

The SAS %RRC Macro

Xiaomei Liao, Samuela Pollack and Donna Spiegelman

May 24, 2018

Abstract:

The macro %RRC uses the risk set regression calibration (RRC) method to correct the point and interval estimate of the relative risk in the Cox proportional hazard regression model for bias due to measurement error in one or more baseline or time-varying exposures, including time-varying variables that are functions of the exposure history such as the 12-month moving average exposure, cumulative average exposure, cumulative total exposure, etc. An external and internal validation study designs are available to use this macro. Technical details are given in Liao et al. (2011) and Liao et al. (2018).

Keywords: SAS, macro, measurement error, Cox proportional hazard model, time-varying covariates, a-b time unit moving average exposure (e.g. 12-month moving average), cumulatively updated exposure, simple updated exposure.

Contents

- 1 Description
- 2 Invocation and details
- 3 Examples
- 4 Warning

1. Description

The RRC method for the Cox proportional hazard model is designed to use under the following circumstances:

1. Five types of time-varying exposure metrics are supported:
 - 1) Cumulatively average updated exposure (`type=1`);
 - 2) a-b time unit ($a \leq b$) moving average exposure (`type=2`) (e.g. $a=1$, $b=12$ means 12-month moving average exposure);
 - 3) Cumulatively total updated exposure (`type=3`);
 - 4) Simple updated exposure (`type=4`);
 - 5) Baseline time-independent exposure (`type=5`).
2. Two kinds of measurement error models are supported for cumulatively average updated exposure (`type=1`): general linear error model (`emodel=1`) and classical additive error model (`emodel=2`). The default model is general linear error model.
3. Outcome is time-to-event (left truncation is supported).

4. All model covariates measured with error are continuous (perfectly measured covariates may be either categorical or continuous).
5. For all types of exposure, please provide the individual exposure measurements at each time unit considered (e.g. each questionnaire cycle, or, each month, et al.). The program will create the corresponding types of time-varying exposure metrics internally according to the “type” you choose.
6. The validation study dataset can’t contain any missing data. For type=1, 3, 4, 5, the main study dataset can’t contain any missing data as well. If you have missing data in your data sets, depending on your preference, you may delete them, or use the Last Value Carry Forward (LVCF) to fill in the values. However, for type=2, the missing values are allowed at the beginning of the follow-up, and the program will calculate a-b time unit moving average exposure using the formula $x(t_s) = \frac{\sum_{t=t_s-b+1}^{t_s-a+1} c(t)I(t)}{\sum_{t=t_s-b+1}^{t_s-a+1} I(t)}$, where $I(t)$ is the indicator function for missing, i.e. $I(t) = 0$ if missing, otherwise $I(t) = 1$; $c(t)$ is the individual exposure measurements at each time unit; $x(t_s)$ is the a-b time unit moving average exposure at time t_s .
7. For the internal validation study design, true exposure values can be used if available. For that, the user needs to create a variable which will indicate the corresponding true observations and declare it in the “useind=” statement.

2. Invocation and details

You need to prepare a SAS data set for the main study and another data set for external validation study. Please specify the variable name for “id=”, which is the subject identification for both the main study and the validation study. Both the main study and validation study must have the same variable names for covariates with measurement error (for “surrogate=”). The main study data set must have a variable that defines the occurrence of the event (for “case=”). It is optional to have one or more perfectly measured covariates in both main and validation datasets (for “confounder=”). The same confounders with the same variable names must be available in both the main study and the validation study. The parameter “time=” is the time-to-event variable at which the covariates are measured till the end of follow-up (i.e. got an event or censored). The main study data is in counting process structure (i.e. one person per multiple records indexed by “time” variable, please see the example for the data structure).

The %rrc macro works for five types of exposures. If the exposure of interest is the cumulatively average updated exposure, please set parameter Type=1. In this scenario, you need to provide the individual exposure measurements at each questionnaire cycle; the program will create the cumulatively updated average internally. For fitting the general linear measurement error, i.e. $c = \alpha_0 + \alpha_1 C + e$, where c is the true exposure and C is the surrogate, please set Type=1, emodel=1. For fitting the classical measurement error model, i.e. $C = c + e$, which is not a realistic model for most data in nutritional and environmental epidemiology though, please set Type=1, emodel=2. In this case, you need to provide a value for parameter icc (i.e. ρ_I with $0 < \rho_I < 1$), which is the correlation of the repeated measurements of the surrogate exposure; and the value for parameter dampfactor (i.e. θ with $0 < \theta < 1$), which is the dampening factor of the covariance matrix for surrogate exposure, where we assume that the correlation between repeated measures

on the same exposure within a study participant, C_{ij} and $C_{ij'}$, is given by $\text{corr}(C_{ij}, C_{ij'}) = \rho_1^{\theta|j-j'|}$. If $\theta = 0$, the covariance matrix has the compound symmetry structure; if $\theta = 1$, the covariance matrix has AR(1) structure; if $0 < \theta < 1$, the covariance matrix has the damped exponential structure (DEX) (Alvaro Munoz, V. C., et al. *Biometrics*, 48, 733-742, Sep, 1992). The `Type=1, emodel=2` scenario works for the exposure only model without covariates.

