EvaluatingPhenotypeAlgorithms

Joel N. Swerdel

2020-03-25

Contents

1 Introduction

The `Phevaluator` package enables evaluating the performance characteristics of phenotype algorithms (PAs) using data from databases that are translated into the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM).

This vignette describes how to run the PheValuator process from start to end in the `Phevaluator` package.

2 Overview of Process

There are several steps in performing a PA evaluation: 1. Creating the extremely specific (xSpec), extremely sensitive (xSens), and prevalence cohorts 2. Creating the Diagnostic Predictive Model using the PatientLevelPrediction (PLP) package 3. Creating the Evaluation Cohort 4. Creating the Phenotype Algorithms for evaluation 5. Evaluating the PAs 6. Examining the results of the evaluation

Each of these steps is described in detail below. For this vignette we will describe the evaluation of PAs for diabetes mellitus (DM).

2.1 Defining the set of cohorts to evaluate

The first step is to define the set of cohorts for which we wish to determine performance characteristics. We do this by creating a data frame with five columns:

- **atlasId**: The cohort ID in ATLAS.
- **atlasName**: The full name of the cohort.
- **cohortId**: The cohort ID to use in the package. Usually the same as the cohort ID in ATLAS.
- **name**: A short name for the cohort, to use to create file names. do not use special characters.
- **washoutPeriod**: washoutPeriod: The mininum required continuous observation time prior to index date for subjects within the cohort to test.

A convenient way to create such a data frame is to create a CSV file, and load it into R. Here is an example table we assume is stored in `c:/myCohortFile.csv`:

atlasId,atlasName,cohortId,name,washoutPeriod
7142,Type 2 Diabetes 2 X,7142,Type2Diabetes,0
8339,Type 2 Diabetes 3 X,8339,Type2Diabetes,0
7143,Type 2 Diabetes 1 X IP,7143,Type2Diabetes,0
7144,Type 2 Diabetes 1 X IP 1st Position,7144,Type2Diabetes,0

We can read the table using:

```
cohortSetReference <- read.csv("c:/myCohortFile.csv")
```

2.2 Creating the Extremely Specific (xSpec), Extremely Sensitive (xSens), and Prevalence Cohorts

The extremely specific (xSpec), extremely sensitive (xSens), and prevalence cohorts are developed using the ATLAS tool. The xSpec is a cohort where the subjects in the cohort are likely to be positive for the health outcome of interest (HOI) with a very high probability. This may be achieved by requiring that subjects have multiple condition codes for the HOI in their patient record. An example of this for DM is included in the OHDSI ATLAS repository. In this example each subject has an initial condition code for DM. The cohort definition further specifies that each subject also has a second code for DM between 1 and 30 days after the initial DM code and 10 additional DM codes in the rest of the patient record. This very specific algorithm for DM ensures that the subjects in this cohort have a very high probability for having the condition of DM. This PA also specifies that subjects are required to have at least 365 days of observation in their patient record.

An example of an xSens cohort is created by developing a PA that is very sensitive for the HOI. The system uses the xSens cohort to create a set of "noisy" negative subjects, i.e., subjects with a high likelihood of not having the HOI. This group of subjects will be used in the model building process and is described in detail below. An example of an xSens cohort for DM is also in the OHDSI ATLAS repository.

The system uses the prevalence cohort to provide a reasonable approximation of the prevalence of the HOI in the population. This improves the calibration of the predictive model. The system will use the xSens cohort as the default if a prevalence cohort is not specified. This group of subjects will be used in the model building process and is described in detail below. An example of an prevalence cohort for DM is also in the OHDSI ATLAS repository.

2.3 Evaluating phenotype algorithms for chronic health conditions

The function createChronicPhenotypeModel allows the user to perform a complete analysis for determining the performance characteristics for one or more phenotype algorithms (cohort defintions) for chronic health conditions. This function initiates the process for the three major steps in PheValuator, namely:
1) Develop a diagnostic predictive model for the health condition.
2) Select a large, random set of subjects from the dataset and use the model to determine the probability of each of the subjects having the health condition.
3) Determine the performance characteristics for one or more phenotype algorithms to be used in studies.

createChronicPhenotypeModel should have as inputs:

