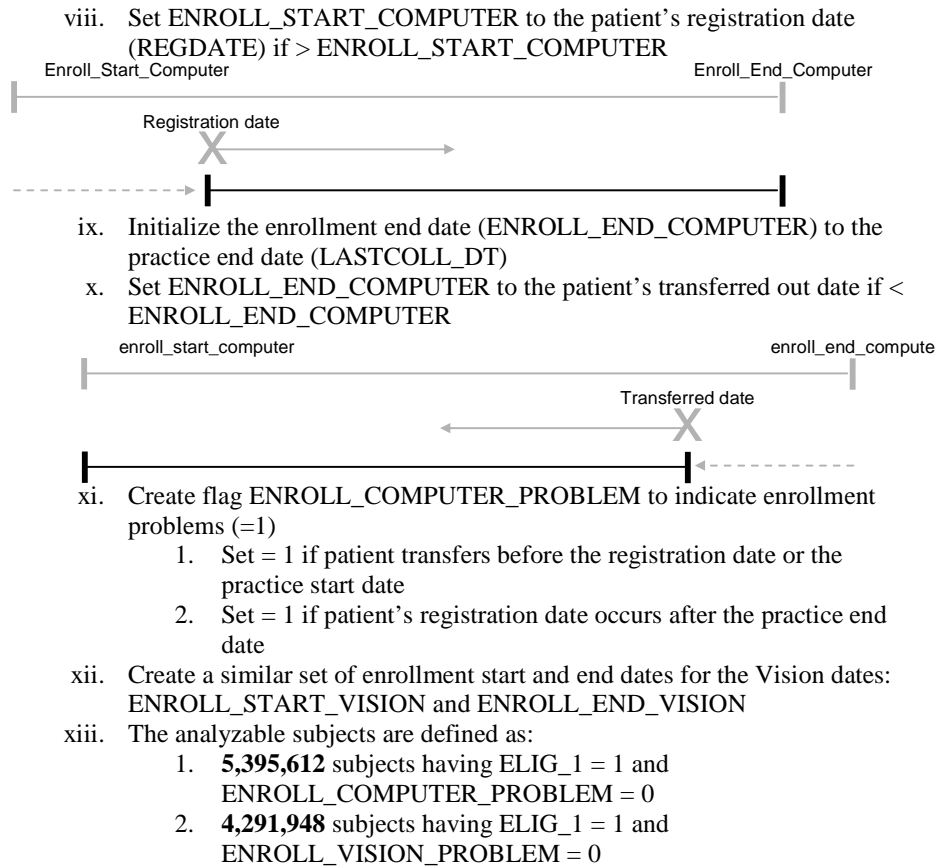


Technical Description: Case-Control analysis

This document is used as a guide to describe the analytic methods and process to conduct a case-control analysis in the THIN database.

Develop a technical plan from the analytic plan: The technical plan is the analysts guide to the analysis. This is a detailed version of the analytic plan where the database is made operational (analyzable), specific fields are described and named, methods are detailed and the exact final analyses are listed. The analytic plan is a summary of the analysis and is typically a more detailed version of the analytic methods based on the broad summary described in the study protocol. The following sections layout a technical plan example >>

