WEB APPENDIX

Standardizing Discrete-Time Hazard Ratios With a Disease Risk Score

David B. Richardson, Alexander P. Keil, Jess Edwards, Alan C. Kinlaw, and Stephen R. Cole

Correspondence to David Richardson, Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA 27599 Phone: 919-966-2675 FAX: 919-966-2089 (email: david.richardson@unc.edu)

Author affiliations: Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill (David B. Richardson, Alexander P. Keil, Jess Edwards, and Stephen R. Cole); Division of Pharmaceutical Outcomes and Policy, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Alan C. Kinlaw).

David B. Richardson was supported by grant R01 OH011409 from the National Institute for Occupational Safety and Health of the Centers for Disease Control and Prevention. Alexander P. Keil received funding support from NICHD DP2-HD-08-4070.

Assume a cohort data set named 'A' includes one record per person with the following variables: i X Z U T d, which correspond to i, X_i, Z_i, τ, T_i and d_i in the main text.

These data could be represented in a person-period structure with contiguous 1-unit periods (i.e,. $L = 1$). Each cohort member, i, contributes one row of data for each time period of observation, letting j index discrete time intervals, from entry time 0 through the administrative end of follow-up. The binary outcome indicator, Y, takes a value of '1' at period $j = T$ if person i experienced the outcome of interest, else '0'. For each record, we define $q = I[j \leq T]$, and $w = I[X = 0]\, q$. The data step in the code below generates a person-period data structure, named 'B'.

A w weighted regression model is fitted to the data and outputs predict value, g, as a function of j, Z and appends these to each record and creates a data structure named 'C'.

Then, an extended data structure, named 'D' is constructed. Each cohort member contributes 2 rows of data for each time period of observation, indexed k = 1,0. If k = 1 then the variable m equals Y, and the variable n equals q. If k = 0 then m equals the calculated value gS, and n equals the calculated value S.

Using this expanded data set, we estimate a standardized rate ratio by defining log of n as an offset, and fitting a generalized linear model (with robust variance estimator) as follows:

```
data B (drop=d T U);
 set A;
  do j = 0 to U ;
   if j=T then Y=d; else Y=0;
   q=(j<= min(T, U));
   w=q*(X=0);
   output;
  end;

proc genmod data=B descending;
            model Y = j Z / link=logit dist=bin;
            weight w;
            output out=C p=g;

data D;
 set C; by i j;
  retain S F;
  if first.i then do; S=1; F=0; end;
  gS=g*S;    F=F+gS;
  k=1; m=Y;  n=max(1e-8,q); offset=log(n); output;
  k=0; m=gS; n=max(1e-8,S); offset=log(n); output;
  S=1-F;

proc genmod data=D; where X>0;
            class i;
            model m = k / link=log dist=POISSON offset=offset;
            repeated subject=i / type=ind; run;
```

Bootstrapping confidence intervals

Confidence intervals may be obtained by bootstrapping the data set as follows:

* The code below uses 200 samples, but this can be changed by the user;

```
* Bootstrap Step 0: Make replicates of same size;
proc freq data= A; table i / noprint out=_ncovals (rename=(count=_ncovals)
drop=percent);run;

proc sql noprint ;
select sum(_ncovals)    into :maxsize    from _ncovals; quit;
%let maxsize= &maxsize;

proc surveyselect noprint data=A method=urs rep=200 n= &maxsize out=boot
seed=123; cluster i;

data boot;
 set boot;
 retain z 1 count 0;
 do k=1 to numberhits; idk=i *1000+k; count=count+1; output; end; run;

*Bootstrap Step 1;
 data B (drop=d T U i k NumberHits z count);
 set boot;
  do j = 0 to U ;
   if j=T then Y=d; else Y=0;
   q=(j<= min(T, U));
   w=q*(X=0);
   output;
  end;

*Bootstrap Step 2  ;
  proc genmod data=B descending; by replicate; class z / ref=first ;
    model Y = j Z / link=logit dist=bin;
    weight w;  output out=C p=g;

*Bootstrap step 3;
data D;
 set C; by replicate idk j;
  retain S F;
  if first.idk then do; S=1; F=0; end;
  gS=g*S;    F=F+gS;
  k=1; m=Y;  n=max(1e-8,q); offset=log(n); output;
  k=0; m=gS; n=max(1e-8,S); offset=log(n); output;
  S=1-F; run;

*Bootstrap step 4;
ods output Estimates=rr_est ;
  proc genmod data=D; where X>0;by replicate;
      model m = k / link=log dist=POISSON offset=offset;
      estimate 'rr' k 1 / exp; run; ods rtf close;

data rr;  set rr_est;  if Label ne 'rr'; epred=LBetaEstimate;
proc univariate data=rr; var epred; output out=rr_cis pctlpts=2.5 97.5
pctlpre=rr_cis; run;
proc print data=rr_cis noobs label; run;
```