

***IU Method Enhancement Program Design  
Observational Medical Outcomes Partnership (OMOP)  
Research Collaborations  
October 10, 2011***

***Xiaochun Li, PhD  
Department of Biostatistics, Indiana University School of Medicine***

**1. Background**

The success of an epidemiology study for either drug safety surveillance or comparative effectiveness (CER) depends largely on the design. In the last round of collaboration with OMOP, we developed a simple cohort approach HSIU in the last grant to screen associations between medications and non-specified conditions in very large clinical and administrative databases. We learned through OMOP's extensive testing of all methods that analytic results are extremely sensitive to design features of methods. To better control the balance between sensitivity and specificity and to improve the overall accuracy through bias reduction, we anticipate that the implementation of a series of options that allow flexibility in the cohort study design and inclusion of user-defined covariates will help to achieve better classification of drug-outcome associations and enhance the computational speed. The flexible design features will also allow investigators to choose a design for a specific investigation of a drug and an outcome based on the characteristics of the drug and the nature of the outcome. Although our enhanced method will allow high-throughput screening of associations between medications and non-specified conditions through looping the program, its main purpose is to perform the 'traditional' type of epidemiology studies but with 'knobs' available to vary the design parameters to facilitate sensitivity analyses to test the robustness of the results.

**2. Scope of Work**

We propose to explore cohort-design methods, and to develop a program that will allow us to study outcomes with flexible design features. Specifically, we would like develop a generic cohort program but beyond the scope of the existing HSIU method by adding or modifying various design features, for example, requirement of certain baseline duration, options for follow-up to better define exposed patient cohort and 'control' patient cohort. For effect estimation in large databases, the prominent issue is to control and account for sources of bias. Therefore it is important to have options to allow user to include any covariates for confounding adjustment in analysis. These design features are for the purpose of flexibly tailoring the designs for bias reduction.

All options will be evaluated in an incremental fashion to assess the added value in terms of correct classification of drug-outcome association test cases.

### 3. Program Design Considerations

a. *Design*

We will allow the options of either a New User design or a prevalent user design with various user specified features, e.g., length of washout, size of surveillance window for events.

b. *Eligibility criteria, analysis covariates and outcomes*

We put these together as they are similar from the programming point of view. We envision a parameter file a la fashion of RICO will enable a user to define a variable (whether eligibility, analysis covariates or outcomes) by specifying the data table(s), time windows and logical rules. Data tables can be any of the following: drug era, condition era, observation, procedure and visit. Time windows are typically relative to an index date (drug or event). Logical rules combine various data elements from different tables for the definition of a variable, e.g., for the definition of a certain outcome, we may require certain ICD9s for diagnoses and some lab test results and/or occurrence of certain procedures.

This is an *important* enhancement. Most current methods are cumbersome to modify to allow adjustment of user-specified covariates. This feature will allow users to flexible select patients cohort(s), adjust for any covariates they wish to (as long as their data supports) and analyze any outcome of their interest (again as long as data supports).

c. *Analysis*: options for plugging in various statistical methods to allow group comparisons with user-specified confounding adjustment.

d. *Output*: SAS data sets of the following

1. Risk estimates and standard errors for each combination of parameter settings.
2. Patient characteristics, such as Table 1 in a typical epidemiology paper.

## 4. Program Design Details

### 4.1 Design- selecting the cohort

#### **Selection and bridging of relevant drug eras**

- a. Obtain the drug era table of all drug eras of the target drug(s) and the comparator drug(s)
- b. Implement bridging if there are multiple drug\_concept\_ids in either the target drug group or the comparator drug group
- c. New user or prevalent user cohort selection from the drug era table of step a.

Target and comparator drug concept ids will be defined as 2 separate lines as the parameter file shown in Figure 2A.