1. **Create an operational data store (ODS) of the THIN database (“Load_Tables.sas”)**
 - a. Store in folder “THIN_0707” – THIN updates come as total refresh of database and will be named according to month and year (ex. July 2007)
 - b. Edit the THIN SAS main data files
 - i. 5 data files containing all 358 practices
 1. 5 data types:
 - a. Medical
 - b. Patient
 - c. Therapy
 - d. Additional health (AHD)
 - e. Dosage
 - f. PVI
 2. Rename the original tables as MEDICAL_ORIG, etc
 - c. Load the reference data files to SAS
 - a. READ codes for medical data
 - b. MULTILEXEID for therapy data
 - d. Run the program “Load_Tables.sas” to create the operational data tables – see Appendix 1 for program
 - i. Create unique patient identifier SUBJID = PATID||PRACID on all tables (this the key field matching patient level data)
 - ii. Convert the date fields to SAS date fields
 - iii. Operational main tables are named without the “_ORIG” suffix
 - iv. Create inclusion flags indicating patients eligible for analysis (ELIG_1 = 1 where PATFLAG = A or C)
 - v. Attach the practice computerization start and end dates - these dates indicate the time period that the practice is live in the database
 1. COMP_DT = the first day the practice starting using a computer system to record all medical information
 2. VISION_DT = the first day the practice starting using the Vision computer system to record all medical information (usually on or after the COMP_DT)
 3. LASTCOLL_DT = the date when the practice stopped recording medical information (this date will change with database updates – moving forward)
 - vi. Create clean “enrollment” periods for each patient based on the overlap of the practice start and end dates and the patient’s registration and transfer dates
 - vii. Initialize the enrollment start date (ENROLL_START_COMPUTER) to the COMP_DT



2. Creating reference tables

("Pgm1_ExtractData.sas")

- a. This process is iterative and will require initial guidance and medical review
- b. Start with key words describing the disease/outcome of interest
- c. Review the outputted data set for additional key words and specifically hierarchical grouping of medical codes
- d. Add indicator flags to the reference table describing the search criteria – this will help the medical reviewers understand how the selection was made
- e. Make an initial extraction based on these codes
 - i. Include the unique patient counts on the reference table – this will help the reviewers when making code list decisions
- f. Output the potential list to an excel table adding filter buttons for querying the list
- g. Have medical reviewers flag the codes of interest
- h. Load the final list to SAS and use in the extraction of data

3. Study population

("Pgm2_Cohorts.sas")

- a. Create a table of subjects who fulfill the following criteria from the [PATIENT] THIN database table
- b. Build a patient level analytic table in the project folder and name [PATIENT_ANALYSIS]
 - i. This is the analytic table for the analysis and all analytic variables will be appended to this table
- c. The table includes patients fulfilling the following criteria:
 - i. ELIG_1 = 1 and ENROLL_COMPUTER_PROBLEM = 0
 - ii. Having at least 6-months of enrollment after Jan 1, 1992 (for this example):

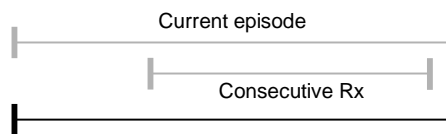
1. (ENROLL_END_COMPUTER – ENROLL_START_COMPUTER + 1) >= 182-days
 2. ENROLL_END_COMPUTER >= Jan 1, 1992
 - iii. < 80-years old on Jan 1, 1992
 1. THIN date of birth does not contain day and month in most cases (Jan 1st imputed for month and day) – children may have month
 2. Subjects turning 80 years old after Jan 1, 1992 will be truncated at the 80th birthday for the study period
 - d. Create the study index date – the later of the following 2 dates:
 1. January 1, 1992
 2. ENROLL_START_COMPUTER + 365
 - e. Create study outcome date as the first occurrence of disease/outcome in the follow-up period: DISEASE_DT
 - f. Create outcome flag if the subject has disease/outcome in the follow-up period: DISEASE_FLAG
 - g. Create the STUDY_END_DT date as the earlier of the following dates:
 - i. Disease/outcome case date
 - ii. Exclusion criteria (as listed in section 3 a-1): baseline diagnosis/drug use
 1. EXCLUSION_DATE = first occurrence of any of these diagnoses
 - iii. 80th birthday
 - iv. ENROLL_END_COMPUTER
4. **Baseline exclusion flags** - flag each anytime prior to index date and in one-year baseline (name each flag as BASE_1YR_DISEASE, BASE_ANY_DISEASE, etc): (“Pgm2b_Cohorts_BaseCovar.sas”)
- a. Disease of interest
 - b. Other diagnosis 1
 - c. Other diagnosis 2
 - d. Other diagnosis 3
 - e. Etc...
5. **Selection of the controls** (“Pgm3_Matching.sas”)
- a. Use N:1 matching of disease/outcome cases to eligible controls – the potential controls come from the set of subjects not selected as disease/outcome cases
 - b. Create table of cases from PATIENT_ANALYSIS that have DISEASE_FLAG = 1
 - c. Create variables:
 - i. CASE_FLAG = 1 for all subjects in this table (eventually will be added with the control subjects)
 - ii. AGE at index date
 - iii. PRACTICE (sub-selected from the SUBJID)
 - d. Merge the disease/outcome cases to the PATIENT_ANALYSIS table selecting the set of potential controls for each case
 - i. Create fields:
 1. CASE_FLAG = 0
 2. AGE at index date
 3. PRACTICE (sub-selected from the SUBJID)
 4. CASE_ID = case SUBJID
 5. CASE_STUDY_END_DT = disease/outcome case date
 6. RANDNUM = randuni(θ)
 - ii. Merging criteria:
 1. SUBJID is not equal to the case ID
 2. SEX equals case’s SEX
 3. AGE is +/-5 from case’s AGE
 4. Practice is the same as case’s practice
 5. The disease/outcome case date occurs with the control’s follow-up period

