

eAPPENDIX – DESCRIPTION OF SIMULATIONS AND SIMULATION RESULTS

Simulations were performed in SAS, version 9.4, software (SAS Institute, Inc., Cary, North Carolina). Simulations were performed to demonstrate the expected presence of selection bias for the effect of race (defined at birth) on Z infection for the approach that corresponds to the researcher's original cross-sectional study where the target population is Black and White residents of City X born in 1936 and the study population is all Black and White City X residents who were born in 1936 and were living when study enrollment occurs in 2016. Simulations were also used to demonstrate how bias could be minimized by changing the target population or the study design. For all examined scenarios, 500 simulations with a target population of 10,000 individuals in the United States were performed and we assumed that migration and Z infections prior to or at birth in City X were non-existent or negligible.

To demonstrate the expected presence of selection bias for the effect of race on Z infection for the approach that corresponds to the researcher's original cross-sectional study, continuous times from birth to Z infection and death for the target population were generated following the method outlined by Bender and colleagues (1). The method outlined by Bender and colleagues (1) was followed so that the proportional hazards assumption held for both the times to Z infection and death. The continuous times were rounded up to the nearest integer so that the times to Z infection and death matched the integer birth and study year. Times from birth to Z infection and death were generated when 30% of the target population was Black (versus White) and 50% had $U=1$ (versus $U=0$). The aforementioned 30% was selected to facilitate an adequate number of Black and White residents in the study population to complete comparisons by race.

Additional specifications for the times from birth to Z infection for the target population included: (a) a constant hazard of Z infection (i.e., exponential distribution); (b) a mean age of Z infection of 67 years old for a White person with $U=0$ in the target population; and (c) hazard ratios for the effect of Black race and U on Z infection of 1 and 2.00, respectively. Additional specifications for the times from birth to death in the target population included: (a) a Gompertz distribution for the distribution of mortality in the United States; (b) a mean and standard deviation for life expectancy of 68 and 20 years old, respectively, for White residents with $U=0$ in the target population informed by the United States vital statistics (2); (c) hazard ratios for the effect of Black race on death of 1.50 when $U=0$ and 2.00 when $U=1$; and (d) hazard ratios for the effect of U on death of 1.50 among White residents and 2.00 among Black residents. All aforementioned hazard ratios were selected to match the relationships shown in the Figure and to ensure strong enough selection bias would be introduced based on one single common cause of mortality and Z infection. Lastly, for simplicity we assumed that any individual in the target population that died in or before 2016 (i.e., at or before reaching 80 years old) died before potential study enrollment and was therefore excluded from the study population. However, assuming that any individual in the target population that died in 2016 (i.e., at 80 years old) died after potential study enrollment and including those who died in 2016 in the study population did not meaningfully change the below described findings.

To demonstrate how survival-related selection bias could be minimized by changing the target population to be Black and White City X residents born in 1976 without changing the cross-sectional study design, the aforementioned simulation specifications were changed such that the mean life expectancy for White residents with $U=0$ in the target population was 83 years old, which was informed by the United States vital statistics (2). In addition, the hazard ratios for

the effect of Black race on death were specified to be 2.00 when $U=0$ and 2.50 when $U=1$ while the hazard ratios for the effect of U on death was specified to be 2.00 among White residents and 2.50 among Black residents. Furthermore, for simplicity we assumed that any individual in the target population that died in or before 2016 (i.e., at or before reaching 40 years old) died before potential study enrollment and was therefore excluded from the study population. However, assuming that any individual in the target population that died in 2016 (i.e., at 40 years old) died after potential study enrollment and including those who died in 2016 in the study population did not meaningfully change the below described findings.

To demonstrate how survival-related selection bias could be minimized by changing the study design to a cohort study without changing the target population, the simulation specifications were the same as outlined above for the approach that corresponded to the researcher's original cross-sectional study. However, the data were assumed to be ascertained by the researcher based on a cohort instead of a cross-sectional study design using a birth registry, Z infection surveillance data, and death data from the National Death Index or another source. Times to Z infection and death as well as other data were assumed to be accurately captured and a time-to-event framework was applied where follow up was the minimum of the simulated times to Z infection, the simulated times to death, and 80, and a Cox proportional hazard model was fit.