- connectionDetails - connectionDetails created using the function createConnectionDetails in the DatabaseConnector package.
- cdmDatabaseSchema - The name of the database schema that contains the OMOP CDM instance. Requires read permissions to this database. On SQL Server, this should specify both the database and the schema, so for example 'cdm_instance.dbo'.
- databaseId - Short name for the database (default="TestDB")
- cohortDatabaseSchema - The name of the database schema that is the location where the cohort data used to define the at risk cohort is available. If cohortTable = DRUG_ERA, cohortDatabaseSchema is not used by assumed to be cdmSchema. Requires read permissions to this database.
- cohortTable - The tablename that contains the at risk cohort. If cohortTable <> DRUG_ERA, then expectation is cohortTable has format of COHORT table: cohort_concept_id, SUBJECT_ID, COHORT_START_DATE, COHORT_END_DATE.
- workDatabaseSchema - The name of a database schema where the user has write capability. A temporary cohort table will be created here.
- modelOutputFileName - A string designation for the training model file Recommended structure: "Model_(xSpec Name)*(CDM)*(Qualifiers)_(Analysis Date)", e.g.,"Model_10XStroke_MyCDM_Age18-62_20190101" to designate the file was from a **Model**, built on the **10 X Stroke** xSpec, using the **MyCDM** database, including ages **18 to 62**, and analyzed on **20190101**.
- evaluationOutputFileName - A string designation for the evaluation cohort file. Recommended structure: "Evaluation_(xSpec Name)*(CDM)*(Qualifiers)_(Analysis Date)", e.g.,"Evaluation_10XStroke_MyCDM_Age18-62_20190101" to designate the file was from an **Evaluation**, built on the **10 X Stroke** xSpec, using the **MyCDM** database, including ages **18 to 62**, and analyzed on **20190101**
- phenotypeEvaluationFileName - A string designation for the .csv file with the results for the phenotype algorithm evaluation
- xSpecCohortId - The number of the "extremely specific (xSpec)" cohort definition id in the cohort table (for noisy positives)
- xSensCohortId - The number of the "extremely sensitive (xSens)" cohort definition id in the cohort table (used to estimate population prevalence and to exclude subjects from the noisy positives)
- prevalenceCohortId - The number of the cohort definition id to determine the disease prevalence, usually a super-set of the exclCohort
- excludedCovariateConceptIds - A list of conceptIds to exclude from featureExtraction which should include all concept_ids used to create the xSpec and xSens cohorts

- includedCovariateIds - A list of covariate IDs that should be restricted to (default=NULL)
- addDescendantsToExclude - Should descendants of excluded concepts also be excluded? (default=FALSE)
- mainPopulationCohortId - The number of the cohort to be used as a base population for the model (default=NULL)
- baseSampleSize - The maximum number of subjects in the evaluation cohort (default=2M)
- lowerAgeLimit - The lower age for subjects in the model (default=NULL)
- upperAgeLimit - The upper age for subjects in the model (default=NULL)
- startDays - The days to include prior to the cohort start date. If the mainPopulationCohortId = 0, this should be 0 (default=0)
- endDays - The days to include after the cohort start date. By default this is set to include all the data in a subject's record (default=10000)
- gender - The gender(s) to be included (default c(8507, 8532))
- startDate - The starting date for including subjects in the model (default=NULL)
- endDate - The ending date for including subjects in the model (default=NULL)
- removeSubjectsWithFutureDates - Should dates be checked to remove future dates (default=TRUE)
- cdmVersion - The CDM version of the database (default=5)
- outFolder - The folder where the output files will be written (default=working directory)
- savePlpData - Determines whether the large PLP data file is saved. Setting this to FALSE will reduce the use of disk space (default=FALSE)
- createModel - Run the function to create the diagnostic predictive model (default=TRUE)
- createEvaluationCohort - Run the function to create the evaluation cohort (default=TRUE)
- cohortDefinitionsToTest - A dataframe with cohorts to analyze. Leave blank to not test any cohort definitions (default=Null). The dataframe must contain the following elements:

1) atlasId: The cohort ID in ATLAS.
2) atlasName: The full name of the cohort.
3) cohortId: The cohort ID to use in the package. Usually the same as the cohort ID in ATLAS.
4) name: A short name for the cohort, to use to create file names. Do not use special characters.
5) washoutPeriod: The mininum required continuous observation time prior to index date for subjects within the cohort to test.

The createChronicPhenotypeModel function will produce one or more of the following artifacts (depending on the flags that were set):
1) A Patient Level Prediction file (in .rds format) containg the information from the model building process
2) A Patient Level Prediction file (in .rds format) containg the information from applying the model to the evaluation cohort
3) A csv file containing the results from the analysis of the phenotype algorithms to be tested

For example:

```
options(fftempdir = "c:/temp/ff") #place to store large temporary files

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

phenoTest <- createChronicPhenotypeModel(connectionDetails = connectionDetails,
                cdmDatabaseSchema = "my_cdm_data",
                databaseId = "TestDB",
                cohortDatabaseSchema = "my_results",
                cohortTable  = "cohort",
                workDatabaseSchema = "scratch.dbo", #a database schema with write access
                modelOutputFileName = "Train_10XDM_MyCDM_18-62_20190101",
                evaluationOutputFileName = "Eval_10XDM_MyCDM_18-62_20190101",
```

```
                xSpecCohortId = 1769699,
                xSensCohortId = 1770120,
                prevalenceCohortId = 1770120,
                excludedCovariateConceptIds = c(201820),
                includedCovariateIds = c(),
                addDescendantsToExclude = TRUE,
                mainPopulationCohortId = 0, #use the entire subject population
                baseSampleSize = 2000000,
                lowerAgeLimit = 18,
                upperAgeLimit = 90,
                startDays = 0, #from the start of the subject's record
                endDays = 10000, #to the end of the subject's record
                gender = c(8507, 8532),
                startDate = "19000101",
                endDate = "21000101",
                removeSubjectsWithFutureDates = TRUE,
                outFolder = "c:/phenotyping",
                savePlpData = FALSE, #will preserve disk space
                createModel = TRUE, #will create a model
                createEvaluationCohort = TRUE, #will create an evaluation cohort
                cohortDefinitionsToTest = cohortSetReference)
```

