start_date,
condition_end_date AS cohort_end_date,
condition_occurrence.person_id AS subject_id
INTO @cohortDatabaseSchema.@outcomeTable
FROM @cdmDatabaseSchema.condition_occurrence
INNER JOIN @cdmDatabaseSchema.visit_occurrence
ON condition_occurrence.visit_occurrence_id = visit_occurrence.visit_occurrence_id
WHERE condition_concept_id IN (
SELECT descendant_concept_id
FROM @cdmDatabaseSchema.concept_ancestor
WHERE ancestor_concept_id = 192671 -- GI - Gastrointestinal haemorrhage
)
AND visit_occurrence.visit_concept_id IN (9201, 9203);
```

This is parameterized SQL which can be used by the `SqlRender` package. We use parameterized SQL so we do not have to pre-specify the names of the CDM and result schemas. That way, if we want to run the SQL on a different schema, we only need to change the parameter values; we do not have to change the SQL code. By also making use of translation functionality in `SqlRender`, we can make sure the SQL code can be run in many different environments.

```
library(SqlRender)
sql <- readSql("vignette.sql")
sql <- render(sql,
              cdmDatabaseSchema = cdmDatabaseSchema,
              cohortDatabaseSchema = cohortDatabaseSchema)
sql <- translate(sql, targetDialect = connectionDetails$dbms)

connection <- connect(connectionDetails)
executeSql(connection, sql)
```

In this code, we first read the SQL from the file into memory. In the next line, we replace the two parameter names with the actual values. We then translate the SQL into the dialect appropriate for the DBMS we already specified in the `connectionDetails`. Next, we connect to the server, and submit the rendered and translated SQL.

4 Specifying hypotheses of interest

The first group of arguments define the exposure and outcome. Here we demonstrate how to create a list of exposure-outcome pairs:

```
diclofenac <- 1124300
negativeControls <- c(705178, 705944, 710650, 714785, 719174, 719311, 735340, 742185,
                      780369, 781182, 924724, 990760, 1110942, 1111706, 1136601,
                      1317967, 1501309, 1505346, 1551673, 1560278, 1584910, 19010309,
                      40163731)

exposuresOutcomeList <- list()
exposuresOutcomeList[[1]] <- createExposuresOutcome(
  outcomeId = 1,
  exposures = list(createExposure(exposureId = diclofenac))
)
for (exposureId in c(negativeControls)) {
  exposuresOutcome <- createExposuresOutcome(
    outcomeId = 1,
    exposures = list(createExposure(exposureId = exposureId, trueEffectSize = 1))
  )
  exposuresOutcomeList[[length(exposuresOutcomeList) + 1]] <- exposuresOutcome
}
```

We defined the outcome of interest to be the cohort with ID 1 we defined in the SQL above. The exposures include diclofenac (concept ID 1124300) and a large number of negative control exposures.

A convenient way to save `exposuresOutcomeList` to file is by using the `saveExposuresOutcomeList` function, and we can load it again using the `loadExposuresOutcomeList` function.

5 Specifying analyses

The second group of arguments are not specific to a hypothesis of interest, and comprise the majority of arguments. For each function that will be called during the execution of the analyses, a companion function is available that has (almost) the same arguments. For example, for the `fitSccsModel()` function there is the `createFitSccsModelArgs()` function. These companion functions can be used to create the arguments to be used during execution:

```
getDbSccsDataArgs <- createGetDbSccsDataArgs(
  useCustomCovariates = FALSE,
  deleteCovariatesSmallCount = 100,
  exposureIds = c(),
  maxCasesPerOutcome = 100000
)

createStudyPopulationArgs <- createCreateStudyPopulationArgs(
  naivePeriod = 180,
  firstOutcomeOnly = FALSE
)

covarExposureOfInt <- createEraCovariateSettings(
  label = "Exposure of interest",
```