If the exposure of interest is a-b time unit moving average exposure, such as 12-month moving average exposure, please set parameter `Type=2`. This scenario requires the user to provide the additional “b” time units data of follow-up before the baseline, so that a-b time unit moving average exposure can be calculated for the first “a-b” time units after the baseline. The parameter “`month1`” is used to set the value of a. The parameter “`month`” is used to set the value of b. In the main study dataset, the “`period`” variable is used to indicate the time period, which has the values of -b+1, -b+2, ..., 0 in the additional “b” time units follow-up before the baseline; and has the values of 1, 2, ..., a, ..., b, b+1, b+2, ... starting from the baseline. It’s possible that the values of exposure are missing for some subjects at the beginning of -b+1 time unit (otherwise Last Value Carry Forward (LVCF) can be used to fill in the missing value when it’s in the middle of the follow-up), hence a missing indicator variable “`missing`” is used to skip the missing values when calculating the a-b time unit moving average using the formula $x(t_s) = \frac{\sum_{t=t_s-b+1}^{t_s-a+1} c(t)I(t)}{\sum_{t=t_s-b+1}^{t_s-a+1} I(t)}$, where $I(t)$ is the indicator function for missing, i.e. $I(t) = 0$ if missing, otherwise $I(t) = 1$; $c(t)$ is the individual exposure measurements at each time unit; $x(t_s)$ is the a-b time unit moving average exposure at time t_s . The value of the missing indicator variable “`missing`” is $1 - I(t)$.

If the exposure of interest is the cumulatively total updated exposure, please set parameter `Type=3`. This scenario is similar as the cumulatively average updated exposure, but we only consider the general linear measurement error here due to the more generality of this type of measurement error model.

If the exposure of interest is the simple updated exposure, please set parameter `Type=4`. In this scenario, the exposure of interest is each study participant’s most recent value of the exposure.

If the exposure of interest is the baseline exposure, which is time-independent, please set parameter `Type=5`. In this case, all model covariates must be baseline only, and the main study has one record per person, as does the validation study. You also need to specify the variable name for “`timestart=`”, which is the variable for the time (e.g. age) at the start of follow-up, and the variable name for “`timeend=`”, which is the variable for the time at the end of follow-up (i.e. got an event or censored). It means that you need to create these two variables in the main study, however, you don’t need to create the variable “`time`” in the main study as in options “`type=1, 2, 3, 4`”. Put the same name of “`time`” variable as in the validation study and the macro will generate this variable in the main study according to “`timestart`” and “`timeend`”. This way will take care of the left truncation in the analysis for time-independent exposure.

The macro also allows the using of non-conformal dimension of measurement error model and primary regression model. In details, in addition to the common confounders specified in the

parameter “confounder”, if some error-free covariates are only associated with the measurement error model, not associated with the outcome in the primary regression model, then they may be only included in the measurement error model, not in the primary regression model, which results in the non-conformal dimension of two models. These covariates are specified in the parameter “extravariablenval”. On the other hand, if some covariates are only associated with the outcome in the primary regression model, not associated with measurement error model, then they may be only included in the primary regression model, not in the measurement error model. These covariates are specified in the parameter “extravariablenmain”.

The macro includes three optional parameters “initialvalue”, “xupper”, “xlower”, which are respectively the initial values, upper bounds and lower bounds for the log(RR) estimates of all the exposure and covariates for the optimization process of the partial likelihood of Cox model. Users are recommended to supply these values for the better and faster convergence of the likelihood. If not provided, the default values for “initialvalue” are zero for all the exposure and covariates. The default values for “xlower” and “xupper” are -5 and 5 for all the exposure and covariates, respectively.

Although the residual method is included as an option in the macro for controlling for confounding with two or more perfectly measured covariates, it may not be a valid method in the context of time-varying covariates, hence it’s not suggested to use “residual=1”. You may increase the value of “groupnum” such that each risk set has more subjects for fitting the measurement error model, and thus stabilize the estimation results.

The required and optional arguments for %rrc are summarized as follows:

```
%macro rrc
(
id=,          /* Variable name of subject identification;
              Required */
main=,        /* The name of the main study SAS dataset;
              Required */
validation=,  /* The name of the validation study SAS dataset;
              Required */
surrogate=,   /* The exposure measured with error in main study and
              validation study; Required */
true=,        /* The exposure measured without error in validation study;
              Required */
case= ,       /* The variable that indicates whether an event occurs or
              not; Required */
time =,       /* The variable for time-to-event outcome;
              Required */
type=,        /* type of data,
              1: Cumulatively average updated exposure (time-varying);
              2: a-b time unit moving average exposure (time-varying);
              3: Cumulatively total updated exposure (time-varying);
              4: Simple updated exposure (time-varying);
              5: Baseline exposure only (time-independent);
              Required */
month=12,     /* Required only for type=2. The value of unit ``b`` in a-b
              time unit moving average (the default is 12, which is
              used for 12-month moving average). Time unit could be
```

```

        month or year or else.
    */
month1=1, /* Required only for type=2. The value of unit ``a'' in a-b
time unit moving average (the default is 1, which is
used for 12-month moving average). */
missing=, /* Required only for type=2. Missing indicator for main
exposure, it's 1 if missing, otherwise it's 0. */
period=, /* Required only for type=2. Period indicator for a-unit
moving average. This variable contains values of
-a+1,-a+2,...,0,1,2,...,a,a+1,a+2,...
*/
residual=0, /* if residual=1,use residual method for dimension
reduction of the models; otherwise, residual=0;
Required */
initialvalue=, /* the initial values for the log(RR) estimates of the
exposure and covariates. The input should be a vector
with numbers separated by space. Optional, default=0
for all */
xupper=, /* the upper bounds for the log(RR) estimates of all the
exposure and covariates. Optional, default=5 for all
*/
xlower=, /* the lower bounds for the log(RR) estimates of all the
exposure and covariates. Optional, default=-5 for all
*/
emodel=1, /*if emodel=1(default), it assumes general linear
measurement error model;
if emodel=2, it assumes classical additive measurement
error model. Only Required for type=1 */
confounder= , /* The perfectly measured covariates adjusted in the
model; Optional */
extravariabileinval=, /* extra variables included in the measurement
error model, NOT in the primary regression
model. */
extravariabileinmain=, /* extra variables included in the primary
regression model, NOT in the measurement error
model. */
groupnum=5, /* The minimum number of subjects in each risk set for
validation study; Optional, default=5*/
increments=1, /* The units in which the RR for true covariate will be
reported; Optional, default=1*/
icc=, /* the correlation of repeated measurements of surrogate
exposure; Only required for type=1, emodel=2 */
dampfactor=, /* the dampening factor of the covariance matrix for
surrogate exposure; Only required for type=1, emodel=2*/
timestart=, /* the start of follow-up time, for time-independent
exposure; Only Required for type=4 */
timeend=, /* the end of follow-up time(an event or censoring), for
time-independent exposure; Only Required for type=4 */
filename=RRCOutput.txt, /* filename for the output file; Optional,
default=RRCOutput.txt */
csvfilename=none, /* filename for CSV output. The default is no CSV
output if the users don't require it */
path = /tmp /* path of directory to put intermediate data files. The
default directory working on Channing is /tmp. */
useind=, /* whether to use the corrected xhat value or its original
value; 0=use original value, 1=use xhat. This feature is only for type
1 or 3. */
/* The following option works for Type=1,2,3 now */

