**Additional File 1: Data generating mechanisms for simulation study**

**A Comparison of Methods to Estimate the Survivor Average Causal Effect in the Presence of Missing Data: a Simulation Study**

Myra B. McGuinness, MBiostat, Jessica Kasza, PhD, Amalia Karahalios, PhD, Robyn H. Guymer, MBBS, PhD, Robert P. Finger, MBBS, PhD, Julie A. Simpson, PhD

Data generation and all statistical analyses were performed using Stata/SE version 15.1 (StataCorp LP, College Station, TX, USA) with the Mersenne Twister random number generator and a starting seed of 13073199.[1]

For each scenario, a dataset of 10,000 observations was simulated 1,200 times. The interval between the baseline and follow-up wave was assumed to be equal for all participants. The program was written to disregard any dataset generated without sufficient cases of age-related macular degeneration (AMD) for analysis and loop until 1,200 valid datasets had been generated. However no datasets were disregarded during this simulation study.

The parameters used to generate the data are provided in Table 1 below.

Reference

1. Matsumoto M, Nishimura T: **Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator**. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 1998, **8**(1):3-30.

Table S1: Parameters used to generate data for simulation study

|  |  |
| --- | --- |
| **Variable** | **Population model** |
| Sex  0 Male  1 Female |  |
| Mean-centred age |  |
| Genotype  0 Absent  1 Present |  |
| Residence  0 High socioeconomic area  1 Low socioeconomic area |  |
| Iron intake  0 Low iron intake  1 High iron intake |  |
| Potential survival when  0 Deceased  1 Alive |  |
| Potential survival when  0 Deceased  1 Alive | If compliant with the monotonicity assumption:  if  If monotonicity assumption violated: |
| Strata  0 Never-survivor  1 Always-survivor  2 Compliant-survivor  3 Defiant-survivor | if &  if &  if &  if & |
| Survival status when  0 Deceased  1 Alive | if ( & ) or ( & )  if ( & ) or ( & ) |
| Potential outcome when  0 Disease absent  1 Disease present | if |
| Potential outcome when  0 Disease absent  1 Disease present | if |
| Non-missing outcome data  0 Missing outcome data  1 Outcome data not missing | if  if |
| Age-related macular degeneration  0 Absent  1 Present | if &  if &  if or |
| Model coefficients: or ; , or ; or | |

Generate sex , mean-centred age , an indicator for poor health and an indicator of area of residence for 10,000 participants.

Generate potential survival for high iron intake conditional on , and .

**For scenarios compliant with the monotonicity assumption:**

Generate potential survival for low iron intake conditional on , and for participants who would survive following high iron intake (i.e. for those with )

**For scenarios that violate the monotonicity assumption:**

Generate potential survival under low iron intake conditional on , and for all participants.

Derive observed survival status according potential survival status and and exposure [

Generate potential outcomes for low iron intake conditional on , and for participants who would survive under low iron intake (i.e. for those with )

Generate potential outcomes for high iron intake conditional on , and for participants who would survive under high iron intake (i.e. for those with )

Allocate observed outcome for age-related macular degeneration deterministically according to , , and potential outcome (i.e. or

Generate the exposure, iron intake , conditional on and .

Generate an indicator of non-missing outcome data [ conditional on , and for survivors (i.e. those with ).

Allocate strata according to potential survival status and

Figure S1: Flow chart of data generation process. Datasets were generated and analysed 1,200 times under each scenario

**Stata code for simulation study**

// Generate single dataset

capture program drop sim\_iron\_data

program define sim\_iron\_data, rclass

syntax, vm(numlist max = 1) uy(numlist max = 1) uz(numlist max = 1)

clear

quietly set obs 10000

tempvar pz z0 z1 strata py y0 y1

\* Exposure and covariates

quietly gen U = uniform() < invlogit(ln(0.5)) // unmeasured genotype

quietly gen V1 = uniform() < invlogit(0) // sex

quietly gen V2 = floor(30 \* runiform() - 15) // Mean centred age [-14,15]

quietly gen A = uniform() < invlogit((ln(0.75)\*V1) + (ln(0.95)\*V2)) // iron

quietly gen D = uniform() < invlogit(0) // residence

\* Survival

quietly gen Z = .

quietly gen `pz' = ln(1.5) + (ln(2) \* V1) + (ln(0.95) \* V2) + (ln(`uz') \*U)

quietly gen `z1' = uniform() < invlogit(`pz' + ln(2))

quietly gen `z0' = 0

if `vm' == 0 {

quietly replace `z0' = uniform() < invlogit(`pz') if `z1' == 1

}

if `vm' == 1 { // Violation of monotonicity

quietly replace `z0' = uniform() < invlogit(`pz')

