Bayesian Logistic Regression for medical claims data

Bayesian logistic regression for medical claims data is a novel statistical approach that possesses the advantages of the regression analysis such as being resistance to confounding by co-medication (“innocent bystander problem”) and adjustment for masking effect [1, 5].

Statistical Model

Let's assume that Y is a response variable that takes on two possible values, 0 and 1, and $x_1,\ x_2,...,x_n$ are covariate variables. In our program Y is an indicator of the outcome of interest (condition) and $x_1,\ x_2,...,x_n$ are indicators of drugs and, possibly, of some other categorical variables (like age or sex groups). For n explanatory variables the model is:

$$\log\frac{\Pr(Y=1)}{1-\Pr(Y=1)} = \beta_0 + \beta_1 x_{1'} + ... + \beta_n x_n .$$

Estimation of the i-th coefficient, $\hat{\beta}_i$, can be interpreted as an approximation to the log-odds ratio, $\log\left(\frac{\Pr(Y=1|x_i=1)}{1-\Pr(Y=1|x_i=1)} / \frac{\Pr(Y=1|x_i=0)}{1-\Pr(Y=1|x_i=0)}\right)$, that characterizes the influence of the i-th covariate adjusted for all other explanatory variables in the model.

In the Bayesian interpretation one puts priors on the regression coefficients. Such an approach allows to avoid overfitting by assuming that each regression coefficient β_i is likely to be close to 0; see, for example, [2, 5] for more details.

Input Data

As in other statistical methods for analysis of medical claims data (see, for example, the Disproportionality Analysis Method specification at the OMOP website at http://omop.fnih.org/MethodsLibrary), mapping the original data into the form applicable for statistical analysis is an essential step. The software allows two options when dealing with conditions: incident condition type or prevalent condition type. An incident condition reflects the first occurrence of the event, while prevalent condition type allows multiple occurrences of the condition for the same patient. *Index date* of a condition is defined as a date of the first occurrence of the condition for incident cases, and as a date of occurrence for prevalent cases. In either situation, input data presented in the *common data model* (CDM) format is mapped into the set of reports. Each report contains a condition identifier (Condition_concept_id in the CDM format) and a drug, or a set of drugs. Each drug from this set has the condition index-date inside of its drug-era or a derivative of drug-era (if the risk window is different from zero):

DRUG_ERA_START_DATE + DAYS_FROM_THE_DRUG_ERA_START <=
<= INDEX_DATE <= DRUG_ERA_END_DATE + RISK_WINDOW (1)

RISK_WINDOW parameter may take on integer values. If RISK_WINDOW = -30, then the drug is included into the report only if the condition INDEX_DATE is within 30 days from the drug initiation date (*DRUG_ERA_START_DATE* in the CDM terminology).

Analysis of each condition requires fitting a separate regression model. Data set for regression analysis is formed by ‘condition’ reports and ‘no-condition’ reports. Each patient either contributes to the ‘condition’ part of the input data set or to ‘no-condition’ part, but can not contribute to both of them simultaneously. Number of ‘no-condition’ reports is controlled by the input parameter (see the description of the program input parameters below.)

Bayesian Regression Software

Having prepared input data using SAS, we fit the logistic regression for the high-dimensional data. For this purpose we utilize an open-source Bayesian Binary Regression (BBR) program [2,3,4]. BBR allows two choices of priors: Gaussian and Laplace. Another parameter that has to be specified is the variance of the prior. More details on the BBR software can be found at [3,4]. Computer system should have the SAS software and the BBR program installed to carry out the analysis described here.

Description of the parameter file, BLR_parameters.txt

Sample parameter file for the Bayesian Logistic Regression program:

<BEGIN FILE “BLR_parameters.txt”>

```
CONDITION_TYPE_PREVALENT_1_INCIDENT_2: 1
PRIOR_TYPE_LAPLACE_1_NORMAL_2: 1
PRIOR_VARIANCE: 1
INCLUDE_AGE_0_(NO)_OR_1_(YES): 1
INCLUDE_SEX_0_(NO)_OR_1_(YES): 1
SIZE_OF_THE_NO_CONDITION_PART: 1000000
RISK_WINDOW_IN_DAYS: 30
DRUG_PERSISTENCE_WINDOW_DAYS_0_OR_30: 30
CONDITION_PERSISTENCE_WINDOW_DAYS_0_OR_30: 30
NUMBER_OF_SPLITS: 5
DAYS_FROM_THE_DRUG_ERA_START: 1
DRUG_ERA_TABLE: OMOP_DRUG_ERA
CONDITION_ERA_TABLE:  CONDITION_ERA
DATABASE_NAME: NULL
```

<END_OF_FILE “BLR_parameters.txt”>

The input parameters for the BLR program are the following:

1) Conditions type: whether to consider all occurrences (1- Prevalent) or first occurrence (2- Incident) of conditions as potential outcomes.

