SAS coding examples for case-cohort designs

"Simple" scenario (Table, row 1) All cases selected, selection probability of sub-cohort = x%
Example: O'Brien et al. (2017) Serum Vitamin D and Risk of Breast Cancer within Five Years. *Environ Health Persp*

cc1 = a data set containing the case-cohort data, including the following variables
- subcohort = 1 if in subcohort; 0 if not
- case = 1 if a case; 0 if not
- age_enrollment = age at enrollment
- age_eof = age at end of follow-up (e.g. event time or censoring time)
- exp = exposure of interest
- covar1, covar2, covar3 = covariates of interest (coded as categories)
- ID = identification variable

sampling_rate= number of particants in sub-cohort / number of participants in full eligible cohort

```
*test code;
%LET epsilon=0.01; *or any number smaller than your smallest time
unit;
%LET sampling_rate=0.05; *for the example data set cc1;

*restructure data set so that cases in sub-cohort weighted differently
according to time (will appear as two entries);
DATA ccnew1;
      SET wcc.cc1;
      *cases within subcohort - contribute fully until just before
      diagnosis;
      IF subcohort=1 AND case=1 THEN DO;
            start = age_enrollment;
            stop= age_eof - &epsilon;
            event = 0; *considered a censored observation;
      wt= 1/&sampling_rate; *inverse probability of sampling weight;
      OUTPUT;
      END;

      *all cases contribute person-time right before event, count as
      event;
      IF case=1 THEN DO;
            start = age_eof - &epsilon;
            stop = age_eof;
            event = 1;
            wt=1;
      OUTPUT;
      END;

      *non-cases within subcohort - contribute full person time,
      censored;
      ELSE IF subcohort=1 AND case=0 THEN DO;
            start = age_enrollment;
            stop = age_eof;
```

```
            event = 0;
            wt= 1/&sampling_rate;
            *inverse probability of sampling weight;
      OUTPUT;
      END;
RUN;

PROC PHREG DATA=ccnew1 covs(aggregate);
      CLASS covar1 covar2 covar3;
      MODEL (start,stop)*event(0) = exp covar1 covar2 covar3;
      WEIGHT wt;
      ID ID;
      HAZARDRATIO exp;
RUN;
```

Covariate-stratified case-cohort (Table, row 2) All cases selected, Sub-cohort selection probabilities of x_A% (Group A) and x_B% (Group B)
Example: Niehoff et al. (*in review*) Metals and breast cancer risk: a prospective study using toenail biomarkers

cc2 = a data set containing the case-cohort data, including the following variables
subcohort = 1 if in subcohort; 0 if not
case = 1 if a case; 0 if not
age_enrollment = age at enrollment
age_eof = age at end of follow-up (e.g. event time or censoring time)
exp = exposure of interest
covar2, covar3 = covariates of interest (coded as categories)
ID = identification variable
groupA=1 if in group A; 0 if in group B

sampling_rateA= number in sub-cohort from group A / number in full cohort from group A
sampling_rateB= number in sub-cohort from group B / number in full cohort from group B

```
%LET sampling_rateA=0.08; *for the example data set cc2;
%LET sampling_rateB=0.15; *for the example data set cc2;

*restructure data set so that cases in sub-cohort weighted differently
according to time (will appear as two entries);
DATA ccnew2;
      SET wcc.cc2;
*cases within subcohort - contribute fully until just before
diagnosis;
      IF subcohort=1 AND case=1 THEN DO;
            start = age_enrollment;
            stop= age_eof - &epsilon;
            event = 0; *considered a censored observation;
      IF groupA=1 THEN wt= 1/&sampling_rateA;
      ELSE IF groupA=0 THEN wt= 1/&sampling_rateB;
*inverse probability of sampling weights;
      OUTPUT;
      END;

      *all cases contribute person-time right before event, count as
event;
      IF case=1 THEN DO;
            start = age_eof - &epsilon;
            stop = age_eof;
            event = 1;
            wt=1;
      OUTPUT;
      END;

      *non-cases within subcohort - contribute full person time,
censored;
      ELSE IF subcohort=1 AND case=0 THEN DO;
```

```
            start = age_enrollment;
            stop = age_eof;
            event = 0;
IF groupA=1 THEN wt= 1/&sampling_rateA;
ELSE IF groupA=0 THEN wt= 1/&sampling_rateB;
*inverse probability of sampling weights;
      OUTPUT;
      END;
RUN;

PROC PHREG DATA=ccnew2 covs(aggregate);
      CLASS covar2 covar3;
      MODEL (start,stop)*event(0) = exp groupA covar2 covar3;
      WEIGHT wt;
      ID ID;
      HAZARDRATIO exp;
RUN;
```

