EvaluatingPhenotypeAlgorithms

Joel N. Swerdel

2020-03-09

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 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.

2.2 Creating the Diagnostic Predictive Model

The function createPhenotypeModel develops the diagnostic predictive model for assessing the probability of having the HOI in the evaluation cohort.

createPhenotypeModel should have as inputs:

- connectionDetails - connectionDetails created using the function createConnectionDetails in the DatabaseConnector package.
- xSpecCohort - The number of the "extremely specific (xSpec)" cohort definition id in the cohort table (for noisy positives)
- 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'.

- 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.
- cohortDatabaseTable - 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.
- outDatabaseSchema - 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**.
- xSensCohort - 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)
- prevalenceCohort - The number of the cohort definition id to determine the disease prevalence, usually a super-set of the exclCohort
- excludedConcepts - A list of conceptIds to exclude from featureExtraction which should include all concept_ids used to create the xSpec and xSens cohorts
- addDescendantsToExclude - Should descendants of excluded concepts also be excluded? (default=FALSE)
- mainPopulationCohort - The number of the cohort to be used as a base population for the model (default=NULL)
- lowerAgeLimit - The lower age for subjects in the model (default=NULL)
- upperAgeLimit - The upper age for subjects in the model (default=NULL)
- 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)
- checkDates - Should observation period dates be checked to guard against errors such as dates in the future? (default=TRUE)
- cdmVersion - The CDM version of the database (default=5)
- outFolder - The folder where the output files will be written (default=working directory)

The createPhenotypeModel function creates a PLP model to be used for determining the probability of the HOI in the evaluation cohort.

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 <- PheValuator::createPhenotypeModel(connectionDetails = connectionDetails,
                               xSpecCohort = 1769699,
                               xSensCohort = 1770120,
                               cdmDatabaseSchema = "my_cdm_data",
                               cohortDatabaseSchema = "my_results",
                               cohortDatabaseTable = "cohort",
                               outDatabaseSchema = "scratch.dbo", #a database schema with write access
                               modelOutputFileName = "Train_10XDM_MyCDM_18-62_20190101",
                               prevalenceCohort = 1770119, #the cohort for prevalence determination
                               excludedConcepts = c(201820),
                               addDescendantsToExclude = TRUE,
```

```
                                mainPopulationCohort = 0, #use the entire subject population
                                lowerAgeLimit = 18,
                                upperAgeLimit = 90,
                                startDate = "20100101",
                                endDate = "20171231",
                                checkDates = TRUE,
                                outFolder = "c:/phenotyping")
```

In this example, we used the cohorts developed in the "my_results" cdm, specifying the location of the cohort table (cohortDatabaseSchema, cohortDatabaseTable - "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. With the parameters above, the name of the predictive model output from this step will be: "c:/phenotyping/Train_10XDM_MyCDM_18-62_20190101.rds"

2.3 Creating the Evaluation Cohort

The function createEvaluationCohort uses the PLP function applyModel to produce a large cohort of subjects, each with a predicted probability for the HOI.

createEvaluationCohort should have as inputs:

- connectionDetails - connectionDetails created using the function createConnectionDetails in the DatabaseConnector package.
- xSpecCohort - The number of the "extremely specific (xSpec)" cohort definition id in the cohort table (for noisy positives)
- xSensCohort - 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)
- cdmDatabaseSchema - The name of the database schema that contains the OMOP CDM instance. Requires read permissions to this database. On SQL Server, this should specifiy both the database and the schema, so for example 'cdm_instance.dbo'.
- cohortDatabaseSchema - The name of the database schema that is the location where the cohort data used to define the at risk cohort is available. Requires read permissions to this database.
- cohortDatabaseTable - The tablename that contains the at risk cohort. The expectation is cohortTable has format of COHORT table: cohort_concept_id, SUBJECT_ID, COHORT_START_DATE, COHORT_END_DATE.
- outDatabaseSchema - The name of the database schema that is the location where the data used to define the outcome cohorts is available. Requires read permissions to this database.
- 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**
- modelOutputFileName - A string designation for the training model file
- mainPopulationCohort - The number of the cohort to be used as a base population for the model (default=NULL)
- 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 (default=-10000)
- endDays - The days to include after the cohort start date (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)

- 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)

For example:

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


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

evalCohort <- PheValuator::createEvaluationCohort(connectionDetails = connectionDetails,
                                xSpecCohort = 1769699,
                                xSensCohort = 1770120,
                                cdmDatabaseSchema = "my_cdm_data",
                                cohortDatabaseSchema = "my_results",
                                cohortDatabaseTable = "cohort",
                                outDatabaseSchema = "scratch.dbo",
                                evaluationOutputFileName = "Eval_10XDM_MyCDM_18-62_20190101",
                                modelOutputFileName = "Train_10XDM_MyCDM_18-62_20190101",
                                mainPopulationCohort = 0,
                                lowerAgeLimit = 18,
                                upperAgeLimit = 90,
                                startDate = "20100101",
                                endDate = "20171231"
                                outFolder = "c:/phenotyping",
                                savePlpData = FALSE)
```

