Cohort method

by Ivan Zorych, Patrick Ryan, David Madigan

Last revised: May 18, 2012

Our cohort approach to medical claims data includes selection of new-user cohorts based on two sets of drugs, target and comparator drugs. There are several scenarios available for such selection (see parameters section below for more details). Then we utilize propensity based adjustment for outcome models.

There are three ways to select covariates and calculate propensity score in our implementation. One is based on high-dimensional propensity adjustment approach as described in [1]. Our implementation works with up to three data dimensions. Default dimension includes drugs (Drug_era table) and, optionally, conditions (Condition_era table) and procedures (Procedures_occurrence table) may be included. Following [1], we identify empirical covariates, assess their recurrence, prioritize and select covariates, estimate propensity score, and, finally, estimate exposure-outcome association adjusting for propensity. The second way to calculate propensity is to use Bayesian regression that includes all variables from up to three dimensions as covariates. The third way follows covariate selection procedure implemented in [2]. This procedure is based on calculating propensity utilizing the most prevalent covariates from all selected data dimensions (drugs, conditions, procedures).

We calculate several estimates adjusted for propensity: stratified (by propensity groups) odds ratio using the Mantel-Haenszel method, adjusted odds ratio using multivariate logistic regression to model outcome as a function of exposure and indicator variables for groups of propensity score, and adjusted odds ratio from the similar multivariate logistic regression model where propensity participates as a continuous variable.

All three approaches to selecting covariates and calculating propensity scores can be carried out without or with trimming of propensity scores. If trimming is used, the range of propensity scores is trimmed according to percentiles of propensity scores in target cohort at the lower end and in the comparator cohort at the upper end [3].

Other available features include nesting within indication when only patients that had an indication are included into each, target or comparator, cohort. A number of additional covariates may be included into the model that estimates propensity: number of drugs that the person has taken, number of person's conditions, number of visits, Charlson comorbidity index [4].

Parameters

Parameters defined in the CM_parameters.txt file specify how the method is executed. Sample parameter file for the CM program:

<BEGIN FILE "CM_parameters.txt">
RUN_NUMBER: 1
WASH_OUT_PERIOD_DAYS: 180
COVARIATE_ELIGIBILITY_DAYS: 180
SURVEILLANCE_WINDOW_DAYS: 30
FREQUENCY_BY_PEOPLE: 0
TOP_N_PREVALENT_CODES: 0
TOP_CONFOUNDERS: 0
NUMBER_OF_GROUPS_BY_POPENSITY: 5 20
DRUG_ERA_TABLE: DRUG_ERA
CONDITION_ERA_TABLE: OMOP_CONDITION_ERA
DRUG_PERSISTENCE_WINDOW_DAYS_0_OR_30: 30
CONDITION_PERSISTENCE_WINDOW_DAYS_0_OR_30: 30
SELECTION_PROCEDURE_ONE_1_YES_0_NO: 0

SELECTION_PROCEDURE_TWO_1_YES_0_NO: 1

SELECTION_PROCEDURE_THREE_1_YES_0_NO: 0

PRIOR_TYPE_LAPLACE_1_NORMAL_2: 2

PRIOR_VARIANCE: 1

USE_CONDITIONS_AS_DIMENSION_1_YES_0_N: 1

USE_PROCEDURES_AS_DIMENSION_1_YES_0_N: 1

NUMBER_OF_PARALLEL_RUNS: 4

DELETE_INTERMEDIATE_FILES_1_YES_0_NO: 1

USE_SPDE_1_YES_0_NO: 0

USE_IML_1_YES_0_NO_FOR_SELECTION_ONE: 1

TRIMMING: 0

TRIM_PNTL: 0

USE_YEAR: 1

REDEFINE_DRUG_ERA: 0

DRUG_PERIOD_BACK: 0

NESTING: 0

AGE_STRATA_FILE: 0

USE_CHARLESON_IND: 1

USE_DRUG_NUM: 1

USE_VISITS_NUM: 1

USE_PROCS_NUM: 1

DATABASE_NAME: NULL

COMPARATOR: 1

<END FILE "CM_parameters.txt">

RUN_NUMBER: Parameter that identifies current run. It helps to manage multiple CM runs. The parameter is reflected in the file name of each output file.

WASH OUT PERIOD: The amount of time in days a person must be in the database in order to be included in the analysis.

COVARIATE ELIGIBILITY: Period prior to exposure during which covariates are counted.

SURVEILLANCE WINDOW: Period of time (in days) a patient is inferred to be 'at-risk' and therefore counting occurrence of conditions as outcomes (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).

FREQUENCY BY PEOPLE: Minimal number of people that should have a covariate (for the covariate to be included in the analysis).

TOP N PREVALENT CODES: N (number) most prevalent covariates to be included in the analysis (for each dimension).

TOP CONFOUNDERS: Maximal number of covariates to be included in calculation of propensity.

NUMBER OF GROUPS BY PROPENSITY SCORE: Number of groups by propensity (for example, parameter value 10 corresponds to grouping subjects by propensity deciles). Several values may be specified (separated by a blank). Example:

NUMBER_OF_GROUPS_BY_PROPENSITY_SCORE: 5 10 20

DRUG ERA TABLE: Name of the DRUG_ERA table in the CDM format, possible values: OMOP_DRUG_ERA or DRUG_ERA.

