

eElectronic Appendix 1:

Output of a screening tool for close correlates of treatment choice that are not related to the study outcome (coxib example).

Nsaid - Cox. Outcome = GI complication
K500 N200 Dx3 Generic 12:31 Wednesday, June 25, 2008

Dimension	code	frequency_ type	rr_ce	rr_cd	rr_ce_ rankStars
x3_ambdx	274	once	0.57413	1.33750	1.0 ***
dx3_mddx	274	once	0.59632	1.55786	2.0 *
dx3_hospdx	618	once	0.68079	1.29027	3.0 ***
dx3_ambdx	618	once	0.68825	1.20109	4.0 ****
dx3_ambdx	162	sporadic	0.75304	1.23305	6.0 ****
generic_drugs	warfarin sodium	frequent	1.29333	1.33805	12.0 ***
dx3_nhdx	715	frequent	1.28900	1.34443	14.0 ***
prcdr_mdproc	85610	sporadic	1.28382	1.56258	16.0 *
prcdr_ambproc	863	once	0.77982	1.27691	17.0 ****
generic_drugs	warfarin sodium	sporadic	1.28004	1.34342	20.0 ***
dx3_mddx	714	frequent	1.27896	1.05056	21.0 *****
dx3_ambdx	714	frequent	1.27299	1.46914	23.0 **
generic_drugs	warfarin sodium	once	1.26605	1.39066	25.0 **
dx3_mddx	714	sporadic	1.26183	1.14412	27.0 *****
dx3_ambdx	162	once	0.79712	1.14912	31.0 ****
dx3_mddx	725	once	1.24455	1.15490	35.0 ****
generic_drugs	olanzapine	frequent	0.80962	1.46871	38.0 **
generic_drugs	folic acid	frequent	1.23513	1.52650	39.0 *
prcdr_mdproc	85610	once	1.22844	1.38142	44.0 ***
dx3_hospdx	332	frequent	1.22719	1.24977	45.0 ****
generic_drugs	tramadol hcl	sporadic	1.22716	1.46824	46.0 **
dx3_nhdx	427	frequent	1.22433	1.58125	50.0 *
dx3_mddx	714	once	1.22422	1.22965	51.0 ****
generic_drugs	calcitonin,salmon,synthetic	sporadic	1.22363	1.22610	52.0 ****
dx3_nhdx	V43	once	1.22324	1.21224	53.5 ****
generic_drugs	calcitonin,salmon,synthetic	frequent	1.22253	1.50926	55.0 *
dx3_mddx	715	once	1.22232	1.06860	56.0 *****
dx3_nhdx	414	frequent	1.22148	1.07309	58.0 *****
dx3_nhdx	427	once	1.22083	1.58953	59.0 *
prcdr_hospproc	8154	once	1.21911	1.45755	60.0 **
generic_drugs	lansoprazole	frequent	1.21677	1.51021	62.0 *
dx3_nhdx	724	once	1.21209	1.63779	64.0 *
generic_drugs	lansoprazole	sporadic	1.21093	1.49089	66.0 **
generic_drugs	calcitonin,salmon,synthetic	once	1.20865	1.60861	69.0 *
generic_drugs	tobramycin sulfate/dexamethasone	frequent	0.82740	1.13450	70.0 *****
dx3_ambdx	V58	frequent	1.20771	1.13084	72.0 *****
prcdr_hospproc	4525	once	1.20720	1.28605	73.0 ***
dx3_nhdx	715	once	1.20659	1.50653	74.0 **
dx3_ambdx	847	once	0.82932	1.29134	75.0 **
dx3_mddx	715	frequent	1.20547	1.16486	76.0 ****
dx3_ambdx	188	once	0.83116	1.46430	78.0 **
prcdr_mdproc	20610	once	1.20150	1.38682	80.0 ***
dx3_ambdx	427	frequent	1.20018	1.28203	82.0 **
dx3_nhdx	294	once	1.19915	1.42409	87.0 **
dx3_nhdx	V57	once	1.19865	1.40867	88.0 **
dx3_ambdx	845	once	0.83487	1.56913	91.0 *
dx3_ambdx	714	once	1.19773	1.11960	92.0 *****
dx3_mddx	E92	once	0.83819	1.02634	96.0 *****
dx3_nhdx	820	frequent	1.19120	1.05624	98.0 *****

Dimension = Data dimension

Code = Identified code or drug

Frequ. type = once/sporadic/frequent, as described in the text

rr_ce = association between binary confounder and binary exposure (relative risk)
rr_cd = association between binary confounder and binary outcome (relative risk)
rr_ce_rank = Rank of rr_ce; strongest association (farthest from null, as judged by magnitude of the log of the absolute value of the risk ratio) is ranked 1 (highest)
Stars = Indication of strength of rr_cd, where more stars indicates a weaker association (closer to the null). Stars are calculated by computing decile of rr_cd's distance from null and then identifying the bottom 5 deciles. The decile containing covariates closest to the null will receive 5 stars.

eElectronic Appendix 2:

Documentation for the hd-PS SAS macro.

The hd-PS SAS macro can be downloaded at www.drugapi.org ... links ... downloads.

High Dimensional Propensity Score Macro

Usage Instructions

Macro Version 1.0

SYNTAX

```
%RunHighDimPropScore( parameters );
```

Description

This macro implements a multi-step algorithm to implement high-dimensional proxy adjustment in claims data. The algorithm is described in detail in *High-dimensional proxy adjustment in claims data studies of treatment effects* by Schneeweiss et al. Briefly, the macro's steps include 1) identify data dimensions, e.g. diagnoses, procedures, and medications, 2) empirically identify candidate covariates, 3) assess recurrence, i.e. multiplicity of same code, 4) prioritize covariates, 5) select covariates for adjustment, 6) estimate exposure propensity score, and 7) estimate outcome model.