Outcome-stratified case-cohort (Table, row 3) 100% of type I cases and y% of type 2 cases selected; sub-cohort selection probability x% for all
Example: Sampling 100% of estrogen receptor-negative breast cancers and 20% of estrogen receptor-positive breast cancers, with the desire to look at subtype-specific and overall exposure-disease associations

cc3 = a data set containing the case-cohort data, including the following variables
subcohort = 1 if in subcohort; 0 if not
case = 1 if a case; 0 if not
age_enrollment = age at enrollment
age_eof = age at end of follow-up (e.g. event time or censoring time)
exp = exposure of interest
covar2, covar3 = covariates of interest (coded as categories)
ID = identification variable
Subtype1=1 if case of disease subtype 1; 0 otherwise
Subtype2=1 if case of disease subtype 2; 0 otherwise

sampling_rate= number of particants in sub-cohort / number of participants in full eligible cohort
sampling_rate_subtype1= number of case of subtype 1 selected / total number of subtype 1 cases
sampling_rate_subtype2= number of case of subtype 2 selected / total number of subtype 2 cases

```
%LET epsilon=0.01; *or any number less than your smallest time unit;
%LET sampling_rate=0.05; *for the example data set cc3;
%LET sampling_rate_subtype1=0.20; *20% of subtype1 selected;
%LET sampling_rate_subtype2=1; *100% of subtype2 selected;

*restructure data set so that cases in sub-cohort weighted differently
according to time (will appear as two entries);
DATA ccnew3;
      SET wcc.cc3;
      *selected cases within subcohort - contribute fully until just
      before diagnosis;
      IF subcohort=1 AND (subtype1=1 | subtype2=1) THEN DO;
            start = age_enrollment;
            stop= age_eof - &epsilon;
            event = 0; *considered a censored observation;
            wt= 1/&sampling_rate;
            *inverse probability of sampling weight;
      OUTPUT;
      END;

      *cases contribute person-time right before event only if
      selected, contribute based on weights;
      IF (subtype1=1 | subtype2=1) THEN DO;
            start = age_eof - &epsilon;
            stop = age_eof;
            event = 1;
            IF subtype1=1 THEN wt=1/&sampling_rate_subtype1;
                  ELSE IF subtype2=1 THEN wt=1/&sampling_rate_subtype2;
```

```
      OUTPUT;
      END;

      *non-cases within subcohort - contribute full person time,
censored;
      ELSE IF subcohort=1 AND subtype1=0 AND subtype2=0 THEN DO;
            start = age_enrollment;
            stop = age_eof;
            event = 0;
            wt= 1/&sampling_rate; *inverse probability of sampling
weight;
      OUTPUT;
      END;
RUN;

PROC PHREG DATA=ccnew3 covs(aggregate);
      CLASS covar1 covar2 covar3;
      MODEL (start,stop)*event(0) = exp covar1 covar2 covar3;
      WEIGHT wt;
      ID ID;
      HAZARDRATIO exp;
RUN;
```