2) Prior on the regression coefficients. Two types are allowed: Laplace and Normal (1- Laplace, 2 - Normal).

3) Prior variance: any positive number.

4) Adjustment for age (1=Yes, 0= No). If “1” then age indicator variables will be added to the set of regression covariates.

5) Adjustment for sex (1=Yes, 0= No). If “1” then sex indicator variables will be added to the set of regression covariates.

6) SIZE_OF_THE_NO_CONDITION_PART specifies the number of reports in the input data set that do not contain a condition.

7) RISK_WINDOW_IN_DAYS parameter: the period of time a patient is inferred to be ‘at-risk’ and therefore counting occurrence of conditions as potential events (ex: -30 is used to capture events that happen within 30 days of initiation of exposure; +60 is used to capture events that happen anytime during or within 60 days following the end of exposure)

8) DRUG_PERSISTENCE_WINDOW: the OMOP common data model provides supplementary tables populated with ‘drug eras’ to derive periods of exposure from disparate sources (such as prescription dispensings, procedural administrations, medication history, prescription histories). ‘Drug eras’ were built using a 0-d and 30-day persistence window assumption used to characterize continuous use. This parameter specifies which of the two assumptions to apply.

9) CONDITION_PERSISTENCE_WINDOW parameter: the OMOP common data model provides supplementary tables populated with ‘condition eras’ to derive episodes of care for a given condition, based on available information (such as diagnoses, problem lists) ‘Condition eras’ were built using a 0-d and 30-day persistence window assumption used to aggregate observations that are likely part of one period. This parameter specifies which of the two assumptions to apply.

10) To facilitate calculations, SAS/CONNECT is used to carry out regression analysis in parallel. NUMBER_OF_SPLITS parameter defines number of concurrent SAS jobs and depends on technical capabilities of the particular computer facility. Possible values of the parameter are 1,2,3….

11) DAYS_FROM_THE_DRUG_ERA_START parameter defines after how many days of exposure to a drug, a condition is considered together with the drug in the analysis:
DRUG_ERA_START_DATE + DAYS_FROM_THE_DRUG_ERA_START ≤ CONDITION_ERA_START_DATE ≤ DRUG_ERA_END_DATE + RISK_WINDOW_IN_DAYS.

12) DRUG_ERA_TABLE: name of the DRUG_ERA table in the CDM format, possible values: OMOP_DRUG_ERA or DRUG_ERA.

13) CONDITION_ERA_TABLE name of the CONDITION_ERA table in the CDM format, possible values: OMOP_CONDITION_ERA or CONDITION_ERA.

14) Database Name [alphanumeric abbreviation]. Database name parameter that is used in the name of the output file. If parameter is set to 'NULL' , program will use database name defined in the body of the SAS code.

References

1. Hauben, M., Madigan, D., Gerrits, C.M., Walsh, L, Van Puijenbroek, E.P., (2005), The role of data mining in pharmacovigilance, Expert Opin. Drug Saf., #4: 929–948.

2. Genkin, A., Lewis, D., Madigan, D., (2007). Large-Scale Bayesian Logistic Regression for Text Categorization, Technometrics, Vol. 49, No. 3: 291–304.

3. BBR: Bayesian Logistic Regression Software,
http://www.stat.rutgers.edu/~madigan/BBR/

4. Bayesian Logistic Regression (BBR, BMR, BXR),
http://www.bayesianregression.org/

5. Caster, O., Noren, G.N., Madigan, D., and Bate, A. (2010). Large-Scale Regression-Based Pattern Discovery: The Example of Screening the WHO Global Drug Safety Database. Statistical Analysis and Data Mining, to appear.