```

```

truepredict=, /* Name of the variable for predicted true value */
surrogatefunc=, /* Name of the variable for the function (such as 12
                month) of surrogate exposure for output */
outdata= /* Name of the output data for the main study with both the
          surrogate and the predicted true exposure */
);

```

3. Examples

Example 1: Low-carbohydrate diet scores and risk of type II diabetes in the Health Professionals Follow-up Study. Three different analysis have been carried out using %rrc macro:

1(a). Analysis for cumulatively updated diet scores;

1(b). Analysis for simple updated diet scores;

1(c). Analysis for diet scores at baseline.

Reference: L. de Koning, T. T Fung, X. M. Liao, S. Chiuve, E. B. Rimm, W. C. Willett, D. Spiegelman, F. B. Hu, *American Journal of Clinical Nutrition*, 2011; 93(4): 844-50.

Three low-carbohydrate diet scores were discussed in this paper --- low-carbohydrate diet with high total protein and fat, low-carbohydrate diet with high animal protein and fat, and low-carbohydrate diet with high vegetable protein and fat. We will use the first one, low-carbohydrate diet with high total protein and fat to illustrate the use of %rrc macro.

1(a): Analysis for cumulatively updated diet scores.

First, we are interested in the cumulatively updated low-carbohydrate diet with high total protein and fat. The diet score has been calculated according to Halton TL, et al. *N Engl J Med*, 2006, for both the main study and the validation study. BMI is a potential confounder in the analysis.

The main study dataset can be read in as follows:

```

Data all;
infile
"/udd/stxia/RRCmacro/TestforChanning07312012/hpfsMain2_withBMI_po
int.dat";
input id ageyr allt2d ttlcarb bmi;
run;

```

And the validation dataset can be read in as follows:

```

Data vali;
infile
"/udd/stxia/RRCmacro/TestforChanning07312012/hpfsVal_dscore_bmi.d
at";
input id ageyr ttlcarbdr ttlcarb bmi;
run;

```

Note: In both the main study and the validation study, the input value for exposure of interest (diet score) is the individual diet score based on the FFQ at each questionnaire cycle or based on DR. Both the exposure and covariates could have some values missing and it's suggested to use "Last Value Carried Forward (LVCF)" to fill in the missing values before doing the analysis. The cumulatively updated exposure at time t is the average of the individual diet score up to time t .

For example, the following are some data records from the main study dataset:

Obs	id	ageyr	allt2d	ttlcarb	bmi
1	100005	73	0	4	22.6
2	100005	74	0	4	21.9
3	100005	76	0	8	22.4
4	100005	78	0	8	22.4
5	100005	80	0	8	23.2
6	100005	82	0	8	23.2
7	100005	84	0	8	23.2
8	100006	63	0	28	26.5
9	100006	65	0	28	24.8
10	100006	67	0	22	26.5
11	100006	70	0	22	26.5
12	100006	71	0	23	26.5
13	100006	73	0	23	25.8
14	100006	75	0	18	25.8
15	100006	77	0	18	25.1
16	100006	79	0	21	25.3
17	100006	81	0	21	25.8

Here are some data records from the validation study dataset:

Obs	id	ageyr	ttlcarbdr	ttlcarb	bmi
1	100965	67	23.0	27	22.3510
2	101463	60	5.5	11	21.6997
3	102750	43	12.0	13	22.2422
4	102781	43	6.0	0	21.8556
5	114258	61	15.0	17	27.9793
6	114408	59	17.5	10	25.5231
7	114722	58	20.0	17	25.8270
8	114776	56	9.5	15	26.4689
9	114974	53	7.5	20	30.5152
10	116073	73	4.5	5	22.8072
11	116217	69	15.0	12	27.2520
12	117761	74	14.0	4	23.4931
13	120133	61	18.0	20	23.3521
14	123647	67	11.5	12	28.6967
15	129563	72	20.0	24	26.4563

The general linear measurement error is an appropriate measurement error model for dietary exposures, so this measurement error model is always our first choice. We will first illustrate the use of the macro in the crude analysis, which adjusts only for age in years as the time scale. We call the macro as:

```
%rrc
(
id=id,
main=all,
validation=vali,
surrogate=tlcarb,
```

```

true=tlcarbdr,
confounder=, /* leave it blank */
case=allt2d,
time=ageyr,
groupnum=5,
increments=20, /* difference of 90th and 10th percentile */
type=1, /* For cumulatively updated exposure */
emodel=1, /* For the general linear measurement error */
timestart=, /* not needed here, optional */
timeend=, /* not needed here, optional */
residual=0,
filename=Outputgen_crude);

```

Then the text file "Outputgen_crude" showed:

```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
This analysis is for cumulatively average updated exposure
with general linear measurement error model.

