Case-control estimation specification

The case-control design [1, 2, 3] has been commonly applied in retrospective studies to compare patients with a condition to those who don't have it by looking back into their medical history. The case-control approach has been suggested as an efficient design for studying adverse drug reactions within observational data, such as administrative claims and electronic health records.

This implementation of the case-control estimation approach leverages the basic design of a case-control study to enable estimates of drug-condition associations across a large set of drugs and conditions. The algorithm extracts the information necessary to yield an odds ratio, but can be applied simultaneously to multiple conditions (each acting as distinct case definitions for case-control sub-studies), and allows for all exposures to be evaluated for each outcome. In this manner, multi-set case-control estimation can be used to study specific drug-condition relationships, but is also scalable to be applied within an active surveillance context to both monitoring of Health Outcomes of Interest, and identification of non-specified conditions.

Cases

For each condition under study, cases are defined as 'incident events', were we consider the first occurrence of each condition and denote the date of such occurrence as the *index date*. All patients with a given condition create a set of cases. Other variables that are recorded for patients in the case group are sex, age, and, optionally, location, race, etc. To be selected to the case group, a patient has to be observed for at least M days prior to the index date. The default value for M is 183 (days).

Controls

All patients in the database who did not experience the condition and were enrolled/observed for at least M days are potential controls. To create a set of controls for a given case set, we divide all cases into sub-groups by sex and age.

For each case group, set we create an index-pool that contains index-dates of all the patients in the case group. Each index-date from this index-pool is associated with a sex-age group of corresponding case patients. For each index-date we select, using randomization, controls from the pool of potential controls of the same sex and age who were observed on this *index date* and were enrolled into the medical plan for at least M days prior to the *index date.* If control is selected, it is assigned this index-date as a "control index-date." We stop looking for controls for the particular index-date if some fixed number of controls (parameter Number_of_matches) is reached. We can do up to several passes (parameter Number_of_passes) through a subset of potential controls to select this number of controls. If such selection is not successful or if the selected number of controls is smaller than the desired number of matches, we move to the next index-date.

The described procedure creates case and control sets that are cross-classified into sub-groups by the matching variables.

Analysis

We analyze cases and controls selected above by employing many:to:many category matching analysis [1]. Many:to:many matching analysis is the most general matching analysis approach that is suitable for practical situations when the number of cases and controls varies depending on a condition of interest.

A drug contributes to the exposure to drug count for a particular patient in the case group or control group if the *index date* (*control index date* for a control) is within some interval of time that is related to a drug era. Input parameter *risk window* controls this relationship. Possible values of *risk window* are -30, 0,1,2,... . If $risk\ window = -30$, then drug contributes to the count only if the *index date* is within 30 days from the drug initiation date (*drug_era_start_date* in the CDM terminology). Otherwise, a drug contributes to the "exposure to drug count" if the *index date (control index date* for the controls*)* is within a period of time when the drug is taken by the patient + additional *risk_window* days:

$$drug\_era\_start\_date \le index\_date \le drug\_era\_end\_date + risk\_window.$$

Following [1] we estimate odds ratio of exposure to drug vs. no exposure as

$$\psi = \frac{\sum T_{ik}^{(rs)}}{\sum B_{ik}^{(rs)}},$$

where $T_{ik}^{(rs)} = k(s-i+k)m_{ik}^{(rs)}/(r+s)$, and $B_{ik}^{(rs)} = (i-k)(r-s)m_{ik}^{(rs)}/(r+s)$, here $m_{ik}^{(rs)}$ is the number of matched sets (sub-groups) with r cases and s controls with i exposures to the drug, including k exposed cases.

Additional features

The current version of the method includes a number of new features such as selecting controls by matching on visit dates, restricting analysis to the first occurrence of each drug, nesting within an indication, and using an option for either conditional logistic regression or Bayesian logistic regression. When the regression is used for analysis, a number of additional covariates may be included into the model: number of drugs that the person has taken, number of person's conditions, number of visits, Charlson comorbidity index [4]. See technical specification below for the set of parameters that pertain to each available option.