CONDITION_ERA_TABLE: Name of the CONDITION_ERA table in the CDM format, possible values: OMOP_CONDITION_ERA or CONDITION_ERA.

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, and prescription histories). 'Drug eras' were built using a 0-day and 30-day persistence window assumption. This parameter specifies which of the two assumptions applies.

CONDITION PERSISTENCE WINDOW: 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-day and 30-day persistence window assumption used to aggregate observations that are likely a part of one period. This parameter specifies which the two assumptions applies.

SELECTION PROCEDURE ONE: If 1 is selected, then the program runs the covariate selection procedure defined in [1]. Parameter can take on two values: 0 and 1.

SELECTION PROCEDURE TWO: If 1 is selected, then the program uses Bayesian regression to calculate the propensity score. Parameter can take on two values: 0 and 1.

SELECTION PROCEDURE THREE: If 1 is selected, then the program uses the selection procedure from [2]. Parameter can take on two values: 0 and 1.

PRIOR TYPE: Prior that is used in the Bayesian regression. 1 means Laplace prior, 2 – Normal prior.

PRIOR VARIANCE: the value of the hyperparameter in the Bayesian regression.

USE CONDITIONS AS DIMENSION: Include conditions (1) or exclude conditions (0) as a dimension in the analysis.

USE PROCEDURES AS DIMENSION: Include procedures (1) or exclude procedures (0) as a dimension in the analysis.

NUMBER OF PARALLEL RUNS: To facilitate calculations, SAS/CONNECT is used to carry out analysis in parallel. This parameter defines a number of concurrent SAS jobs and depends on technical capabilities of a particular computer facility. Possible values of the parameter are 1,2,3….

DELETE INTERMEDIATE FILES: Delete intermediate files if set to 1, do not delete if set to 0.

USE SPDE: Should be set to 0 (experimental parameter).

USE_IML: Use SAS proc IML if set to 1, do not use if set to 0. Recommended value is 1.

TRIMMING: Do not trim propensity scores (0), use a trimming option (1).

TRIM_PNTL: How much to trim (in percents).

USE_YEAR: (1) use year in the model that calculates the propensity score, (0) do not use year.

REDEFINE_DRUG_ERA: experimental feature, should be set to 0.

DRUG_PERIOD_BACK: experimental feature, should be set to 0.

NESTING: (1) nesting within an indication, (0) do not use nesting.

AGE_STRATA_FILE: experimental feature, should be set to 0.

USE_CHARLESON_IND: (1) use Charlson index, (0) do not use Charlson index.

USE_DRUG_NUM: (1) use number of drugs, (0) do not use number of drugs.

USE_VISITS_NUM: (1) use number of visits, (0) do not use number of visits.

USE_PROCS_NUM: (1) use number of procedures, (0) do not use number of procedures.

DATABASE NAME: Database Name [alphanumeric abbreviation]. Database name parameter which is used in the name of the output files. If parameter is set to 'NULL' , program will use database name defined in the body of the SAS code.

COMPARATOR: numbers 1 to 9. Values from 1 to 7 are reserved for different sets of target and comparator drugs under the usual cohort selection procedure. If comparator is set to be equal 8, then a comparator cohort will contain all people with indication(s) who do not belong to the target cohort, the target cohort contains people with indication who were exposed to drug(s) of interest. Comparator value 9 is an option that creates comparator cohort of negative control drugs for each target drug-condition combination.

Output files

Program produces the following output files:

Propensity score stratification using Mantel Haenszel adjustment over 20 strata (sample file name: cohort_method_run_1_comp_1_DBN_NAME**_MH_pg20**.txt)

Propensity score stratification using Mantel Haenszel adjustment over 5 strata (sample file name: cohort_method_run_1_comp_1_ DBN_NAME**_MH_pg5**.txt)
Unadjusted odds ratio from univariate logistic regression predicting outcome from exposure (sample file name: cohort_method_run_1_comp_1_ DBN_NAME**_OR**.txt)
Propensity score adjustment using 20 strata as indicator variables in logistic regression outcome model
(sample file name: cohort_method_run_1_comp_1_DBN_NAME**_PS_pg20**.txt)
Propensity score adjustment using 5 strata as indicator variables in logistic regression outcome model
(sample file name: cohort_method_run_1_comp_1_DBN_NAME**_PS_pg5**.txt)
Propensity score adjustment using propensity score as continuous variable in logistic regression outcome model (cohort_method_run_1_comp_1_DBN_NAME**_PS2**.txt)

Name of each output file contains information about the run number ("run_1" in the sample file name above), comparator ("comp_1" in the example above) and database abbreviation ("DBN_NAME").

References

1. Schneeweiss, S., Rassen, J., Glynn, R., Avorn, J., Mogun, H., and Brookhart, A., (2009), High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data, *Epidemiology,* vol. 20, pp. 512- 522.
2. Alan Brookhart, Incident User Design (IUD-HOI) Method, http://omop.fnih.org/MethodsLibrary.
3. Sturmer, T., Rothman, K.J., Avorn, J., Glynn, R.J., (2010), Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution – a simulation study, *American Journal of Epidemiology*, vol. 172, #7, pp. 843-854.

Contact: Ivan Zorych, e-mail: **please use Contact page**