```

The increment for each exposure is: 20.0

There is only one exposure in the analysis and no covariate.
The main exposure is measured with error.

```

# of subjects in the main study: 41539
# of person-year observations in the main study: 380337
# of cases in the main study: 2790
# of unique failure time: 53

```

The unique failure times are:

```

39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
89.0 90.0 92.0

```

```

# of subjects in the validation study: 127
# of person-year observations in the validation study: 127
# of risk sets in the validation study: 19

```

```

-----
Information about the measurement error model
fitting in the validation study:
-----

```

Risk_set	Age_low	Age_up	Person_year	Intercept	slope	R ²
1	39.0	40.0	9	0.425	0.496	0.195
2	41.0	42.0	12	0.427	0.471	0.414
3	43.0	43.0	5	0.367	0.395	0.713
4	44.0	44.0	6	0.394	0.491	0.601
5	45.0	46.0	7	0.515	0.772	0.464
6	47.0	48.0	5	0.228	0.598	0.789
7	49.0	50.0	5	0.718	0.179	0.101
8	51.0	53.0	6	0.839	-0.029	0.000
9	54.0	55.0	8	0.099	0.846	0.963
10	56.0	57.0	6	-0.139	0.959	0.876

11	58.0	59.0	11	0.635	0.314	0.320
12	60.0	61.0	7	0.321	0.479	0.218
13	62.0	62.0	5	0.831	-0.178	0.041
14	63.0	63.0	5	0.087	0.835	0.556
15	64.0	65.0	5	0.493	0.167	0.101
16	66.0	67.0	8	0.500	0.488	0.297
17	68.0	68.0	5	0.062	0.554	0.240
18	69.0	71.0	6	0.816	-0.093	0.007
19	72.0	92.0	6	0.352	0.444	0.585

Results for uncorrected estimates (naive cox)
and corrected estimates (RRC):

Uncorrected result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
ttlcarb	0.817	0.062	0.000	2.264	[2.01, 2.56]

RRC result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
ttlcarb	1.509	0.237	0.000	4.523	[2.84, 7.19]

%%

To control for confounding by BMI, we call the macro as follows:

```
%rrc
(
id=id,
main=all,
validation=vali,
surrogate=tlcarb,
true=tlcarbdr,
confounder=bmi, /* BMI is a confounder in this analysis*/
case=allt2d,
time=ageyr,
groupnum=40, /* Increase the size of groupnum to stabilize
the results */
increments=20, /* difference of 90th and 10th percentile */
type=1, /* For cumulatively updated exposure */
emodel=1, /* For the general linear measurement error */
timestart=, /* not needed here, optional */
timeend=, /* not needed here, optional */
residual=0,
filename=Outputgen_BMI);
```

Then the text file "Outputgen_BMI" showed:

%%

This analysis is for cumulatively average updated exposure
with general linear measurement error model.

The increment for each exposure is: 20.0

There is one exposure and one or more covariates in the analysis. The exposure is measured with error, but the covariates are error-free.

```
# of subjects in the main study: 41539
# of person-year observations in the main study: 380337
# of cases in the main study: 2790
# of unique failure time: 53
```

```
The unique failure times are:
39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
89.0 90.0 92.0
```

```
# of subjects in the validation study: 127
# of person-year observations in the validation study: 127
# of risk sets in the validation study: 3
```

```
-----
Information about the measurement error model
fitting in the validation study:
-----
```

Risk_set	Age_low	Age_up	Person_year	Intercept	slope ¹	R ²
1	39.0	47.0	42	0.308	0.482	0.364
2	48.0	60.0	42	0.258	0.534	0.407
3	61.0	92.0	43	0.074	0.343	0.160

¹ This slope is for the main exposure --- diet score. Since it's a two-dimensional analysis, there is a slope for both diet score and BMI. Due to format limitations, %RRC only displays the slope for the main exposure.

```
-----
Results for uncorrected estimates (naive cox)
and corrected estimates (RRC):
-----
```

```
Uncorrected result:
```

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
ttlcarb	0.586	0.062	0.000	1.796	[1.59, 2.03]
bmi	0.090	0.002	0.000	1.095	[1.09, 1.10]

```
RRC result:
```

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
ttlcarb	0.922	0.765	0.228	2.515	[0.56, 11.27]
bmi	0.082	0.008	0.000	1.085	[1.07, 1.10]

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

If BMI is only a risk factor in the primary regression model, not associated with the measurement error model, then we call the macro as follows:

```
%rrc
```

```
(
id=id,
main=all,
validation=vali,
surrogate=tlcarb,
true=tlcarbdr,
extravariablenmain=bmi, /* BMI is a risk factor in the primary
                           regression model only */
case=allt2d,
time=ageyr,
groupnum=40, /* Increase the size of groupnum to stabilize
               the results */
increments=20, /* difference of 90th and 10th percentile */
type=1, /* For cumulatively updated exposure */
emodel=1, /* For the general linear measurement error */
timestart=, /* not needed here, optional */
timeend=, /* not needed here, optional */
residual=0,
filename=Outputgen_BMIa);
```

Then the text file "Outputgen_BMIa" showed:

```
%%%%%%%%%%
This analysis is for cumulatively average updated exposure
with general linear measurement error model.
```

The increment for each exposure is: 20.0

There is one exposure and one or more covariates in the analysis. The exposure is measured with error, but the covariates are error-free.

```
# of subjects in the main study: 41539
# of person-year observations in the main study: 380337
# of cases in the main study: 2790
# of unique failure time: 53
```

```
The unique failure times are:
39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
89.0 90.0 92.0
```

```
# of subjects in the validation study: 127
# of person-year observations in the validation study: 127
# of risk sets in the validation study: 3
```

```
-----
Information about the measurement error model
fitting in the validation study:
-----
```

```
Risk_set  Age_low  Age_up  Person_year  Intercept  slope  R^2
```