```
  includeEraIds = "exposureId",
  start = 1,
  end = 0,
  endAnchor = "era end",
  profileLikelihood = TRUE,
  exposureOfInterest = TRUE
)

createSccsIntervalDataArgs1 <- createCreateSccsIntervalDataArgs(
  eraCovariateSettings = covarExposureOfInt
)

fitSccsModelArgs <- createFitSccsModelArgs()
```

Any argument that is not explicitly specified by the user will assume the default value specified in the function. Note that for several arguments for concept or cohort definition IDs we can use the `exposureIdRef` (default = `"exposureId"`) in the `Exposure` objects that we used in `createExposuresOutcome()`. In this case, we defined the argument `includeEraIds` to get the value of the `exposureId` variable, meaning it will take the value of the diclofenac concept ID, or any of the negative control IDs. Also note that we set `exposureOfInterest = TRUE`, which will cause the estimate for this covariate to be included in the result summary later on.

We can now combine the arguments for the various functions into a single analysis:

```
sccsAnalysis1 <- createSccsAnalysis(
  analysisId = 1,
  description = "Simplest model",
  getDbSccsDataArgs = getDbSccsDataArgs,
  createStudyPopulationArgs = createStudyPopulationArgs,
  createIntervalDataArgs = createSccsIntervalDataArgs1,
  fitSccsModelArgs = fitSccsModelArgs
)
```

Note that we have assigned an analysis ID (1) to this set of arguments. We can use this later to link the results back to this specific set of choices. We also include a short description of the analysis.

We can easily create more analyses, for example by including adjustments for age and seasonality, or for including other drugs in the model. For example, here we create a covariate representing drugs that are used in preventing the outcome of interest, in this case proton pump inhibitors (PPIs):

```
ppis <- c(911735, 929887, 923645, 904453, 948078, 19039926)

covarProphylactics <- createEraCovariateSettings(
  label = "Prophylactics",
  includeEraIds = ppis,
  start = 1,
  end = 0,
  endAnchor = "era end"
)

createSccsIntervalDataArgs2 <- createCreateSccsIntervalDataArgs(
  eraCovariateSettings = list(
    covarExposureOfInt,
    covarProphylactics
```

```
  )
)

sccsAnalysis2 <- createSccsAnalysis(
  analysisId = 2,
  description = "Including prophylactics",
  getDbSccsDataArgs = getDbSccsDataArgs,
  createStudyPopulationArgs = createStudyPopulationArgs,
  createIntervalDataArgs = createSccsIntervalDataArgs2,
  fitSccsModelArgs = fitSccsModelArgs
)

ageSettings <- createAgeCovariateSettings(ageKnots = 5)

seasonalitySettings <- createSeasonalityCovariateSettings(seasonKnots = 5)

covarPreExp <- createEraCovariateSettings(
  label = "Pre-exposure",
  includeEraIds = "exposureId",
  start = -30,
  end = -1,
  endAnchor = "era start"
)

createSccsIntervalDataArgs3 <- createCreateSccsIntervalDataArgs(
  eraCovariateSettings = list(
    covarExposureOfInt,
    covarPreExp,
    covarProphylactics
  ),
  ageCovariateSettings = ageSettings,
  seasonalityCovariateSettings = seasonalitySettings,
  eventDependentObservation = TRUE
)

sccsAnalysis3 <- createSccsAnalysis(
  analysisId = 3,
  description = "Including prophylactics, age, season, pre-exposure, and censoring",
  getDbSccsDataArgs = getDbSccsDataArgs,
  createStudyPopulationArgs = createStudyPopulationArgs,
  createIntervalDataArgs = createSccsIntervalDataArgs3,
  fitSccsModelArgs = fitSccsModelArgs
)

covarAllDrugs <- createEraCovariateSettings(
  label = "Other exposures",
  excludeEraIds = "exposureId",
  stratifyById = TRUE,
  start = 1,
  end = 0,
  endAnchor = "era end",
  allowRegularization = TRUE
)
```