SAS implementation

SAS program CC_v5.sas conducts case-control estimation using the described approach.

Input Specification of the program

Sample parameter file for the Case Control program:

<BEGIN FILE "CC_parameters.txt">

1.NUMBER_OF_SPLITS: 10

2.NUMBER_OF_PASSES: 10

3.NUMBER_OF_MATCHES: 10

4.DAYS_ENROLLED: 30

5.RISK_WINDOW_IN_DAYS: 30

6.DRUG_PERSISTENCE_WINDOW_DAYS_0_or_30: 30

7.CONDITION_PERSISTENCE_WINDOW_DAYS_0_or_30: 30

8.DELETE_INTERMEDIATE_FILES_0_no_1_yes: 0

9.DAYS_FROM_THE_DRUG_ERA_START: 1

10.DRUG_ERA_TABLE: Drug_era

11.CONDITION_ERA_TABLE: Omop_condition_era

12.USE_VISITS_0_NO_1_YES: 0

13.VISITS_DAYS_WITHIN: 180

14.NEST_0_NO_1_YES: 0

15.USE_REGRESSIONS_0_No_1_YES: 0

16.BLR_0_No_1_YES: 0

17.BLR_0_No_1_YES_Conditions: 0

18.BLR_0_No_1_YES_Drgs_counts: 0

19.BLR_0_No_1_YES_Cnds_counts: 0

20.BLR_0_No_1_YES_Vists_counts: 0

21.BLR_0_No_1_YES_CHA_index: 0

22.BLR_HYPERPARAMETER: 1

23.CLR_0_NO_1_YES: 0

24.UNI_0_OTHER_1: 0

25.CHARLSON_WINDOW: 9999
26.COVARIATES_WINDOW: 30
27.DRUGS_OCCURRENCE_0_ALL_1_FIRST: 1
28.DATABASE_NAME: NULL
29.RUN_NUMBER: 56
<END_OF_FILE "cc_parameters.txt">

The input parameters for the CC program are:
I1. To facilitate calculations, data is divided into several parts (NUMBER_OF_SPLITS parameter). Possible values 1,2,3….

I2. Number of passes through the set of potential controls (NUMBER_OF_PASSES parameter). Possible values are 1,2,3…

I3. Number of matches for each case person (NUMBER_OF_MATCHES parameter). Possible values are 1,2,3. Recommended values are 3-8.

I4. Person can be selected as a control after being enrolled into the medical plan for some period of time (DAYS_ENROLLED parameter, default value is 183 days).

I5. Risk window. (RISK_WINDOW_IN_DAYS parameter). A single integer indicating length of the at-risk period for drugs following the CDM drug era (in days). A value of 0 indicates only on-drug events. A value of 30 indicates that the at-risk period extends 30 days after the CDM drug era. A large value (e.g. 99,999) indicates an indefinite at-risk period. If risk window is -30 (negative 30) then drug contributes to the count only if the *index date* is within 30 days from the drug initiation date (*drug_era_start_date* in the CDM terminology). [possible values: -30, 0,1,2,…]

I6-I7. Persistence windows for drug_eras and condition_eras. (DRUG_PERSISTENCE_WINDOW and CONDITION_PERSISTENCE_WINDOW parameters) [number: 0 or 30] See CDM documentation for more details.

I8. Option that allows either to keep intermediate files (DELETE_INTERMEDIATE_FILES=0) or delete them (DELETE_INTERMEDIATE_FILES=1). Intermediate files are helpful during the initial runs or runs that use subsets of data.

I9. The parameter defines if the first day on a drug is considered as a day of exposure. If DAYS_FROM_THE_DRUG_ERA_START=0, then the first day on a drug is the first day of exposure. If DAYS_FROM_THE_DRUG_ERA_START=1, then the method considers a person to be exposed to the drug starting from the second day of the drug era.

I10. Name of the drug era dataset

I11. Name of the condition era dataset.

I12. USE_VISITS_0_NO_1_YES: 1=Match on visits , 0=do not match on visits.