1	39.0	47.0	42	0.435	0.498	0.364
2	48.0	60.0	42	0.327	0.540	0.407
3	61.0	92.0	43	0.435	0.359	0.160

Results for uncorrected estimates (naive cox)
and corrected estimates (RRC):

Uncorrected result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
ttlcarb	0.586	0.062	0.000	1.796	[1.59, 2.03]
bmi	0.090	0.002	0.000	1.095	[1.09, 1.10]

RRC result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
ttlcarb	1.266	0.230	0.000	3.547	[2.26, 5.56]
bmi	0.090	0.001	0.000	1.094	[1.09, 1.10]

%%%

Although the classical linear measurement error model is not a realistic model for the dietary exposures, we use it below for the illustration of the use of the macro with `type=1`, `emodel=2` in the crude analysis. We call the macro as follows:

```
%rrc
(
id=id,
main=all,
validation=vali,
surrogate=tlcarb,
true=tlcarbdr,
confounder=, /* leave it blank */
case=allt2d,
time=ageyr,
groupnum=5,
increments=20, /* difference of 90th and 10th percentile */
type=1, /* For cumulatively updated exposure*/
emodel=2, /* For classical measurement error model */
icc=0.6, /* correlation for the surrogate exposure */
dampfactor=0.0, /* try compound symmetry covariance */
residual=0,
filename= Outputadd_crude);
```

Then the text file "Outputadd_crude" showed:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
This analysis is for cumulatively average updated exposure
with classical additive measurement error model.
```

The increment for each exposure is: 20.0

There is only one exposure in the analysis and no covariate.
The main exposure is measured with error.
It used compound symmetry covariance matrix

with ICC= 0.6 .

of subjects in the main study: 41539
of person-year observations in the main study: 380337
of cases in the main study: 2790
of unique failure time: 53

The unique failure times are:
39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
89.0 90.0 92.0

of subjects in the validation study: 127
of person-year observations in the validation study: 127
of risk sets in the validation study: 19

Information about the measurement error model
fitting in the validation study:

Risk_set	Age_low	Age_up	Person_year	Intercept	slope	R^2
1	39.0	40.0	9	0.425	0.496	0.195
2	41.0	42.0	12	0.427	0.471	0.414
3	43.0	43.0	5	0.367	0.395	0.713
4	44.0	44.0	6	0.394	0.491	0.601
5	45.0	46.0	7	0.515	0.772	0.464
6	47.0	48.0	5	0.228	0.598	0.789
7	49.0	50.0	5	0.718	0.179	0.101
8	51.0	53.0	6	0.839	-0.029	0.000
9	54.0	55.0	8	0.099	0.846	0.963
10	56.0	57.0	6	-0.139	0.959	0.876
11	58.0	59.0	11	0.635	0.314	0.320
12	60.0	61.0	7	0.321	0.479	0.218
13	62.0	62.0	5	0.831	-0.178	0.041
14	63.0	63.0	5	0.087	0.835	0.556
15	64.0	65.0	5	0.493	0.167	0.101
16	66.0	67.0	8	0.500	0.488	0.297
17	68.0	68.0	5	0.062	0.554	0.240
18	69.0	71.0	6	0.816	-0.093	0.007
19	72.0	92.0	6	0.352	0.444	0.585

Results for uncorrected estimates (naive cox)
and corrected estimates (RRC):

Uncorrected result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
t11carb	0.817	0.062	0.000	2.264	[2.01, 2.56]

RRC result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
t11carb	1.000	0.082	0.000	2.718	[2.31, 3.19]

%%

1(b): Analysis for simple updated diet scores.

Here, we look at the effect of the simple updated low-carbohydrate diet with high total protein and fat. BMI is a potential confounder in the analysis.

We can set up the main study dataset and the validation study dataset exactly the same as in Example 1(a). Then call the macro as:

```
%rrc
(
id=id,
main=all,
validation=vali,
surrogate=tlcarb,
true=tlcarbdr,
confounder=bmi, /* BMI is a confounder */
case=allt2d,
time=ageyr,
groupnum=50,
increments=20,
type=4, /* For simply updated exposure */
residual=0,
filename=OutputSimple_BMI);
```

The text file “OutputSimple_BMI” is produced by the call above to %RRC:

%%

This analysis is for simple updated exposure.

The increment for each exposure is: 20.0

There is one exposure and one or more covariates in the analysis. The exposure is measured with error, but the covariates are error-free.

```
# of subjects in the main study: 41539
# of person-year observations in the main study: 380337
# of cases in the main study: 2790
# of unique failure time: 53
```

The unique failure times are:

```
39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
89.0 90.0 92.0
```

```
# of subjects in the validation study: 127
# of person-year observations in the validation study: 127
# of risk sets in the validation study: 2
```

```
-----
Information about the measurement error model
fitting in the validation study:
-----
```

Risk_set	Age_low	Age_up	Person_year	Intercept	slope	R^2
1	39.0	51.0	51	0.421	0.456	0.341
2	52.0	92.0	76	0.028	0.459	0.296

```
-----
Results for uncorrected estimates (naive cox)
and corrected estimates (RRC):
-----
```

Uncorrected result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
t1lcarb	0.465	0.055	0.000	1.592	[1.43, 1.77]
bmi	0.091	0.002	0.000	1.095	[1.09, 1.10]

RRC result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
t1lcarb	0.856	0.265	0.001	2.353	[1.40, 3.96]
bmi	0.080	0.007	0.000	1.083	[1.07, 1.10]

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

1(c): Analysis for diet scores at baseline.

Here, we examine the effect of low-carbohydrate diet with high total protein and fat at baseline. BMI is the potential confounder in the analysis.

You need to prepare the main study dataset and the validation dataset with time-independent baseline covariates. Then read in the main study dataset as follows:

```
Data all_baseline;
infile "hpfsMain2_withBMI_baseline.dat";
input id allt2d t1lcarb bmi ageyr_start ageyr_end;
run;
```

And the validation dataset can be read in as follows:

```
Data vali;
infile "hpfsVal_dscore_bmi.dat";
input id ageyr t1lcarbdr t1lcarb bmi;
run;
```

Note: The validation study dataset is the same as in Example 1(a) and 1(b). However, the main study dataset is different. There is only one observation per subjects. The exposure “t1lcarb” and the confounder “bmi” are both baseline measurements. The time variable for event or censoring, “ageyr”, is included in the validation study dataset, but not in the main study dataset. Instead, the main study dataset has another two variables, “ageyr_start” for the age at the

start of follow-up, and “ageyr_end” for the age at the end of follow-up (i.e. got an event or censored).

Then we can call the macro as:

```
%rrc
(
id=id,
main=all_baseline,
validation=vali,
surrogate=tlcarb,
true=tlcarbdr,
confounder=bmi, /* BMI is a confounder */
case=allt2d,
time=ageyr, /* the same name as the time variable in the
validation study */
groupnum=50,
increments=20,
type=5, /* For baseline exposure */
timestart=ageyr_start,
timeend=ageyr_end,
residual=0,
filename=Outputbaseline_BMI);
```

The text file “Outputbaseline_BMI” is produced by the call above to %RRC:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
This analysis is for baseline exposure.
```

The increment for each exposure is: 20.0

There is one exposure and one or more covariates in the analysis. The exposure is measured with error, but the covariates are error-free.

```
# of subjects in the main study: 41539
# of person-year observations in the main study: 711494
# of cases in the main study: 2790
# of unique failure time: 53
```

```
The unique failure times are:
39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
89.0 90.0 92.0
```

```
# of subjects in the validation study: 127
# of person-year observations in the validation study: 127
# of risk sets in the validation study: 2
```

Information about the measurement error model fitting in the validation study:

```
-----
```

Risk_set	Age_low	Age_up	Person_year	Intercept	slope	R ²
1	39.0	51.0	51	0.421	0.456	0.341
2	52.0	92.0	76	0.028	0.459	0.296

```
-----
```

Results for uncorrected and corrected relative risks (RRC):

Uncorrected result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
ttlcarb	0.430	0.055	0.000	1.538	[1.38, 1.71]
bmi	0.087	0.002	0.000	1.091	[1.09, 1.10]

```
-----
```

RRC result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
ttlcarb	0.785	0.251	0.002	2.193	[1.34, 3.58]
bmi	0.077	0.007	0.000	1.080	[1.07, 1.09]

Example 2: Measurement error correction for the association between 12-month moving average PM 2.5 exposure and all cause mortality in the Nurses' Health Study air pollution cohort using %rrc macro.

Reference: J. E. Hart, X. M. Liao, B. L. Hong, R. C. Puett, J. D. Yanosky, H. Suh, M. A. Kioumourtzoglou, D. Spiegelman and F. Laden, The Association of Long-Term Exposure to PM_{2.5} on All-Cause Mortality in the Nurses' Health Study and the Impact of Measurement-Error Correction. *Environmental Health*, 14(1): 38.

First we need to read in the main study and validation study datasets. Since the exposure metrics we are interested in here is 12-month moving average exposure, we need to have an additional 12 months follow-up before the baseline of the main study. The main study dataset and the additional 12-month follow-up dataset can be read in as follows:

```
/* The main study dataset*/
/* It includes the period variable as the indicator for each
   period */
data main;
    set XXXX1;
run;

/* Additional 12-month follow-up before the baseline */
/* Change the period variable to -11 to 0 */
data main_ext;
    set XXXX2;
    period=period-12;
run;

/* Put two data sets together */
data main_all;
    set main main_ext;
run;
```

```

/* Set the missing indicator */
data main_all;
    set main_all;
    missind=0;
    if pmlavg=. then missind=1;
run;

/* The validation study dataset */
data vali;
    set YYYY;
run;

/* Delete the missing values in the validation study dataset */
data vali1;
    set vali;
    if pmlavg=. or trueexp=. or agemo=. then delete;
run;

```

Call the macro as:

```

%rrc
(
id=id,
main=main_all, /* Name of main dataset, required */
validation=vali1, /* Name of validation dataset, required */
surrogate=pmlavg, /* predictor with error in main dataset,
required */
true=trueexp, /* predictor without error in main and
validation dataset, required */
confounder=reg2 reg3 reg4 heatsn, /* covariates adjusted in
both models, optional */
extravariabileinmain=yr, /* Covariates adjusted in the primary
regression model only */
extravariabileinval=season_jeff, /* Covariates adjusted in the
measurement error model only */
missing=missind, /* missing indicator */
period=period,
case= caseall, /* variable that indicates whether an event
occurs or not, required */
time= agemo, /* variable for time-to-event outcome, required */
groupnum=50, /* the minimum number of subjects in each risk set
*/
increments=10.0, /* the unit of true predictor, optional,
default=1*/
type=2, /* type of data, 2: 12-month moving average exposure */
month1=1,
month=12,
emodel=1,
residual=0,
initialvalue=0.181 -0.03 -0.06 -0.06 0.03 0.07,
filename=Outputgen_whole_crude00-06_num50_12ave_ambor_new,
csvfilename=Outputgen_whole_crude00-06_num50_12ave_ambor_new.csv

```

);

The text file "Outputgen_whole_crude00-06_num50_12ave_ambor_new" is produced by the call above to %RRC:

%%%

This analysis is for 1 - 12 month average exposure with general linear measurement error model.