The simulation results indicated that for the approach that corresponded to the researcher's original cross-sectional study, the study population on average yielded estimates with some bias. Specifically, the estimates obtained from the researcher's original cross-sectional study population indicated that Black residents were less likely to be infected with the Z virus than White residents even though in the target population Black and White residents were

equally likely to be infected with the Z virus. The direction of this bias is likely due to the fact that Black residents and residents with $U=1$ were more likely to die by 2016 and in turn be excluded from the cross-sectional study population. The abovementioned exclusion likely created an inverse relationship between Black race and having a $U=1$ in the cross-sectional study population that does not exist in the target population. This inverse relationship was confirmed in the simulated data and would mean that in the cross-sectional study population, Black residents would be less likely to be infected with the Z virus because they were less likely to have a $U=1$ where having a $U=1$ was specified to increase the likelihood of Z infection in the simulations (3). Consistent with our expectation, changing the target population or study design as described above served to on average move effect estimates closer to the true value of zero (i.e., a null effect of race) compared to the estimates obtained from the approach that corresponds to the original cross-sectional study conducted by the researcher.

eAPPENDIX – SIMULATION CODE

```
*****Simulating data to generate researcher's original target population and cross-sectional study population*****;
data a1936;
  call streaminit(3);
  do n=10000; *Specifying that sample size of target population is 10,000;
    do k=1 to 500; *Specifying 500 simulations each with a target population sample size of 10,000;
      do i=1 to n;
        pA=.3; *Specifying that 30% of target population is Black and 70% is White;
        pU=.5; *Specifying that 50% of target population has a value of U=1;
        A=rand("bern",pA); *Generating variable A that is an indicator for Black (versus White) race;
        U=rand("bern",pU); *Generating variable U;
        AU=A*U;
        Unif_t=rand('UNIFORM');
        Unif_c=rand('UNIFORM');

        *Following the method of Bender et al SIM 2005 to generate Z infection that follows an exponential
        distribution;
        lambda_t=0.015; *Specifying that the baseline hazard function for Z infection (i.e., White person with
        U=0)corresponds to mean time to infection of 67 years old where 1/lambda_t is the mean
        time to infection;
        gammat1=0; *Specifying that the relative hazard of Z infection as a function of A is 1.0;
        gammat2=0.693; *Specifying that the relative hazard of Z infection as a function of U is 2.0;
        T=- (log(unif_t)/(lambda_t*exp(gammat1*A+gammat2*U))); *Generating the variable T that is the time from
        birth to Z infection;

        *Following the method of Bender et al SIM 2005 to generate mortality that follows a Gompertz
        distribution;
        alpha=3.14159/(sqrt(6)*20); *Specifying a standard deviation of 20 years for life expectancy for White
        residents with U=0 who were born in 1936;
        lambda_c=alpha*exp(-0.5772-alpha*68); *Specifying a mean life expectancy of 68 years for White residents
        with U=0 who were born in 1936;
        gammac1=0.405; *Specifying that the relative hazard of death as a function of A is 1.50 when U=0;
        gammac2=0.405; *Specifying that the relative hazard of death as a function of U is 1.50 when A=0;
        gammac3=0.288; *Specifying that the relative hazard of death is 2.00 as a function of A when U=1 and
        2.00 as a function of U when A=1;
        inverse_alpha=1/alpha;
        a_log_unif=alpha*(log(Unif_c));
        lambda_exp_beta_x=lambda_c*exp(gammac1*A+gammac2*U+gammac3*AU);
        C=inverse_alpha*log(1-a_log_unif/lambda_exp_beta_x); *Generating the variable C that is the time from
        birth to death;
```

```

                *Rounding (up) times to Z infection and death to be integers to match the time scale of follow up which
                is in integers;
                T_ceil=ceil(T);
                T=T_ceil;
                C_ceil=ceil(C);
                C=C_ceil;

                *Creating a variable for the observed follow up time (i.e., Ymin) as well as the corresponding indicator
                of having Z infection at the end of follow up in the presence of censoring due to death as well as
                administrative censoring (i.e., Dobs);
                Ymin=min(T,C,80);
                Dobs=0;
                if T<=80 then do; Dobs=1; end;
                if C<T<=80 then do; Dobs=0; end;

                *Creating indicator, S, of being alive in 2016 to be included in researcher's cross-sectional study;
                S=0;
                if C>80 then do; S=1; end;
                output;
            end;
        end;
    end;
run;

*Sorting data by simulation;
proc sort data=a1936; by k; run;

*Calculating the average proportion of individuals in the target population who are alive when the researcher's original cross-
sectional study would have begun enrollment;
proc means data=a1936 noprint; by k; var s; output out=s_mean1936 mean=s_mean; run;
proc means data=s_mean1936 n min max mean p25 median p75; var s_mean; run;

*Calculating the average ln prevalence ratio observed in study population for effect of Black race on Z infection;
proc genmod data=a1936;
    by k;
    class i;
    where s=1;
    model Dobs=a /link=log dist=poisson; ods output GEEEmpPEst=a1936_s;
    repeated subject=i/type=ind;
run;