I13. Parameter defines time period (days) for matching on visit date. Control visit date should be within VISITS_DAYS_WITHIN days from the case index date.

I14. Nesting on indication: 1 - Yes, 0 - No (nesting).

I15. Technical parameters that switches on a regression part of the code.

I16. Use Bayesian regression: 1 - Yes, 0 - No.

I17. BLR_0_No_1_YES_Conditions: include conditions as covariates, 1 - Yes, 0 - No.

I18. BLR_0_No_1_YES_Drgs_counts: include drugs count as a covariate, 1 - Yes, 0 - No.

I19. BLR_0_No_1_YES_Cnds_counts: include conditions count as a covariate, 1 - Yes, 0 - No.

I20. BLR_0_No_1_YES_Vists_counts: include visits count as a covariate, 1 - Yes, 0 - No.

I21. BLR_0_No_1_YES_CHA_index: include Charlson index as a covariate, 1 - Yes, 0 - No.

I22. BLR_HYPERPARAMETER: hyperparameter values in the Bayesian regression model.

I23. CLR_0_NO_1_YES: Use conditional logistic regression: 1 - Yes, 0 - No.

I24. To use additional covariates, such as Charlson index or drugs/visits/conditions counts in the conditional logistic regression model this parameter should be set to 1, otherwise=0.

I25. Technical window parameter to calculate Charlson index, default value should be 9999.

I26.COVARIATES_WINDOW [days]. It defines how additional parameters are calculated in case if regression is used. For example, if this parameter is set to 30, then the code will calculate number of visits before the index date (the same true for number of drugs, conditions).

I27.DRUGS_OCCURRENCE_0_ALL_1_FIRST. If 0 is selected, then all occurrences of a drug participate in the analysis. If 1 is selected, then only the first drug era of a drug is used.

I28. Database Name (NAME_OF_DATABASE_4_LETTERS). [alphanumeric (4 letter abbreviation)].

I29. RUN_NUMBER. It is a technical parameter that helps to manage multiple runs. Run number will appear in the name of the output file.

Other important input files/parameters:

Targets. [txt files] Text file(s) – column of drug identifiers, <drugs_of_interest.txt>, and column of condition identifiers, <conditions_of_interest.txt>. File <pairs_of_interest.txt> contains two columns, drug_concept_id and condition_concept_id, file restricts all calculations to the specified pairs thus improving the performance.
<indications.txt> contains two columns, a drug id and its indication.
Two additional files, <CharlsonConcepts.txt> and <CharlsonScoring.txt>, are needed to incorporate Charlson index into the analysis.

Output Specification

CC programs outputs text file and SAS file with a row per drug-event combination. Each row comprises <drug identifier>,<condition identifier>, <score> where score is a real-valued number.

Each output filename begins with an identifier for the method (CC case-control method) followed by a module version number (for example, "v5"), the target database name, and the values of the module-specific parameters.

Thus, for example, the CC program might output a file with the name:
CCv5_OSIM_s30d30c0p3m8_vis_1_nest_0_drug_1.txt (CC program, version 5, risk window 30, drug persistence window 30, condition persistence window 0, passes 3, matches 8, match on visits, no nesting, use the first drug era only).

The program also creates summary text file Summary_*.txt that contains the following information for each drug/condition pair of interest: drug concept id, condition concept id, number of exposed cases, number of exposed controls, number of cases, number of controls.

References

1. Woodward, Mark, (1999), Epidemiology: Study design and data analysis, Chapman & Hall /CRC.

2. Agresti, Allan, (2002), Categorical Data Analysis, Wiley series in Probability and Statistics.

3. Breslow, Norman E., Day, Nicholas E., (1993), Statistical Methods in Cancer Research: The Analysis of Case-control Studies Vol .1, International Agency for Research on Cancer.

4. Charlson, M.E., Pompei P., Ales, K.L., MacKenzie, C.R., (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chron Dis, 40(5): 373-383.

Contact: Ivan Zorych, e-mail: **zorych@gmail.com**