- iii. See Appendix 2 for SAS code
6. **Cohort flags** – during the follow-up time (“Pgm3b_CC_Exposure.sas”)
- a. Exposed cohort: flag subjects for the following drug use between the index date and the study end date
 - i. Drug A exposure
Drug B exposure
 - ii. Unexposed cohort - subjects not having either of the exposed flags
 - iii. Use field names:
 - 1. EXPOSED_DRUGA = 1 for subjects having drug A between index date and study end date
 - 2. EXPOSED_DRUGB = 1 for subjects having drug B between index date and study end date
 - 3. Unexposed cohort have EXPOSED_DRUGA = 0 and EXPOSED_DRUGB = 0
 - b. Sub-cohort - flag subjects having “exposure diagnosis” prior to index date (these subjects can be exposed and unexposed subjects): EXPOSURE_DX
7. **Extract cases for validation:**
We will evaluate the number of disease/outcome cases we have in our analysis and determine if all cases or a random set are to be sent to EPIC for validation review. If there are 1,100 or more cases, we will select 1,000 random cases. If there are < 1,100, we will send all disease/outcome cases in for review.
8. **Covariates** – create the following fields or indicators on the patient level table to be used for controlling confounding in the models (“Pgm3c_CC_Covariates.sas”)
- a. Age at index date
 - i. Continuous (AGE)
 - ii. Categories (AGE_CAT)
 - 1. ≤5
 - 2. 6-9
 - 3. 10-14
 - 4. 15-19
 - 5. 20-29, 30-39, 40-49, ... 70-79
 - b. Gender (SEX = 1 males, = 0 females)
 - c. Diagnosis of disease covariate of interest 1
 - i. Flag anytime in baseline (BASE_ANY_dx1)
 - ii. Flag anytime in follow-up (FOLUP_ANY_dx1)
 - d. Diagnosis of disease covariate of interest 2
 - i. Flag anytime in baseline (BASE_ANY_dx2)
 - ii. Flag anytime in follow-up (FOLUP_ANY_dx2)
 - e. Severity of “exposure diagnosis”
 - i. This flag is determined on the visits to a specialist after diagnosis of exposure diagnosis. Investigate the number visits and time to first visit to make a determination.
9. **Exposure to medications**
 (“Pgm4_Exposure_Episodes.sas”)
 (“Pgm5_Exposure_Categories.sas”)
- a. Therapy episodes: create course of therapy periods for each patient based on days supplied and fill dates
 - b. Defining the days supplied
 - i. Initially defined as the prescription quantity divided by the units per day (= PRSCQTY / DOSGVAL)
 - ii. Review the frequency distribution to evaluate outlying values