```
createSccsIntervalDataArgs4 <- createCreateSccsIntervalDataArgs(
  eraCovariateSettings = list(
    covarExposureOfInt,
    covarPreExp,
    covarAllDrugs
  ),
  ageCovariateSettings = ageSettings,
  seasonalityCovariateSettings = seasonalitySettings,
  eventDependentObservation = TRUE
)

sccsAnalysis4 <- createSccsAnalysis(
  analysisId = 4,
  description = "Including all other drugs",
  getDbSccsDataArgs = getDbSccsDataArgs,
  createStudyPopulationArgs = createStudyPopulationArgs,
  createIntervalDataArgs = createSccsIntervalDataArgs4,
  fitSccsModelArgs = fitSccsModelArgs
)
```

These analyses can be combined in a list:

```
sccsAnalysisList <- list(sccsAnalysis1, sccsAnalysis2, sccsAnalysis3, sccsAnalysis4)
```

A convenient way to save `sccsAnalysisList` to file is by using the `saveSccsAnalysisList` function, and we can load it again using the `loadSccsAnalysisList` function.

6 Executing multiple analyses

We can now run the analyses against the hypotheses of interest using the `runScsAnalyses()`function. This function will run all specified analyses against all hypotheses of interest, meaning that the total number of outcome models is `length(sccsAnalysisList) * length(exposuresOutcomeList)`. (If we want, we can skip some of these combinations using the `analysesToExclude` argument.)

```
multiThreadingSettings <- createDefaultSccsMultiThreadingSettings(parallel::detectCores() - 1)

referenceTable <- runSccsAnalyses(
  connectionDetails = connectionDetails,
  cdmDatabaseSchema = cdmDatabaseSchema,
  exposureDatabaseSchema = cdmDatabaseSchema,
  exposureTable = "drug_era",
  outcomeDatabaseSchema = cohortDatabaseSchema,
  outcomeTable = outcomeTable,
  cdmVersion = cdmVersion,
  outputFolder = outputFolder,
  combineDataFetchAcrossOutcomes = TRUE,
  exposuresOutcomeList = exposuresOutcomeList,
  sccsAnalysisList = sccsAnalysisList,
  sccsMultiThreadingSettings = multiThreadingSettings
)
```

In the code above, we first specify how many parallel threads `SelfControlledCaseSeires` can use. Many of the computations can be computed in parallel, and providing more than one CPU core can greatly speed up the computation. Here we specify `SelfControlledCaseSeires` can use all the CPU cores detected in the system (using the `parallel::detectCores()` function).

We call `runSccsAnalyses`, providing the arguments for connecting to the database, which schemas and tables to use, as well as the analyses and hypotheses of interest. The `outputFolder` specifies where the outcome models and intermediate files will be written.

6.1 Restarting

If for some reason the execution was interrupted, you can restart by re-issuing the `runSccsAnalyses()` command. Any intermediate and final products that have already been completed and written to disk will be skipped.

7 Retrieving the results

The result of the `runSccsAnalyses()` is a data frame with one row per exposure-outcome-analysis combination. It provides the file names of the intermediate and end-result files that were constructed. For example, we can retrieve the fitted model for the combination of our drug of interest, outcome, and first analysis:

```
sccsModelFile <- result$sccsModelFile[result$exposureId == 1124300 &
                                      result$outcomeId == 1 &
                                      result$analysisId == 1]
sccsModel <- readRDS(file.path(outputFolder, sccsModelFile))
sccsModel
```

```
## SccsModel object
##
## Outcome ID: 1
##
## Outcome count:
##   outcomeSubjects outcomeEvents outcomeObsPeriods
## 1           77354        252483             77524
##
## Estimates:
## # A tibble: 1 x 7
##   Name                                ID Estimate LB95CI UB95CI LogRr SeLogRr
##   <chr>                            <dbl>    <dbl>  <dbl>  <dbl> <dbl>   <dbl>
## 1 Exposure of interest: diclofenac  1000     1.23   1.16   1.30 0.205  0.0292
```

Note that some of the file names will appear several times in the table. For example, all analysis share the same sccsData object.

We can always retrieve the file reference table again using the `getFileReference()` function:

```
result <- getFileReference(folder)
```

We can get a summary of the results using `getResultsSummary()`:

```
resultsSum <- getResultsSummary(outputFolder)
head(resultsSum)
```

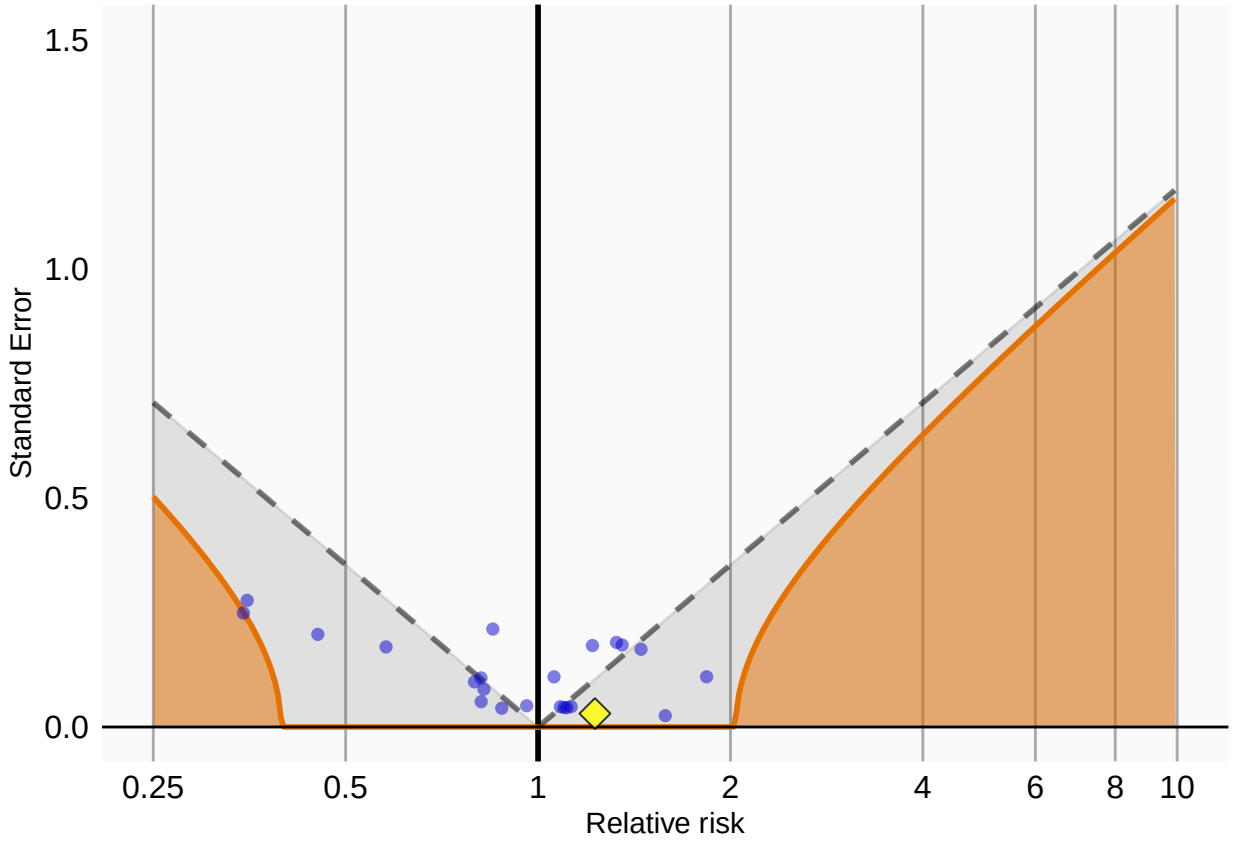
```
## # A tibble: 6 x 29
##   exposuresOu~1 outco~2 analy~3 covar~4 covar~5  eraId trueE~6 outco~7 outco~8 outco~9 covar~* covar~
##           <dbl>   <dbl>   <dbl>   <int> <chr>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl
## 1             1       1       1    1000 Exposu~ 1.12e6      NA   77354  252483   77524    8416  98958
## 2             2       1       1    1000 Exposu~ 7.05e5       1   77354  252483   77524     281  17231
## 3             3       1       1    1000 Exposu~ 7.06e5       1   77354  252483   77524    1802  83938
## 4             4       1       1    1000 Exposu~ 7.11e5       1   77354  252483   77524     142   4870
## 5             5       