```

```

data a1936_s_;
  set a1936_s;
  if parm='A';
  keep k parm estimate;
run;
proc means data=a1936_s_; var estimate; output out=a1936_s_mean_ (keep=a_s_mean) mean=a_s_mean; run;
data a1936_s_mean;
  set a1936_s_mean_;
run;
proc print data=a1936_s_mean; run;

```

*Calculating the average ln hazard ratio observed in study population for effect of Black race on Z infection when the study design is changed to a cohort study without changing the target population;

```

proc phreg data=a1936;
  by k;
  model Ymin*Dobs(0)= a /ties=efron rl; ods output ParameterEstimates=a1936_surv_tte;
run;
data a1936_surv_tte_;
  set a1936_surv_tte;
  keep k parameter estimate;
run;
proc means data=a1936_surv_tte_; var estimate; output out=a1936_surv_tte_mean_ (keep=a_surv_tte_mean) mean=a_surv_tte_mean; run;
data a1936_surv_tte_mean;
  set a1936_surv_tte_mean_;
run;
proc print data=a1936_surv_tte_mean; run;

```

*Calculating the bias and mean squared error observed in the study population for the researcher's original cross-sectional study as well as for the cohort study;

```

data a1936_s_m;
  set a1936_s_;
  rename estimate=est_a_s;
  drop parm;
run;
data a1936_surv_tte_m;
  set a1936_surv_tte_;
  rename estimate=est_a_surv_tte;
  drop parameter;
run;

```

```

proc sort data=a1936_s_m; by k; run;
proc sort data=a1936_surv_tte_m; by k; run;
data combined1936;
    merge a1936_s_m a1936_surv_tte_m;
    by k;
    bias_s=est_a_s-0;
    bias_surv_tte=est_a_surv_tte-0;
run;
proc print data=combined1936; run;

proc means data=combined1936; var bias_s bias_surv_tte;
output out=bias1936_ (keep= bias_s_mean bias_surv_tte_mean)
mean=bias_s_mean bias_surv_tte_mean;
run;
data bias1936;
    set bias1936_;
    v=1;
run;
proc sort data=bias1936; by v; run;

proc means data=combined1936; var est_a_s est_a_surv_tte;
output out=sd1936_ (keep= est_a_s_std est_a_surv_tte_std)
std=est_a_s_std est_a_surv_tte_std;
run;
data sd1936;
    set sd1936_;
    v=1;
run;
proc sort data=sd; by v; run;

data bias_mse1936;
    merge bias1936 sd1936;
    by v;
    mse_s=bias_s_mean**2 + est_a_s_std**2;
    mse_surv_tte=bias_surv_tte_mean**2 + est_a_surv_tte_std**2;
    keep bias_s_mean bias_surv_tte_mean mse_s mse_surv_tte;
run;
proc print data=bias_mse1936; run;

```

*Printing the average ln prevalence ratio or average ln hazard ratio, the bias, and the mean squared error observed in the study population for the researcher's original cross-sectional


```

study as well as for the cohort study;
proc print data=a1936_s_mean; run;
proc print data=a1936_surv_tte_mean; run;
proc print data=bias_mse1936; run;

```

```

*****Simulating data to generate target population that is Black and White City X residents born in 1976 without changing the
cross-sectional study design where the study design corresponds to all Black and White City X residents who were born in 1976 and who
were living when study enrollment occurs in 2016*****;

```

```

data a1976;
  call streaminit(3);
  do n=10000; *Specifying that sample size of target population is 10,000;
    do k=1 to 500; *Specifying 500 simulations each with a target population sample size of 10,000;
      do i=1 to n;
        pA=.3; *Specifying that 30% of target population is Black and 70% is White;
        pU=.5; *Specifying that 50% of target population has a value of U=1;
        A=rand("bern",pA); *Generating variable A that is an indicator for Black (versus White) race;
        U=rand("bern",pU); *Generating variable U;
        AU=A*U;
        Unif_t=rand('UNIFORM');
        Unif_c=rand('UNIFORM');

        *Following the method of Bender et al SIM 2005 to generate Z infection that follows an exponential
        distribution;
        lambda_t=0.015; *Specifying that the baseline hazard function for Z infection (i.e., a White person with
        U=0) corresponds to mean time to infection of 67 years old where 1/lambda_t is the mean
        time to infection;
        gammat1=0; *Specifying that the relative hazard of Z infection as a function of A is 1.0;
        gammat2=0.693; *Specifying that the relative hazard of Z infection as a function of U is 2.0;
        T=- (log(unif_t)/(lambda_t*exp(gammat1*A+gammat2*U))); *Generating the variable T that is the time from
        birth to Z infection;

        *Following the method of Bender et al SIM 2005 to generate mortality that follows a Gompertz
        distribution;
        alpha=3.14159/(sqrt(6)*20); *Specifying a standard deviation of 20 years for life expectancy for White
        residents with U=0 who were born in 1976;
        lambda_c=alpha*exp(-0.5772-alpha*83); *Specifying a mean life expectancy of 83 years for White residents
        with U=0 born in 1976;
        gammac1=0.693; *Specifying that the relative hazard of death as a function of A is 2.00 when U=0;
        gammac2=0.693; *Specifying that the relative hazard of death as a function of U is 2.00 when A=0;
        gammac3=0.223; *Specifying that the relative hazard of death is 2.50 as a function of A when U=1 and