}

foreach b of numlist 0/1 {

quietly replace Z = `z`b'' if A == `b'

}

\* Strata

quietly gen `strata' = 2 // compliant-survivors

quietly replace `strata' = 0 if `z0' == 0 & `z1' == 0 // never-survivors

quietly replace `strata' = 1 if `z0' == 1 & `z1' == 1 // always-survivors

quietly replace `strata' = 3 if `z0' == 1 & `z1' == 0 // defiant-survivors

\* Outcome

quietly gen Y = .

quietly gen `py' = ln(0.11) + (ln(`uy') \* U) + (ln(1.5) \* V1)+(ln(1.05)\*V2)

quietly gen `y0' = uniform() < invlogit(`py') if `z0' == 1

quietly gen `y1' = uniform() < invlogit(`py' + ln(0.6)) if `z1' == 1

foreach b of numlist 0/1 {

quietly replace Y = `y`b'' if A == `b'

}

\* Empirical value of tau

quietly logistic `y1' i.`strata' i.(V1) V2 if inlist(`strata',1,2)

return scalar tau = exp(\_b[2.`strata'])

\* Non-missing data

quietly gen R = .

quietly replace R = uniform() < invlogit(ln(4) + (ln(1.4) \* A) + ///

(ln(2) \* V1) + (ln(0.9) \* V2) + (ln(0.8) \* D)) if Z == 1

quietly replace Y = . if R == 0

end

// Analyse single dataset

capture program drop analyse\_sace\_iron

program define analyse\_sace\_iron, eclass

version 14

preserve

tempvar ax sigmax as ar ssw g0 g1 h0 h1 gh0 gh1

tempname mat

\* Propensity for exposure

quietly logistic A i.V1 V2

quietly predict `ax'

sum `ax' if A == 1, meanonly

quietly gen `sigmax' = r(mean)

sum `ax' if A == 0, meanonly

quietly replace `sigmax' = r(mean) if e(sample)

quietly replace `ax' = 1 - `ax' if A == 0

\* Propensity for survival

quietly logistic Z i.(A V1) V2

quietly predict `as'

foreach i of numlist 0 1 {

quietly logistic Z i.V1 V2 if A == `i'

quietly predict `g`i''

}

\* Propensity for non-missing data

quietly logistic R i.(A V1) V2

quietly predict `ar' if e(sample)

\* Generate weights

quietly gen `ssw' = (`sigmax') /(`ax' \* `as' \* `ar')

\* Marginal structural models

quietly logistic Y i.(A V1) V2 [pweight = `ssw']

local essw = \_b[1.A]

\* Sensitivity approach

foreach b of numlist 0/1 {

sum `g`b'', meanonly

local nu\_`b' = r(mean) // marginal probability of survival

quietly logistic Y i.V1 V2 if A == `b'

quietly predict `h`b''

quietly gen `gh`b'' = `g`b'' \* `h`b''

sum `gh`b'' if Z == 1, meanonly

local xi\_`b' = r(mean) // marginal probability of both survival and AMD

}

local sa1 = ln((`xi\_1') \* (`nu\_0'-`xi\_0') / (`xi\_0' \* (`nu\_1'-xi\_1')))

foreach t of numlist 0.5 2 { // sensitivity parameters

if `t' == 2 {

local ta = 2

}

else {

local ta = 0

}

local q\_`ta' = sqrt(((`nu\_0' + `xi\_1') \* (1 - `t') + (`t' \* `nu\_1'))^2 ///

+ (4 \* `xi\_1' \* `nu\_0' \* (`t' - 1)))

local sa`ta' = ln(((`nu\_0' - `xi\_0') / `xi\_0') \* ///

((`nu\_0' + `xi\_1') \* (`t' - 1) - (`t' \* `nu\_1') + `q\_`ta'') / ///

((`nu\_0' - `xi\_1') \* (`t' - 1) + (`t' \* `nu\_1') - `q\_`ta''))