1       1    1000 Exposu~ 7.15e5       1   77354  252483   77524    2158 110924
## 6             6       1       1    1000 Exposu~ 7.19e5       1   77354  252483   77524    2225 123023
## # ... with 5 more variables: calibratedCi95Ub <dbl>, calibratedP <dbl>, calibratedLogRr <dbl>, calibr
## #   2: outcomeId, 3: analysisId, 4: covariateId, 5: covariateName, 6: trueEffectSize, 7: outcomeSubje
## #   *: covariateEras, *: covariateOutcomes, *: observedDays, *: calibratedRr, *: calibratedCi95Lb
```

This tells us, per exposure-outcome-analysis combination, the estimated relative risk and 95% confidence interval, as well as the number of subjects (cases) and the number of events observed for those subjects. The only covariates included in this summary are those we marked with `exposureOfInterest = TRUE` when calling `createEraCovariateSettings()` earlier.

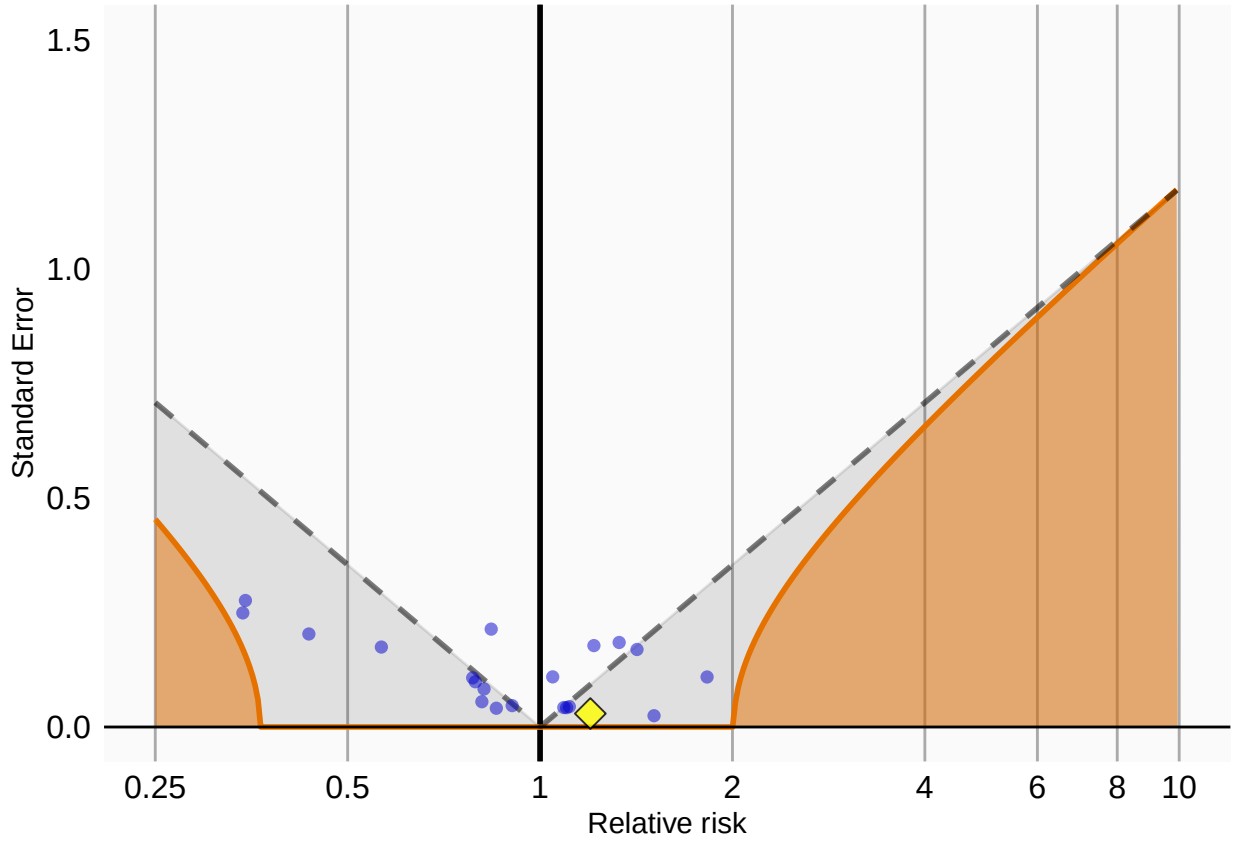
7.1 Empirical calibration

Now that we have produced estimates for all outcomes including our negative controls, we can perform empirical calibration to estimate the bias of the various analyses included in our study. We will create the calibration effect plots for every analysis ID. In each plot, the blue dots represent our negative control exposures, and the yellow diamond represents our exposure of interest: diclofenac. An unbiased, well-calibrated analysis should have 95% of the negative controls between the dashed lines (ie. 95% should have $p > .05$).

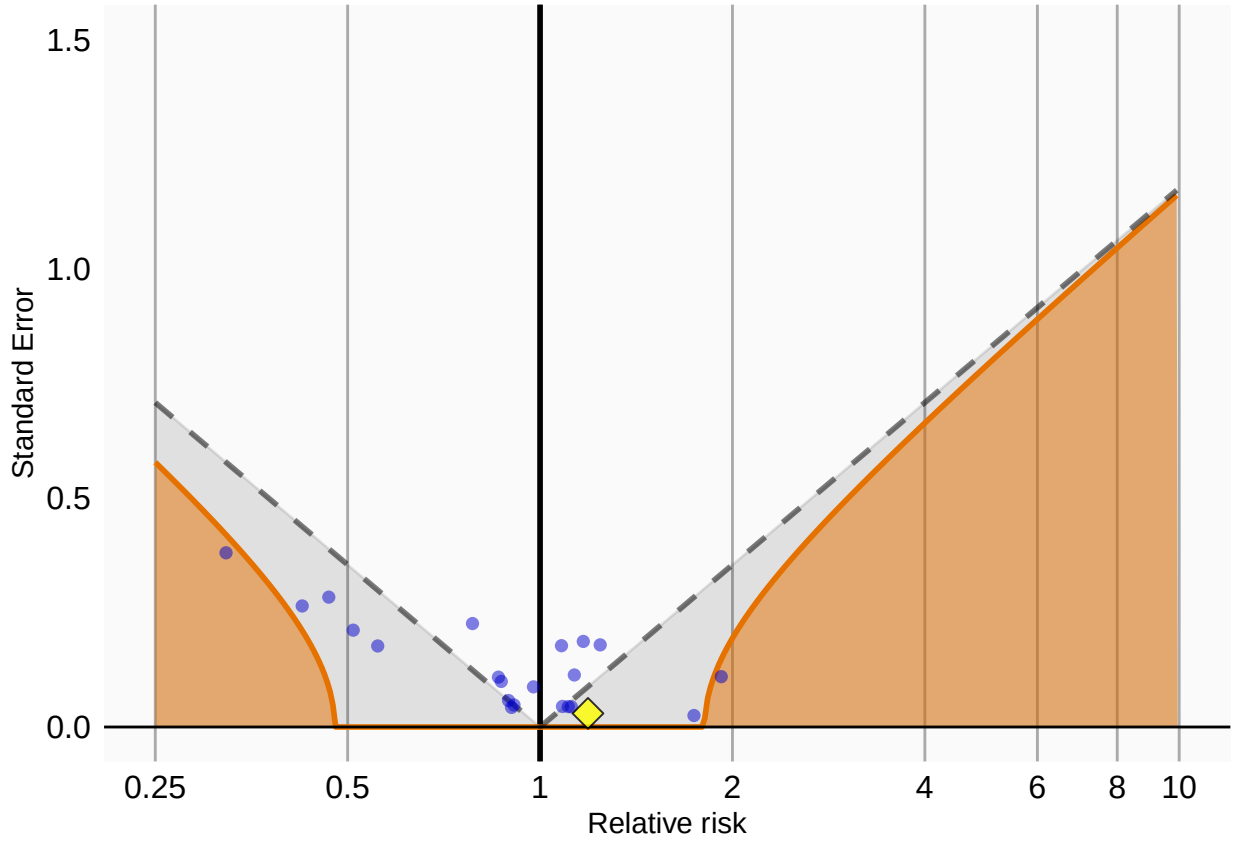
```
install.packages("EmpiricalCalibration")
library(EmpiricalCalibration)