#### **New user cohort selection**

- a. Identify the first ever drug era(s) of each patient drug, include those started during the study period
- b. Identify the drug eras in a which start after the washout period from observation start date.
- c. For non-switchers (a switcher is someone who switches between a target drug and a comparator drug), index\_start\_date is defined as the drug\_era\_start\_date and index\_end\_date as the earlier of the drug\_era\_end\_date and study end date.
- d. For switchers, both drug eras were excluded if target drug era and comparator drug\_era started on the same day (effectively such switchers are dropped from subsequent analysis). Otherwise, the first drug era is kept: index\_start\_date is defined as the drug\_era\_start\_date, and index\_end\_date is the earliest of the drug\_era\_end\_date, the switch date(the drug\_era\_start\_date of the drug switched to) and the study\_end\_date;

#### **Prevalent user cohort selection**

- a. Identify patients who switched drug during the study period (took both the target drug and the comparator drug)
- b. For non-switchers, all drug eras started during the study period were included: index\_start\_date is defined as the drug\_era\_start\_date, index\_end\_date as the earlier of drug\_era\_end\_date and the study end date;
- c. For switchers, only drug eras started during the study period but before the switch date are included: index\_start\_date is defined as the drug\_era\_start\_date, index\_end\_date as the earlier of drug\_era\_end\_date and the switch date.

The resulting data set of cohort of either design will have the format shown below:

CN	PID	ISD	IED	OSD	OED
P	43	3/10/10	4/11/10	1/29/08	12/21/10
P	35	9/1/11	11/21/11	7/31/10	3/24/12
P	71	4/13/11	6/24/11	7/25/10	10/14/11
P	23	8/8/11	8/25/11	3/14/11	4/27/12
P	9	1/31/10	2/9/10	10/21/08	3/16/10
P	55	6/19/09	6/25/09	4/22/08	12/8/09
P	74	8/22/10	10/18/10	4/3/09	4/20/11
P	39	5/21/10	6/10/10	11/6/07	9/27/10
P	44	4/6/09	5/8/09	1/1/08	1/8/10
P	9	4/7/10	6/6/10	2/18/08	6/14/10
P	3	4/24/10	5/22/10	3/20/09	6/6/10
P	90	7/25/09	8/20/09	10/30/07	4/30/10
P	60	8/6/10	9/23/10	6/9/09	12/3/10
P	87	5/28/11	7/1/11	2/24/10	7/23/11
P	23	8/30/10	9/10/10	5/6/10	12/13/10
P	15	4/29/11	5/29/11	11/13/09	12/5/11
P	2	8/27/09	8/31/09	1/5/07	9/5/09
P	83	12/15/10	1/13/11	12/2/08	5/12/11

CN = cohort name
PID = patient id
ISD = index start date
IED = index end date
OSD = observation start date
OED = observation end date

Figure 1: SAS dataset of cohort.

The data set of new user design have only one line for a given patient, and the data set of prevalent design may have multiple lines of records for a given patient.

#### 4.2 Eligibility criteria, analysis covariates and outcomes

Eligibility criteria, analysis covariates and outcomes are defined in a parameter file, see Figure 2A for an example. Lines are numbered for easy understanding of the file format but they are not present for the actual program parameter file.

Assumptions used in this parameter file:

1. If data\_table is NULL, operation is on APEX\_PERSONS\_FULL in Figure 4
2. Cov\_catgy:
  - a. if S (for primary selection to Set 1), 'OR' logic among Ss
  - b. if E (for eligibility), 'AND' logic among Es
  - c. if A (for analysis), concatenate with sep=' ' and put into a string; that will be the right side of model (default) unless otherwise specified
  - d. if blank, those are intermediate variables
3. Occ\_type: '0', '30', 'inpatient', 'outpatient'

In Figure 2A, the first line of the parameter file lists the parameter names. Lines 2 and 3 of the parameter file in Figure 2A define two drug cohorts, patients who were exposed to sitagliptin, and patients who were exposed to the second generation of sulfonylureas (glimepiride, glipizide and glyburide).

It is possible to have additional eligibility criteria. All cohort eligibility criteria correspond to lines ends in 'E' (cov\_catgy=' E ') in the parameter file, e.g., the 9<sup>th</sup> line of the parameter file in Figure. Since the eligibility criteria may involve certain

covariates, the eligibility criteria will be applied in the last step in forming the analysis data-set.