In this example, we used the cohorts developed in the "my_results" cdm, specifying the location of the cohort table (cohortDatabaseSchema, cohortTable - "my_results.cohort") and where the model will find the conditions, drug exposures, etc. to inform the model (cdmDatabaseSchema - "my_cdm_data"). The subjects included in the model will be those whose first visit in the CDM is between January 1, 2010 and December 31, 2017. We are also specifically excluding the concept ID 201826, "Type 2 diabetes mellitus", which was used to create the xSpec cohort as well as all of the descendants of that concept ID. Their ages at the time of first visit will be between 18 and 90. The dataframe, cohortSetReference (see first section of this vignette), is now used to define the cohort definitions for which we want performance characteristics.

In this example, the parameters specify that the function will create the model file:
"c:/phenotyping/Train_10XDM_MyCDM_18-62_20190101.rds",

produce the evaluation cohort file:
"c:/phenotyping/Eval_10XDM_MyCDM_18-62_20190101.rds"

and produce the phenotype algorithm performance characteristics result file:
"c:/phenotyping/PerformanceResultsDiabetes_TestDB.csv"

2.4 Evaluating phenotype algorithms for acute health conditions (those requiring an in-patient hospital visit)

The function createAcutePhenotypeModel allows the user to perform a complete analysis for determining the performance characteristics for one or more phenotype algorithms (cohort defintions) for acute health conditions. This function initiates the process for the three major steps in PheValuator, namely:
1) Develop a diagnostic predictive model for the health condition.
2) Select a large, random set of subjects from the dataset and use the model to determine the probability of each of the subjects having the health condition.
3) Determine the performance characteristics for one or more phenotype algorithms to be used in studies.

createAcutePhenotypeModel should have the same inputs as described for createAcutePhenotypeModel above but with more attention paid to the following parameters:

- startDays - The days to include prior to the cohort start date. For acute health conditions, this will be the number of days preceding the in-patient hospital visit. (default=0)

- endDays - The days to include after the cohort start date. (default=7, i.e., 7 days after the start of the hospital visit)
- mainPopulationCohortIdStartDay - When specifying a mainPopulationCohortId, the number of days relative to the mainPopulationCohortId cohort start date to begin including visits. (default=0, i.e., 0 days before the start of the mainPopulationCohortId cohort start date)
- mainPopulationCohortIdEndtDay - When specifying a mainPopulationCohortId, the number of days relative to the mainPopulationCohortId cohort start date to end including visits. (default=0, i.e., 0 days after the start of the mainPopulationCohortId cohort start date)
- visitLength - The minimum length of index visit for noisy negative comparison (default=3 days). As a guideline, use the average length of a hospital visit for those with the health condition.

For example:

```
options(fftempdir = "c:/temp/ff") #place to store large temporary files

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

phenoTest <- createAcutePhenotypeModel(connectionDetails = connectionDetails,
                cdmDatabaseSchema = "my_cdm_data",
                databaseId = "TestDB",
                cohortDatabaseSchema = "my_results",
                cohortTable  = "cohort",
                workDatabaseSchema = "scratch.dbo", #a database schema with write access
                modelOutputFileName = "Train_Pneumonia_MyCDM_18-62_20190101",
                evaluationOutputFileName = "Eval_Pneumonia_MyCDM_18-62_20190101",
                xSpecCohortId = 1769699,
                xSensCohortId = 1770120,
                prevalenceCohortId = 1770120,
                excludedCovariateConceptIds = c(255848),
                includedCovariateIds = c(),
                addDescendantsToExclude = TRUE,
                mainPopulationCohortId = 0, #use the entire subject population
                mainPopulationCohortIdStartDay = 0,
                mainPopulationCohortIdEndDay = 0,
                baseSampleSize = 2000000,
                lowerAgeLimit = 18,
                upperAgeLimit = 90,
                startDays = 0, #from the start of the subject's record
                endDays = 7, #to the end of the subject's record
                visitLength = 3,
                gender = c(8507, 8532),
                startDate = "19000101",
                endDate = "21000101",
                removeSubjectsWithFutureDates = TRUE,
                outFolder = "c:/phenotyping",
                savePlpData = FALSE, #will preserve disk space
                createModel = TRUE, #will create a model
                createEvaluationCohort = TRUE, #will create an evaluation cohort
                cohortDefinitionsToTest = cohortSetReference)
```

In this example, we are using only the expected value ("EV"). Given that parameter setting, the output from this step will provide performance characteristics (i.e, sensitivity, specificity, etc.) at each prediction

threshold as well as those using the expected value calculations as described in the Step 2 diagram. The evaluation uses the prediction information for the evaluation cohort developed in the prior step. The data frames produced from this step may be saved to a csv file for detailed examination.