# Analysis 1: Simplest model
negCons <- resultsSum[resultsSum$analysisId == 1 & resultsSum$eraId != 1124300, ]
ei <-  resultsSum[resultsSum$analysisId == 1 & resultsSum$eraId == 1124300, ]
null <- fitNull(negCons$logRr,
                negCons$seLogRr)
plotCalibrationEffect(logRrNegatives = negCons$logRr,
                      seLogRrNegatives = negCons$seLogRr,
                      logRrPositives = ei$logRr,
                      seLogRrPositives = ei$seLogRr,
                      null)
```



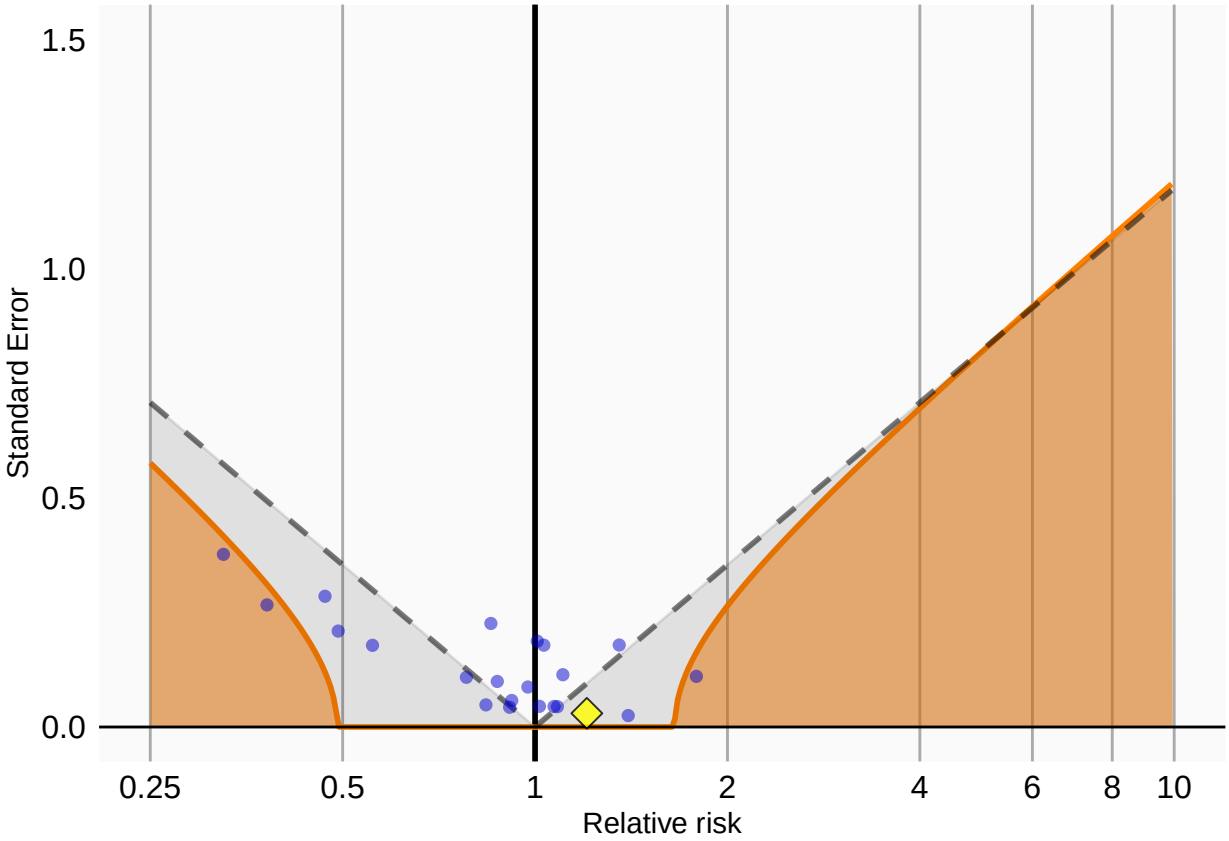