Analysis covariates are in general 'baseline' covariates, meaning that they are defined in some pre-specified time interval relative to the index of exposure. Analysis covariates (cov\_catgy='' or 'A', where null corresponds to an intermediate variable not in the analysis and 'A' corresponds to an analysis variable) defined prior to the index date of exposure (note this is different from the index date of outcome) can be specified by supplying values to the following parameters, concept\_id, data\_table, occ\_type, min\_days\_to\_index, max\_days\_to\_index and rule. Lines 4-6 in Figure 2A shows an example of defining copd related variables. Parameter 'cov\_group' helps programming efficiency by passing through data once, rather than three times for 'copd' related variables. Also we need to input only once concept ids related to copd in the parameter file. Here 'copdlyr' is an intermediate covariate (not used in the analysis but is useful for computing covariates that will be used in the analysis) defined as whether a patient had any of the condition concept ids in the year prior to the index exposure date. The two copd analysis covariates are defined in the next two lines, copd resulting in inpatient stays within 30 days of index exposure and copd resulting in inpatient stays beyond 30 days but within 365 days of index exposure or copd diagnosed from an outpatient visit. Rules in the field of 'rule' are valid SAS statements and will be copied verbatim into the program.

```
1. cov_group, cov_name,
   cov_type, concept_id, data_table, occ_type, min_days_to_index, max_days_to_index, rule, cov_
   catgy
2. ,sitagliptin, occurrence, 1580747, drug_era, 30, , , , S
3. ,sulfony2, occurrence, 1560171 1597756 1559684, drug_era, 30, , , , S
4. copd, copdlyr, occurrence, 255573 255841 257004 257905 258780 259043 261325 261889
   265304 448478 448908, condition_occurrence, , -365, 0, ,
5. copd, copd30d, occurrence, , condition_occurrence, inpatient, -30, 0, , A
6. copd, copd_not_30d, occurrence, , , , , , copdlyr-copd30d, A
7. ,age_ge_18, occurrence, , , , , , age>=18, E
```

**Figure 2A Parameters to APEX**

The resulting data set of covariates will be in the 'long and narrow' format shown below

CN	PID	ISD	CVN	CVV
P	43	3/10/10	C2	1
P	43	3/10/10	C3	1
P	43	3/10/10	C4	1
P	43	3/10/10	C5	1
P	35	9/1/11	C1	1
P	35	9/1/11	C2	1
P	35	9/1/11	C3	1
P	35	9/1/11	C4	1
P	35	9/1/11	C5	1
P	71	4/13/11	C1	1
P	71	4/13/11	C2	1
P	71	4/13/11	C5	1

CN = cohort name
PID = patient id
ISD = index start date
CVN = covariate name
CVV = covariate value

Figure 3: Long and narrow SAS dataset of covariates.

The macro for generating outcomes (y variables) has a parameter file in an almost identical format as the covariate macro, except it has two additional fields 'pw' as persistence window in days for bridging if needed and 'sw' for the definition of surveillance window in days. The parameter 'sw' allows both negative and positive values. Specifically,

- If negative, it means that the outcome is sought within 'sw' days from the index exposure (not including index\_start\_date);
- If positive but less than 9999, it means that the outcome is sought in the interval(s) of index\_start\_date and index\_end\_date + sw. For example, 30 or 60 days. Again, not including index\_start\_date.
- If 9999, it means that the outcome is sought post the index\_start\_date.

Types of outcome variable are specified in field 'cov\_type', which can be 'occurrence', 'quantity\_value' and 'survival'. These categories are explained in the examples in Figure 2B.