Covariate and outcome-stratified case-cohort (Table, row 4) 100% of type I cases and y% of type 2 cases selected; Sub-cohort selection probabilities of x_A% (Group A) and x_B% (Group B)
NOTE: This assumes that case status and subgroup status are selected independently; if this is not true, weights can be re-calculated for each subgroup/subtype combination (= a product of the specified weights)
Example: Oversampling for African-American women and estrogen receptor-negative breast cancers

cc4 = a data set containing the case-cohort data, including the following variables
- subcohort = 1 if in subcohort; 0 if not
- case = 1 if a case; 0 if not
- age_enrollment = age at enrollment
- age_eof = age at end of follow-up (e.g. event time or censoring time)
- exp = exposure of interest
- covar1, covar2, covar3 = covariates of interest (coded as categories)
- ID = identification variable
- groupA=1 if in group A; 0 if in group B
- Subtype1=1 if case of disease subtype 1; 0 otherwise
- Subtype2=1 if case of disease subtype 2; 0 otherwise

sampling_rateA= number in sub-cohort from group A / number in full cohort from group A
sampling_rateB= number in sub-cohort from group B / number in full cohort from group B
sampling_rate_subtype1= number of case of subtype 1 selected / total number of subtype 1 cases
sampling_rate_subtype2= number of case of subtype 2 selected / total number of subtype 2 cases

```
%LET epsilon=0.01; *or any number less than your smallest time unit;
%LET sampling_rateA=0.08; *for the example data set cc4;
%LET sampling_rateB=0.15; *for the example data set cc4;
%LET sampling_rate_subtype1=0.20; *20% of subtype1 selected;
%LET sampling_rate_subtype2=1; *100% of subtype2 selected;

*restructure data set so that cases in sub-cohort weighted differently
according to time (will appear as two entries);
DATA ccnew4;
      SET wcc.cc4;
      *selected cases within subcohort - contribute fully until just
      before diagnosis;
      IF subcohort=1 AND (subtype1=1 | subtype2=1) THEN DO;
            start = age_enrollment;
            stop= age_eof - &epsilon;
            event = 0; *considered a censored observation;
            IF groupA=1 THEN wt= 1/&sampling_rateA;
                  ELSE IF groupA=0 THEN wt=1/&sampling_rateB;
            *inverse probability of sampling weight;
      OUTPUT;
      END;

      *cases contribute person-time right before event only if
      selected, contribute based on weights;
```

```
      IF (subtype1=1 | subtype2=1) THEN DO;
            start = age_eof - &epsilon;
            stop = age_eof;
            event = 1;
            IF subtype1=1 THEN wt=1/&sampling_rate_subtype1;
                  ELSE IF subtype2=1 THEN wt=1/&sampling_rate_subtype2;
      OUTPUT;
      END;

      *non-cases within subcohort - contribute full person time,
      censored;
      ELSE IF subcohort=1 AND subtype1=0 AND subtype2=0 THEN DO;
            start = age_enrollment;
            stop = age_eof;
            event = 0;
            IF groupA=1 THEN wt= 1/&sampling_rateA;
                  ELSE IF groupA=0 THEN wt=1/&sampling_rateB;
            *inverse probability of sampling weight;
      OUTPUT;
      END;
RUN;

PROC PHREG DATA=ccnew4 covs(aggregate);
      CLASS covar2 covar3;
      MODEL (start,stop)*event(0) = exp groupA covar2 covar3;
      WEIGHT wt;
      ID ID;
      HAZARDRATIO exp;
RUN;
```

Case-independent designs (Table, row 5) v% cases and z% of non-cases included in case-cohort sample; want to measure the association between previously measured exposure ("exp") and a second exposure ("exp2"), independent of case status
Example: Lawrence et al. (2020) Association of neighborhood deprivation with epigenetic aging using four clock methodologies. *JAMA Open*

Sampling_rate_cases= number of selected cases / total number of cases
sampling_rate_subcohort= number selected into subcohort / total number in cohort

```
%LET sampling_rate_cases=1; *for the example data set cc5 (all cases);
%LET sampling_rate_subcohort=0.05; *5% of cohort selected into
subcohort;

DATA wcc.cc5;
      SET wcc.cc5;
      IF case=1 THEN wt= 1/&sampling_rate_cases;
            ELSE IF case=0 THEN wt= 1/&sampling_rate_subcohort;
RUN;

PROC GLM DATA=wcc.cc5;
      CLASS exp covar1 covar2 covar3 / DESC;
      MODEL exp2 = exp age_enrollment covar1 covar2 covar3 / SOLUTION
      CLPARM;
      WEIGHT wt;
RUN;
QUIT;
```