```

```

                2.50 as a function of U when A=1;
inverse_alpha=1/alpha;
a_log_unif=alpha*(log(Unif_c));
lambda_exp_beta_x=lambda_c*exp(gammac1*A+gammac2*U+gammac3*AU);
C=inverse_alpha*log(1-a_log_unif/lambda_exp_beta_x); *Generating the variable C that is the time from
                                                    birth to death;

*Rounding (up) times to Z infection and death to be integers to match the time scale of follow up which
is in integers;
T_ceil=ceil(T);
T=T_ceil;
C_ceil=ceil(C);
C=C_ceil;

*Creating a variable for the observed follow up time (i.e., Ymin) as well as the corresponding indicator
of having Z infection at the end of follow up in the presence of censoring due to death as well as
administrative censoring (i.e., Dobs);
Ymin=min(T,C,40);
Dobs=0;
if T<=40 then do; Dobs=1; end;
if C<T<=40 then do; Dobs=0; end;

*Creating indicator, S, of being alive in 2016 to be included in the researcher's cross-sectional study;
S=0;
if C>40 then do; S=1; end;
output;
end;
end;
end;
run;

*Sorting data by simulation;
proc sort data=a1976; by k; run;

*Calculating the average proportion of individuals in target population who are alive when enrollment begins in cross-sectional study;
proc means data=a1976 noprint; by k; var s; output out=s_mean1976 mean=s_mean; run;
proc means data=s_mean1976 n min max mean p25 median p75; var s_mean; run;

*Calculating the average ln prevalence ratio observed in study population for effect of Black race on Z infection;
proc genmod data=a1976;
by k;

```

```

        class i;
        where s=1;
        model Dobs=a /link=log dist=poisson; ods output GEEmpPEst=a1976_s;
        repeated subject=i/type=ind;
run;
data a1976_s_;
    set a1976_s;
    if parm='A';
    keep k parm estimate;
run;
proc means data=a1976_s_; var estimate; output out=a1976_s_mean_ (keep=a_s_mean) mean=a_s_mean; run;
data a1976_s_mean_;
    set a1976_s_mean_;
run;
proc print data=a1976_s_mean_; run;

*Calculating the bias and mean squared error observed in study population for cross-sectional study;
data a1976_s_m;
    set a1976_s_;
    rename estimate=est_a_s;
    drop parm;
run;
proc sort data=a1976_s_m; by k; run;
data a1976_s_m_;
    set a1976_s_m;
    bias_s=est_a_s-0;
run;
proc print data=a1976_s_m_; run;

proc means data=a1976_s_m_; var bias_s;
output out=bias1976_ (keep= bias_s_mean)
mean=bias_s_mean;
run;
data bias1976;
    set bias1976_;
    v=1;
run;
proc sort data=bias1976; by v; run;

proc means data=a1976_s_m_; var est_a_s;
output out=sd1976_ (keep= est_a_s_std)

```

```
std=est_a_s_std;
run;
data sd1976;
    set sd1976_;
    v=1;
run;
proc sort data=sd1976; by v; run;

data bias_mse1976;
    merge bias1976 sd1976;
    by v;
    mse_s=bias_s_mean**2 + est_a_s_std**2;
    keep bias_s_mean mse_s;
run;
proc print data=bias_mse1976; run;

*Printing the average ln prevalence ratio, the bias, and the mean squared error observed in study population based on target
population born in 1976;
proc print data=a1976_s_mean; run;
proc print data=bias_mse1976; run;
```

REFERENCES

1. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine* 2005;24(11):1713-1723.
2. Arias E, Heron M, Xu JQ. United States life tables, 2012. National vital statistics reports. . Hyattsville, MD: National Center for Health Statistics, 2016,
3. VanderWeele TJ, Robinson WR. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* 2014;25(4):473-484. doi: 410.1097/EDE.000000000000105.