Figure 2B gives examples for a few definitions of outcome variables related to acute myocardial infarction (410.x0 and 410.x1). Again, lines are numbered for easy understanding of the file format but they are not present for the actual program parameter file.

```

cov_group, cov_name,
cov_type,concept_id,data_table,occ_type,pw,sw,min_days_to_index,max_days_to_index,rule
1. ,ami,occurrence,312327 326261 326308 326411 326448 434376 436706 438170 438438 438447
   441579,condition_era,30,,30,,,
2. ,ami_cnt,quantity_value,312327 326261 326308 326411 326448 434376 436706 438170 438438
   438447 441579,condition_occurrence,inpatient,30,30,,,
3. ,time2amil,survival,312327 326261 326308 326411 326448 434376 436706 438170 438438 438447
   441579,condition_occurrence,inpatient,,,,,
4. ,time2death,survival,444562 446371 444592,condition_era,30,,,,,
5. ,time2ami,survival,,,,,,min(time2amil, time2death)
6. ,time2ami_event,survival,312327 326261 326308 326411 326448 434376 436706 438170 438438
   438447 441579,condition_occurrence,inpatient,,,,,

```

**Figure 2B Parameters to APEX for outcome**

Line 1 defines an outcome variable 'ami' as type 'occurrence', that is, presence or absence of 'ami' as defined by the occurrence of any of the concept ids corresponding to ICD9s 410.x0 and 410.x1. Since the outcome is sought in the 'condition\_era' table, no bridging is needed. The surveillance window is 30 days, which means, the outcome is sought in the drug era(s) with the era end date(s) extended by 30 days.

Line 2 defines outcome variable 'ami' as type 'quantity\_value', that is, number of 'ami' episodes as defined by the concept ids corresponding to ICD9s 410.x0 and 410.x1. When an outcome is defined by multiple condition\_concept\_ids, bridging is needed to count *distinct* episodes of the outcome. The persistence window used in bridging is specified by 'pw' (30 days in this example). The surveillance window is 30 days, which means, the outcome is sought in the drug era(s) with the era end date(s) extended by 30 days.

Lines 3 to 6 define a time-to-event type of outcome, 'time2ami'. Line 3 instructs to calculate 'time2ami1', which is time since the index exposure to the earliest of dates of events specified in the concept ids and the index\_end\_date. Line 4 instructs to calculate 'time2death', which is time since the index exposure to the earliest of dates of events specified in the concept ids and the index\_end\_date. Line 6 sets the censoring flag for the time-to-event variable (1=presence of event) as time2ami\_event = 1 if min(dates of events as specified by the concept ids) is not missing and <= index\_end\_date, otherwise 0.

**APEX\_PERSONS\_FULL** as shown in **Figure 4** can be formed by transposing the covariate dataset and merging with the dataset of cohort. It is possible that a subset of covariates are absent from the covariate dataset, for example if none of the patients in the cohort had been hospitalized due to COPD within 30 days of index exposure, then the covariate 'copd30d' is not in the covariate dataset. We only store a specified covariate in the covariate dataset if there is at least 1 patient with nonzero value of this covariate. At the stage of merging, we will set 0 to those covariates absent from the covariate dataset but in the list of analysis covariates in the parameter file. *The merged analysis data set will have ALL analysis covariates specified in the parameter file.*

The final analysis dataset can be obtained by implementing the eligibility criteria specified in the parameters to APEX.

CN	PID	ISD	IED	OSD	OED	C1	C2	C3	C4	C5
P	43	3/10/10	4/11/10	1/29/08	12/21/10	0	1	1	1	1
P	35	9/1/11	11/21/11	7/31/10	3/24/12	1	1	1	1	1
P	71	4/13/11	6/24/11	7/25/10	10/14/11	1	1	0	0	1
P	23	8/8/11	8/25/11	3/14/11	4/27/12	1	1	0	1	1
P	9	1/31/10	2/9/10	10/21/08	3/16/10	1	1	1	1	1
P	55	6/19/09	6/25/09	4/22/08	12/8/09	1	0	1	1	1
P	74	8/22/10	10/18/10	4/3/09	4/20/11	1	1	1	1	1
P	39	5/21/10	6/10/10	11/6/07	9/27/10	1	1	1	1	1
P	44	4/6/09	5/8/09	1/1/08	1/8/10	1	0	1	1	1
P	9	4/7/10	6/6/10	2/18/08	6/14/10	1	1	1	1	1
P	3	4/24/10	5/22/10	3/20/09	6/6/10	1	1	1	1	1
P	90	7/25/09	8/20/09	10/30/07	4/30/10	1	0	1	1	1
P	60	8/6/10	9/23/10	6/9/09	12/3/10	1	1	1	1	0
P	87	5/28/11	7/1/11	2/24/10	7/23/11	1	1	1	1	1
P	23	8/30/10	9/10/10	5/6/10	12/13/10	1	1	0	1	0
P	15	4/29/11	5/29/11	11/13/09	12/5/11	1	1	1	1	1
P	2	8/27/09	8/31/09	1/5/07	9/5/09	1	1	1	1	1
P	83	12/15/10	1/13/11	12/2/08	5/12/11	1	1	1	1	0