The increment for each exposure is: 10.0

There is one exposure and one or more covariates in the analysis. The exposure is measured with error, but the covariates are error-free.

of subjects in the main study: 108765
of person-time observations in the main study: 7538537
of cases in the main study: 8604
of unique failure time: 367

The unique failure times are:

646.0 651.0 652.0 653.0 655.0 659.0 660.0 663.0 664.0 665.0
666.0 667.0 669.0 670.0 671.0 672.0 673.0 674.0 675.0 676.0
677.0 678.0 680.0 681.0 682.0 683.0 684.0 685.0 686.0 687.0
688.0 689.0 690.0 691.0 693.0 694.0 695.0 696.0 697.0 698.0
699.0 700.0 701.0 702.0 703.0 704.0 705.0 706.0 707.0 708.0
709.0 710.0 711.0 712.0 713.0 714.0 715.0 716.0 717.0 718.0
719.0 720.0 721.0 722.0 723.0 724.0 725.0 726.0 727.0 728.0
729.0 730.0 731.0 732.0 733.0 734.0 735.0 736.0 737.0 738.0
739.0 740.0 741.0 742.0 743.0 744.0 745.0 746.0 747.0 748.0
749.0 750.0 751.0 752.0 753.0 754.0 755.0 756.0 757.0 758.0
759.0 760.0 761.0 762.0 763.0 764.0 765.0 766.0 767.0 768.0
769.0 770.0 771.0 772.0 773.0 774.0 775.0 776.0 777.0 778.0
779.0 780.0 781.0 782.0 783.0 784.0 785.0 786.0 787.0 788.0
789.0 790.0 791.0 792.0 793.0 794.0 795.0 796.0 797.0 798.0
799.0 800.0 801.0 802.0 803.0 804.0 805.0 806.0 807.0 808.0
809.0 810.0 811.0 812.0 813.0 814.0 815.0 816.0 817.0 818.0
819.0 820.0 821.0 822.0 823.0 824.0 825.0 826.0 827.0 828.0
829.0 830.0 831.0 832.0 833.0 834.0 835.0 836.0 837.0 838.0
839.0 840.0 841.0 842.0 843.0 844.0 845.0 846.0 847.0 848.0
849.0 850.0 851.0 852.0 853.0 854.0 855.0 856.0 857.0 858.0
859.0 860.0 861.0 862.0 863.0 864.0 865.0 866.0 867.0 868.0
869.0 870.0 871.0 872.0 873.0 874.0 875.0 876.0 877.0 878.0
879.0 880.0 881.0 882.0 883.0 884.0 885.0 886.0 887.0 888.0
889.0 890.0 891.0 892.0 893.0 894.0 895.0 896.0 897.0 898.0
899.0 900.0 901.0 902.0 903.0 904.0 905.0 906.0 907.0 908.0
909.0 910.0 911.0 912.0 913.0 914.0 915.0 916.0 917.0 918.0
919.0 920.0 921.0 922.0 923.0 924.0 925.0 926.0 927.0 928.0
929.0 930.0 931.0 932.0 933.0 934.0 935.0 936.0 937.0 938.0
939.0 940.0 941.0 942.0 943.0 944.0 945.0 946.0 947.0 948.0
949.0 950.0 951.0 952.0 953.0 954.0 955.0 956.0 957.0 958.0
959.0 960.0 961.0 962.0 963.0 964.0 965.0 966.0 967.0 968.0
969.0 970.0 971.0 972.0 973.0 974.0 975.0 976.0 977.0 978.0
979.0 980.0 981.0 982.0 983.0 984.0 985.0 986.0 987.0 988.0
989.0 990.0 991.0 992.0 993.0 994.0 995.0 996.0 997.0 998.0
999.0 1000.0 1001.0 1002.0 1003.0 1004.0 1005.0 1006.0 1007.0 1008.0

1009.0 1010.0 1011.0 1012.0 1013.0 1014.0 1015.0 1016.0 1017.0 1018.0
1019.0 1020.0 1021.0 1022.0 1023.0 1024.0 1044.0

of subjects in the validation study: 118
of person-time observations in the validation study: 215
of risk sets in the validation study: 3

Information about the measurement error model
fitting in the validation study:

Risk_set	Age_low	Age_up	Person_year	Intercept	slope	R^2
1	646.0	792.0	52	0.410	0.336	0.581
2	793.0	862.0	50	0.251	0.368	0.420
3	863.0	1044.0	82	0.179	0.551	0.466

Results for uncorrected estimates (naive cox)
and corrected estimates (RRC):

Uncorrected result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
fine_jeff	0.181	0.039	0.000	1.198	[1.11, 1.29]
reg2	-0.059	0.032	0.063	0.943	[0.89, 1.00]
reg3	-0.060	0.032	0.061	0.942	[0.89, 1.00]
reg4	0.026	0.029	0.369	1.026	[0.97, 1.09]
heatsn	0.074	0.022	0.001	1.076	[1.03, 1.12]
yr	-0.032	0.007	0.000	0.969	[0.96, 0.98]

RRC result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
fine_jeff	0.226	0.078	0.004	1.253	[1.08, 1.46]
reg2	-0.066	0.039	0.093	0.936	[0.87, 1.01]
reg3	-0.045	0.039	0.250	0.956	[0.89, 1.03]
reg4	0.005	0.034	0.892	1.005	[0.94, 1.07]
heatsn	0.074	0.022	0.001	1.077	[1.03, 1.12]
yr	-0.033	0.007	0.000	0.968	[0.96, 0.98]