```
# Analysis 2: Including prophylactics
negCons <- resultsSum[resultsSum$analysisId == 2 & resultsSum$eraId != 1124300, ]
ei <-  resultsSum[resultsSum$analysisId == 2 & resultsSum$eraId == 1124300, ]
null <- fitNull(negCons$logRr,
                negCons$seLogRr)
plotCalibrationEffect(logRrNegatives = negCons$logRr,
                      seLogRrNegatives = negCons$seLogRr,
                      logRrPositives = ei$logRr,
                      seLogRrPositives = ei$seLogRr,
                      null)
```




```
# Analysis 3: Including prophylactics, age, season, pre-exposure, and censoring
negCons <- resultsSum[resultsSum$analysisId == 3 & resultsSum$eraId != 1124300, ]
ei <-  resultsSum[resultsSum$analysisId == 3 & resultsSum$eraId == 1124300, ]
null <- fitNull(negCons$logRr,
                negCons$seLogRr)
plotCalibrationEffect(logRrNegatives = negCons$logRr,
                      seLogRrNegatives = negCons$seLogRr,
                      logRrPositives = ei$logRr,
                      seLogRrPositives = ei$seLogRr,
                      null)
```




```
# Analysis 4: Including all other drugs (as well as prophylactics, age, season, pre-
# exposure, and censoring)
negCons <- resultsSum[resultsSum$analysisId == 4 & resultsSum$eraId != 1124300, ]
ei <-  resultsSum[resultsSum$analysisId == 4 & resultsSum$eraId == 1124300, ]
null <- fitNull(negCons$logRr,
                negCons$seLogRr)
plotCalibrationEffect(logRrNegatives = negCons$logRr,
                      seLogRrNegatives = negCons$seLogRr,
                      logRrPositives = ei$logRr,
                      seLogRrPositives = ei$seLogRr,
                      null)
```



8 Acknowledgments

Considerable work has been dedicated to provide the `SelfControlledCaseSeries` package.

```
citation("SelfControlledCaseSeries")
```

```
##
## To cite package 'SelfControlledCaseSeries' in publications use:
##
##   Schuemie M, Ryan P, Shaddox T, Suchard M (2022). _SelfControlledCaseSeries: Self-Controlled Case S
##   <https://github.com/OHDSI/SelfControlledCaseSeries>.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {SelfControlledCaseSeries: Self-Controlled Case Series},
##     author = {Martijn Schuemie and Patrick Ryan and Trevor Shaddox and Marc Suchard},
##     year = {2022},
##     note = {R package version 4.0.0},
##     url = {https://github.com/OHDSI/SelfControlledCaseSeries},
##   }
```

Further, `SelfControlledCaseSeries` makes extensive use of the `Cyclops` package.

```
citation("Cyclops")
```

```
##
## To cite Cyclops in publications use:
##
##   Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D (2013). "Massive parallelization of serial inf
##   Transactions on Modeling and Computer Simulation_, *23*, 10. <https://dl.acm.org/doi/10.1145/24144
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     author = {M. A. Suchard and S. E. Simpson and I. Zorych and P. Ryan and D. Madigan},
##     title = {Massive parallelization of serial inference algorithms for complex generalized linear m
##     journal = {ACM Transactions on Modeling and Computer Simulation},
##     volume = {23},
##     pages = {10},
##     year = {2013},
##     url = {https://dl.acm.org/doi/10.1145/2414416.2414791},
##   }
```

This work is supported in part through the National Science Foundation grant IIS 1251151.