Figure 4: APEX\_PERSONS\_FULLL.

#### 4.3 Analysis: options for plugging in various statistical methods to allow group comparisons with user-specified confounding adjustment.

The following analytical options/parameters are available:

PS\_USE=|1|2

- If not provided, default to conventional general/generalized multivariable regression, no ps estimated; otherwise
- 1=PS matching (1:k, k fixed, caliper)
- 2=PS stratification (ps\_strat\_k strata)

PSM=|formula, formula for propensity score estimation in the SAS format with LHS and RHS.

PS\_CLASS\_VARS, variables that appear in the class statement in the logistic regression

PS\_TRIM, numeric, e.g., 0.05, trim propensity score at 5th and 95th percentiles

NCONTROLS, the number of controls for each treated

SCALE, the scale to match PS on, whether propensity score itself (ps) or the log odds (logit\_ps)

CALIPER, numeric, e.g., 0.25, which means matching will be done within +/- 0.25\*SD of the propensity score on the chosen scale

EXACT\_MATCH\_VARS, list of variables that require matching beyond PS match, e.g., year and calendar of cohort entry, index\_qtr and index\_year

PS\_STRAT\_K, integer, number of strata of propensity score, e.g., 5, 10.

ORM=|formula, formula for outcome regression (exposure effect estimation) in the SAS format with LHS and RHS

ORM\_CLASS\_VARS: variables that appear in the class statement in the OUTCOME regression

STRAT\_VARS, stratification variable list in the outcome regression in the strata statement

BY\_SUBGROUP, if the outcome analysis needs to be done by subgroups, e.g., two separate analyses for patients with prior CVD present or absent. Any variable in the BY\_SUBGROUP var\_list should NOT appear in the STRAT\_VARS.

```
%APEX_analysis(analysis_set=, ps_use=|1|2,
                /* ps-estimation params */
                PSM=|formula,PS_CLASS_VARS=|var_list,
                PS_TRIM=,
                /* ps-match params */
                ncontrols=|k, scale=ps|logit_ps, caliper=|x,
                exact_match_vars=|varnamelist,
                /* ps-stratification params */
                ps_strat_k=|,
                /*outcome analysis parms*/
                ORM=|formula,
                /*stratification variables in outcome regression */
                strat_vars=|var_list,
                /*by_subgroup*/
                by_subgroup=|var_list)

/* example, 1:1 match */
%APEX_analysis(analysis_set, ps_use=1,
                /* ps-estimation params */
                PSM=cn(event='sitagliptin')=(copd30d copd_not_30d dementia
                depression cancer hx_fracture30d hx_fracture_not_30d)*prior_cvd
                ,PS_CLASS_VARS=,
                PS_TRIM=,
                /* ps-match params */
                ncontrols=1, scale=logit_ps, caliper=0.25,
                exact_match_vars=index_year index_qtr,
                /* ps-stratification params */
                ps_strat_k=,
                /*outcome analysis parms*/
                ORM=surv*surv_flag(1)=cn,
                orm_CLASS_VARS=cn(ref='sulfony2'),
                /*stratification variables in outcome regression*/
                strat_vars=index_year index_qtr prior_cvd,
                /*by_subgroup*/
                by_subgroup=)
```

```

/* the above example, can omit "arg=" if no value to pass */

%APEX_analysis(analysis_set, ps_use=1,
               /* ps-estimation params */
               PSM=cn(event='sitagliptin')=(copd30d copd_not_30d dementia
depression cancer hx_fracture30d hx_fracture_not_30d)*prior_cvd ,
               /* ps-match params */
               ncontrols=1, scale=logit_ps, caliper=0.25,
               exact_match_vars=index_year index_qtr,
               /*outcome analysis parms*/
               ORM=surv*surv_flag(1)=cn,
               orm_CLASS_VARS=cn(ref='sulfony2'),
               /*stratification variables in outcome regression*/
               strat_vars=index_year index_qtr prior_cvd
               )

/* example, 1:1 match but instead of one risk estimate, calculate
one for each subgroup */
%APEX_analysis(analysis_set, ps_use=1,
               /* ps-estimation params */
               PSM=cn(event='sitagliptin')=(copd30d copd_not_30d dementia
depression cancer hx_fracture30d hx_fracture_not_30d)*prior_cvd
               ,PS_CLASS_VARS=,
               /* ps-match params */
               ncontrols=1, scale=logit_ps, caliper=0.25,
               exact_match_vars=index_year index_qtr,
               /* ps-stratification params */
               ps_strat_k=,
               /*outcome analysis parms*/
               ORM=surv*surv_flag(1)=cn,
               orm_CLASS_VARS=cn(ref='sulfony2'),
               /*stratification variables in outcome regression*/
               strat_vars=index_year index_qtr,
               /*by_subgroup*/
               by_subgroup= prior_cvd)