In this example, the parameters specify that the function should use the model file: "c:/phenotyping/Train_10XDM_MyCDM_18-62_20190101.rds" to produce the evaluation cohort file: "c:/phenotyping/Eval_10XDM_MyCDM_18-62_20190101.rds" The evaluation cohort file above will be used the evaluation of the PAs provided in the next step.

2.4 Creating the Phenotype Algorithms for evaluation

The next step is to create the PAs to be evaluated. These are specific to the research question of interest. For certain questions, a very sensitive algorithm may be required; others may require a very specific algorithm. For this example, we will test an algorithm which requires that the subject have a diagnosis code for DM from an in-patient setting where the code was specified as the primary reason for discharge. An example of this algorithm is in the OHDSI ATLAS repository. The output of this function is a list containing 2 data frames, one with the results of the PA evaluation and a second with a set of subject IDs that were determined to be true positives, false positives, or false negatives based on prediction threshold of 50%. A true positive, with this criteria, would be a subject that was included in the PA and also had a predicted value for the HOI of 0.5 or greater. A false positive would be a subject who was included in the PA and whose predicted probability was less than 0.5. A false negative would be a subject who was not included in the PA but had a predicted probability of the HOI or 0.5 or greater.

testPhenotypeAlgorithm should have as inputs:

- connectionDetails - ConnectionDetails created using the function createConnectionDetails in the DatabaseConnector package.
- cutPoints - A list of threshold predictions for the evaluations. Include "EV" for the expected value

- evaluationOutputFileName - The full file name with path for the evaluation file
- phenotypeCohortId - The number of the cohort of the phenotype algorithm to test
- phenotypeText - A string to identify the phenotype algorithm in the output file
- cdmShortName - A string to identify the CDM tested (Default = NULL)
- order - The order of this algorithm for sorting in the output file (used when there are multiple phenotypes to test) (Default = 1)
- modelText - Descriptive name for the model (Default = NULL)
- xSpecCohort - The number of the "extremely specific (xSpec)" cohort definition id in the cohort table (for noisy positives) (Default = NULL)
- xSensCohort - The number of the "extremely sensitive (xSens)" cohort definition id in the cohort table (used to exclude subjects from the base population) (Default = NULL)
- prevalenceCohort - The number of the cohort definition id to determine the disease prevalence, (default=xSensCohort)
- cohortDatabaseSchema - The name of the database schema that is the location where the cohort data used to define the at risk cohort is available. Requires read permissions to this database.
- cohortTable - The tablename that contains the at risk cohort. The expectation is cohortTable has format of COHORT table: cohort_concept_id, SUBJECT_ID, COHORT_START_DATE, COHORT_END_DATE.
- washoutPeriod - The washoutPeriod is used when testing algorithms where there is an enforced prior observation period before the index date

For example:

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

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

phenoResult <- PheValuator::testPhenotypeAlgorithm(connectionDetails = connectionDetails,
                cutPoints = c("EV"),
                evaluationOutputFileName = "c:/phenotyping/lr_results_Eval_10X_DM_MyCDM.rds",
                phenotypeCohortId = 1769702,
                cdmShortName = "myCDM",
                phenotypeText = "All Diabetes by Phenotype 1 X In-patient, 1st Position",
                order = 1,
                modelText = "Diabetes Mellitus xSpec Model - 10 X T2DM",
                xSpecCohort = 1769699,
                xSensCohort = 1770120,
                prevalenceCohort = 1770119,
                cohortDatabaseSchema = "my_results",
                cohortTable = "cohort",
                washoutPeriod = 0)
```

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.