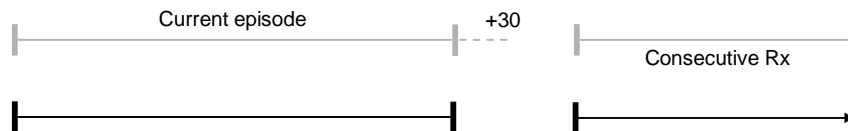
- iii. Reference dosing information offered by the product website helps re-define these values
 - iv. About 15% of the scripts will have an un-definable DOSGVAL (units per day). This field is created by EPIC universally for the database from the text input from the physician's written script
 - v. Match to the dosage reference table for "-1" records to review the written text to help define the units per day
 - vi. Some of the text dosage data will have information to define the units per day and others will read something like "USE AS DIRECTED". These are the records typically coded as "-1" by EPIC in the DOSGVAL field
 - vii. Refer to the product website for duration of therapy for these scripts. Records that are un-definable will require us to make a judgment call on the units per day, prescription quantity and/or days supplied
 - viii. This does not have to be absolutely perfect because we are using these fields to create a proxy for days supplied to ultimately define episodes of therapy use
 - ix. We will be applying a buffer or bridge to link continuous therapy so there is some variability
- c. Prescription end date = prescription fill date (PRSCDATE) + days supplied (computed field)
- d. For each patient, compare consecutive prescriptions (for like drugs) to evaluate if there is continuous therapy
- i. Sort the drug prescriptions by patient ID and prescription fill date (for each drug)
 - ii. Episodes begin with the first prescription
 - iii. If the next prescription begins prior to the end plus 7-days of the first prescription, then the episode is extended to the end date of the 2nd prescription



1. If the 2nd prescription's start and end date are contained within the 1st prescription's therapy period, use the 1st prescription's end date as the marker for the current episode end



- iv. Continue this method of comparing consecutive prescription start and end dates to the current episodes dates
- v. When a consecutive prescription's start date occurs after the current episode's end date (+30-days), then the current episode is complete and a new episode is started with the consecutive prescription's start and dates. Continue the process for each patient.



- e. Defining the therapy episodes for exposures:
- i. DrugA episodes = Drug A episodes not overlapping with Drug B
 - ii. DrugB episodes = Drug B episodes not overlapping with Drug A
 - iii. DrugAB episodes = Drug A and Drug B episodes that overlap
 - iv. None = periods where there is not any Drug A or Drug B therapy
- f. Classifying the therapy use:

- i. One-course users have ≤ 30 -days therapy episode
- ii. Intermittent users have 31+ days of drug therapy covering $\leq 50\%$ of follow-up time
- iii. Persistent users have 31+ days of drug episode therapy covering $> 50\%$ of follow-up time

10. **Analysis**

(“Pgm6_Table1_Counts.sas”)

(“Analysis.do – Stata program”)

- a. Incidence rate of disease/outcome
 - i. Overall
 - ii. Among Exposure diagnosed subjects (sub-groups)
 - iii. By age groups (see above)
 - 1. Also among <20 and ≥ 20 years old
 - iv. Computed as the number of new disease/outcome cases divided by the person-time for subjects at risk (no history of disease/outcome)
- b. Describe the relationships between:
 - i. Drug exposures
 - ii. Severity of disease
 - iii. Occurrence of disease/outcome
 - iv. Use cross-tabs to show distributions
 - v. Run conditional logistic regression to compute a formal relationship computing OR and 95% CI
- c. Tabulate the distribution of the covariates among the controls
 - i. Stratify exposure levels (see section 8.e above)
 - ii. Evaluate the distributions to consider confounders to be controlled for in the logistic regression analyses
- d. Constructing the final adjusted model:
 - i. Run a “one-by-one” stepwise regression model to exclude confounders
 - 1. Delete confounding variable having the smallest impact
 - 2. Use $\leq 10\%$ level for exclusion
 - ii. This is an iterative process with group consideration at each level
- e. Sensitivity analysis: duplicate all analyses including patients with ≥ 6 -months of follow-up only (≥ 182 days)