```

4.4 Output: SAS data sets of the following

- The analysis dataset after cohort selection and covariate calculation is saved in a user specified 'work' folder for further analyses using the analysis module, or custom programs.
- Risk estimates and standard errors for each combination of parameter settings. If stratified analysis is required, there will be an additional column of 'stratum\_name', in which case if the 'stratum\_name' is blank, the row is for the whole analysis set.

Effect_estimate	SE

- Patient summaries
  - Summary of patient selection, which present how the numbers of patients selected are impacted by each selection criteria. The mock-up table appears as below,

**APEX\_CRITERIA\_SUMMARY**

N in Index	18		
N meeting all criteria	9		
Covariate name	N with criterion	% of INDEX	n gained by excluding covariate
C1	17	94%	1
C2	15	83%	3
C3	15	83%	1
C4	17	94%	0
C5	15	83%	2

where

- 'N in index' is the number of patients after primary selection (cov\_catgy="S"),
  - C1-C5 are eligibility criteria (cov\_catgy="E"),
  - 'N meeting all criteria' is the number of patients satisfying C1 & C2 & C3 & C4 & C5,
  - For each row i, 'N with criterion' is the number of patients satisfying this criterion alone, '% of INDEX' is the ratio of 'N with criterion' and 'N in Index', 'n gained by excluding covariate' is the difference of 'N in index' and the number of patients with all criteria but the ith criterion.
- ii) Patient characteristics, such as Table 1 in a typical epidemiology paper. Variables are summarized by their attributes, binary, continuous and categorical as shown below. Such tables will be done for
    - The whole analysis data set APEX\_PERSONS\_FULL, and **additionally**
    - The matched data set, **if** matching is employed in the analysis
    - Each stratum, **if** stratified analysis is required.

**APEX\_COVARIATE\_SUMMARY**

Cohort name	P					
N in Cohort						
<b>Binary covariates</b>						
Covariate name	N	% of cohort				
OBESITY						
HYPERTENSION						
SMOKING						
OP						
ANTIDIAB						
<b>Continuous covariates</b>						
		Among patients with >=1				
Covariate name	N with >=1	Min	Median	Max	Mean	Std Dev
AGE						
HOSP30D						
HOSP1YR						
OUTP1YR						
DRUGS1YR						
<b>Categorical covariates</b>						
Covariate name	Category name	N	% of cohort			
GENDER	Male					
	Female					
	Unknown					
AGE_GROUP	18-40					
	40-65					
	>65					

- iii) Summary of number of outcomes and exposure among cohorts. There may be an additional column of 'stratum name' if stratified analysis is required, in which case rows with this field blank are for the entire cohorts.

	# w outcome	Total persons	Time-at-risk
Cohort 1			
Cohort 2			