MACRO DEFINITION FILE

hdps.mcr

Parameters

`var_patient_id` *required*

Name of unique patient identifier variable.

`var_exposure` *optional; default=exposure*

Name of binary exposure variable. A value of 1 for this variable will be considered exposed; 0 will be considered unexposed/referent.

`var_outcome` *optional; default=outcome*

Name of binary outcome variable. A value of 1 for this variable will be considered outcome positive; 0 will be considered outcome negative.

`vars_demographic`

optional

A list of variables to be considered as demographic information, such as age, sex, and race.

vars_predefined

optional

A list of variables to be considered as pre-defined, such as co-morbidities or past drug usage. **If this option is not specified, all variables in the dataset other than the patient ID, exposure, outcome, and demographic variables will be considered pre-defined.**

vars_force_categorical

optional

An optional list of numeric variables that should be treated as categorical. Year of service is a common example. (See note below.)

vars_ignore

optional

A list of variables to be ignored in the propensity score estimation and simply passed through to output data sets. Examples include follow-up time and other variables important to an analysis but which do not figure into confounder adjustment.

top_n

optional; default=200

The *n* most prevalent empirical covariates to consider from each dimension of data.

k

optional; default=500

The number of empirical covariates to include in the resulting propensity score.

trim_mode

optional; default=BOTH

A flag that controls whether to perform trimmed or untrimmed propensity score analyses. If set to Y or y, the macro will perform only trimmed analyses; if set to N or n, the macro will perform only untrimmed analyses. If left unspecified or specified as BOTH, the macro will perform both trimmed and untrimmed analyses.

percent_trim

optional; default=5

If trim_mode is Y or BOTH, the percentage to trim from the propensity score.

input_cohort

required

Name of the input SAS dataset containing the patients, exposure, outcome, and any fixed covariates. Datasets will not be modified. See note below on format of the data.

input_dimX

one or more dimensions required

This parameter is specified one or more times to indicate the names of the datasets containing the dimension data. Datasets will not be modified. See note below for dataset format.

Each time this parameter is specified, two data items are needed: the name of the dimension dataset and the name of the field containing the code of interest. These two items should be supplied on one line with only spaces in between. (See example.)

The parameters should be numbered sequentially and with no gaps in the numbering (input_dim1, input_dim2, etc.).

output_scored_cohort

optional; default=output_scored_cohort

Name of a SAS dataset that will be created or replaced. This dataset will contain the input cohort (including their exposure, outcome, and fixed covariates), as well as the propensity score calculated. It is intended for use in outcome models for further exploration of the outcome of interest.

output_detailed

optional; default=output_detailed

Name of a SAS dataset that will be created or replaced. This dataset will contain the input cohort (including their exposure, outcome, and fixed covariates), as well as the series of empirical covariates calculated from each dimension of data. It is intended for diagnostic use or further exploration of the generated variables and may be quite large.

results_diagnostic

optional; default=result_diagnostic

Name of a SAS dataset that will be created or replaced. This dataset will contain each of the empirical covariates considered, along with measures of prevalence and potential bias for each empirical covariate. It is intended for diagnostic use or further exploration of the generated variables and serves as a “dictionary” for the variables in output_detailed.

results_estimates

optional; default=result_estimates

Name of a SAS dataset that will be created or replaced. This dataset will contain the point estimates and confidence intervals for each of the outcome models automatically run. If trimmed results have been requested, a second dataset with the extension `_trim` will also be created.

Notes

Input cohort data format. The `input_cohort` dataset must contain only the patient ID, exposure, outcome, and any covariates that should be adjusted for in the outcome model. If the `vars_predefined` option is not specified, covariates in the input data that are not to be included in the outcome model should be dropped before running the macro.

Input dimension data format. Each dimension of data (drugs, inpatient procedures, etc.) should be supplied in a unique dataset. This dataset should have at least two fields: the patient identifier and the dimension code. Patients receiving multiple codes, and/or a single code more than one time, should be listed over multiple rows. (For example, in an outpatient procedure dimension, the dimension code would likely be a CPT code.) Additional fields will be ignored.

Covariates. All character covariates in the input cohort will be treated as categorical. Any numeric covariates that should be treated as categorical (such as year of service) should be either converted to character before running the macro or included in the `vars_force_categorical` list.

EXAMPLE CODE

```
%include "/path/to/macro/directory/hdps.mcr";

Title1 'High-dimensional propensity score adjustment';
Title2 '(study description)';

%RunHighDimPropScore (
    var_patient_id           = id,
    var_exposure             = exposed,
    var_outcome              = outcome,
    vars_demographic         = age sex race,
    vars_force_categorical   = year,
    vars_ignore              = followup_time,
    top_n                    = 200,
    k                        = 200,
    trim_mode                = BOTH,
    percent_trim             = 5,
    input_cohort             = master_file,
    input_dim1               = drug_claims          generic_name,
    input_dim2               = outpatient_diagnoses icd9_dx,
    input_dim3               = inpatient_diagnoses icd9_dx,
    input_dim4               = inpatient_procedures icd9_proc,
    input_dim5               = outpatient_procedures cpt,
    output_scored_cohort     = scored_cohort,
    output_detailed          = detailed_cohort,
    results_estimates        = estimates,
    results_diagnostic       = variable_info
);
```