%%

This macro call also generated the CSV file "Outputgen_whole_crude00-06_num50_12ave_ambor_new.csv" as follows, and it can be read into the user customized SAS program and produce the analysis tables.

```
variable,modeltype,para_es,stderr,pval,hr,lowerci,upperci,groupnum,num  
risksets  
fine_jeff,uncorrected, 0.181, 0.039, 0.000, 1.198, 1.110, 1.294,50,3  
reg2,uncorrected,-0.059, 0.032, 0.063, 0.943, 0.887, 1.003, 50, 3  
reg3,uncorrected,-0.060, 0.032, 0.061, 0.942, 0.885, 1.003, 50, 3  
reg4,uncorrected, 0.026, 0.029, 0.369, 1.026, 0.970, 1.085, 50, 3  
heatsn,uncorrected, 0.074, 0.022, 0.001, 1.076, 1.032, 1.123, 50, 3  
yr,uncorrected,-0.032, 0.007, 0.000, 0.969, 0.956, 0.981, 50, 3  
fine_jeff,rrc, 0.226, 0.078, 0.004, 1.253, 1.076, 1.459, 50, 3  
reg2,rrc,-0.066, 0.039, 0.093, 0.936, 0.867, 1.011, 50, 3  
reg3,rrc,-0.045, 0.039, 0.250, 0.956, 0.886, 1.032, 50, 3
```

```
reg4,rrc, 0.005, 0.034, 0.892, 1.005, 0.940, 1.074, 50, 3
heatsn,rrc, 0.074, 0.022, 0.001, 1.077, 1.032, 1.123, 50, 3
yr,rrc,-0.033, 0.007, 0.000, 0.968, 0.955, 0.980, 50, 3
```

Example 3: Measurement error correction for the association between cumulative radon exposure and all cause mortality in the Uranium Miners study cohort using %rrc macro.

First we need to read in the main study and validation study datasets. Since we have the internal validation study design we can use the value of gold standard exposure if available. The main study dataset and the additional 12-month follow-up dataset can be read in as follows:

```
/* The main study dataset*/
/* It includes the true value availability variable as the
indicator for each observation */
data main;
    set XXXX1;
surf=0;
if type='t' then surf=1;
run;

/* The validation study dataset */
data vali;
    set YYYY;
run;
```

Call the macro as:

```
%rrc(id=id, main=main, /*Name of main dataset, required*/
validation=vali, /*Name of validation dataset, required*/
surrogate=yrexp, /*predictor with error in main dataset,
required*/
true=value, /*predictor without error in main and validation
dataset, required*/
case=case, /*variable that indicates whether an event occurs or
not, required*/
time= age_tv , confounder= ,
increments=100,groupnum=170,
type=3, residual=0,
filename=crude_100_3,
useind=surf,
truepredict=predexpxm,
surrogatefunc=surtotxm,
outdata=outdata, path=.);
```

The text file “crude_100_3” is produced by the call above to %RRC:

```
%%%%%%%%%%
This analysis is for cumulatively total updated exposure
with general linear measurement error model.
```

The increment for each exposure is: 100.0

There is only one exposure in the analysis and no covariate.
The main exposure is measured with error.

of subjects in the main study: 2327
of person-year observations in the main study: 88405
of cases in the main study: 132
of unique failure time: 39

The unique failure times are:
41.0 43.0 44.0 46.0 47.0 48.0 50.0 53.0 54.0 55.0
56.0 57.0 58.0 59.0 60.0 61.0 62.0 63.0 64.0 65.0
66.0 67.0 68.0 69.0 70.0 71.0 72.0 73.0 74.0 75.0
76.0 77.0 78.0 79.0 80.0 83.0 84.0 86.0 88.0

of subjects in the validation study: 862
of person-year observations in the validation study: 2833
of risk sets in the validation study: 3

Information about the measurement error model
fitting in the validation study:

Risk_set	Age_low	Age_up	Person_year	Intercept	slope	R ²
1	41.0	44.0	300	0.011	0.148	0.084
2	46.0	50.0	239	0.007	0.315	0.307
3	53.0	88.0	174	0.007	0.314	0.320

Results for uncorrected estimates (naive cox)
and corrected estimates (RRC):

Uncorrected result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
yrexp	0.379	0.044	0.000	1.461	[1.34, 1.59]

RRC result:

Variable	Para_es	S.E.	p_val	H.R.	95% C.I.
yrexp	1.684	0.318	0.000	5.388	[2.89, 10.04]

%%

This macro call also generated the SAS7BDAT file "outdata.sas7bdat", and it can be read into the user customized SAS program and produce the analysis tables.

4. Warnings

a) Core dumped

If you see a “core” file dumped in your directory, or the error message such as “core dumped”, it means the underlying fortran program has some running-time exception such as “segmentation fault”, or “floating point overflow”, or “floating point exception”. One reason could be that the “groupnum” is too small and there are few subjects in each risk set, which results in the instability of the results. You may increase the value of “groupnum” to stabilize the results.

b) Errors in call to %rrc

```
% rrc
(
id = id,
main= one,
validation= vali,
surrogate= surrogate1 surrogate2 surrogate3,
true= true1 true2,
confounder= ,
case= censoring,
time= timetodeath,
groupnum= 5,
increments= 1 ,
type=1,
emodel=1,
icc=,
dampfactor=,
timestart=,
timeend=,
residual=0,
filename=XXX
);
```

SAS Log:

```
ERROR IN SAS MACRO RRC: *****

***** The number of true predictors and the number of surrogate predictors
should be equal *****
```

More errors report will be updated later.

References

Xiaomei Liao, David M. Zucker, Yi Li and Donna Spiegelman (2011) Survival analysis with error-prone time-varying covariates: a risk set calibration approach. *Biometrics* 67(1), 50-58, 2011, March.

Liao X, Zhou X, Wang M, Hart JE, Laden F, Spiegelman D. Survival analysis with functions of mismeasured covariate histories: the case of chronic air pollution exposure in relation to mortality in the nurses’ health study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2018 Feb 1.

Credits

Written by Xiaomei Liao, Ph.D, Harvard School of Public Health, Boston, MA; Donna Spiegelman, Sc.D., Harvard School of Public Health, Boston, MA. Thank Samuela Pollack for her programming support.

Questions can be directed to Xiaomei Liao (stxia@channing.harvard.edu), Samuela Pollack (spollack@hsph.harvard.edu) and Donna Spiegelman(stdls@hsph.harvard.edu) .