}

\* Covariate balance

quietly xi: pbalchk A V1 V2 U D if Y != ., nocatstandardize

matrix CBu = r(ssmeandiff) // unweighted

quietly xi: pbalchk A V1 V2 U D if Y != ., wt(`ssw') nocatstandardize

matrix CB`ssw' = r(ssmeandiff) // weighted

\* Output

matrix `mat' = `essw', `sa0', `sa1', `sa2', CBu, CBssw

ereturn post `mat'

restore

end

Table S2: Average standardised difference in covariate levels between exposure groups among attending survivors, generated with 1,200 repetitions per scenario.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Scenario**  **number** | **Violation**  **of**  **monotonicity** | **Effect of genotype** | | **Covariate** | | | | | | | |
| **Survival** | **Outcome** | **Sex** | | **Age** | | **Area of residence** | | **Unmeasured genotype** | |
|  |  | **Unweighted** | **Weighted** | **Unweighted** | **Weighted** | **Unweighted** | **Weighted** | **Unweighted** | **Weighted** |
| 1 | No | 0.5 | 0.5 | -0.28 | -0.02 | -0.26 | 0.03 | 0.01 | 0.00 | 0.11 | 0.12 |
| 2 | No | 0.5 | 1.0 | -0.28 | -0.02 | -0.26 | 0.03 | 0.00 | 0.00 | 0.11 | 0.12 |
| 3 | No | 0.5 | 2.0 | -0.28 | -0.02 | -0.26 | 0.03 | 0.01 | 0.00 | 0.11 | 0.12 |
| 4 | No | 2.0 | 0.5 | -0.25 | -0.01 | -0.29 | 0.02 | 0.01 | 0.00 | -0.08 | -0.09 |
| 5 | No | 2.0 | 1.0 | -0.25 | -0.01 | -0.30 | 0.01 | 0.01 | 0.01 | -0.08 | -0.09 |
| 6 | No | 2.0 | 2.0 | -0.25 | -0.01 | -0.30 | 0.01 | 0.01 | 0.00 | -0.08 | -0.09 |
| 7 | Yes | 0.5 | 0.5 | -0.21 | 0.00 | -0.34 | 0.00 | 0.01 | 0.00 | 0.05 | 0.04 |
| 8 | Yes | 0.5 | 1.0 | -0.21 | 0.00 | -0.34 | 0.00 | 0.01 | 0.01 | 0.05 | 0.04 |
| 9 | Yes | 0.5 | 2.0 | -0.21 | 0.00 | -0.34 | 0.00 | 0.01 | 0.01 | 0.05 | 0.04 |
| 10 | Yes | 2.0 | 0.5 | -0.21 | 0.00 | -0.35 | 0.00 | 0.01 | 0.01 | -0.04 | -0.04 |
| 11 | Yes | 2.0 | 1.0 | -0.21 | 0.00 | -0.35 | 0.00 | 0.01 | 0.00 | -0.04 | -0.04 |
| 12 | Yes | 2.0 | 2.0 | -0.21 | 0.00 | -0.35 | 0.00 | 0.01 | 0.00 | -0.04 | -0.04 |
| = odds ratio effect of unmeasured genotype on survival;  = odds ratio effect of unmeasured genotype on outcome | | | | | | | | | | | |

Table S3: Log odds ratio estimates from simulation study scenarios with a null effect of the unmeasured variable on the outcome.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Estimation method** | **Monotonicity** | | | | | | | |
| **Valid** | | | | **Violated** | | | |
| **Estimate\*** | **SE** | **SB (%)** | **MSE** | **Estimate\*** | **SE** | **SB (%)** | **MSE** |
| 1.0 | 0.5 | Average | 1.00 | 0.10 |  |  | 1.00 | 0.09 |  |  |
| Marginal structural model | -0.51 | 0.11 | -3 | 0.01 | -0.52 | 0.10 | -4 | 0.01 |
| Sensitivity analysis |  |  |  |  |  |  |  |  |
| SP = 0.5 | -0.31 | 0.10 | 195 | 0.05 | -0.41 | 0.10 | 106 | 0.02 |
| SP = 1 | -0.52 | 0.10 | -5 | 0.01 | -0.51 | 0.10 | 6 | 0.01 |
| SP = 2 | -0.81 | 0.10 | -296 | 0.10 | -0.66 | 0.10 | -157 | 0.03 |
| 1.0 | 2.0 | Average | 1.00 | 0.09 |  |  | 1.00 | 0.10 |  |  |
| Marginal structural model | -0.51 | 0.10 | -1 | 0.01 | -0.51 | 0.10 | 3 | 0.01 |
| Sensitivity analysis |  |  |  |  |  |  |  |  |
| SP = 0.5 | -0.35 | 0.09 | 173 | 0.03 | -0.42 | 0.09 | 99 | 0.02 |
| SP = 1 | -0.50 | 0.09 | 9 | 0.01 | -0.50 | 0.09 | 17 | 0.01 |
| SP = 2 | -0.73 | 0.09 | -242 | 0.06 | -0.62 | 0.09 | -117 | 0.02 |
| \* Estimates of the log odds ratio have been averaged over 1,200 simulated datasets from each scenario.  MSE = mean square error; SACE = survivor average causal effect; SB = standardized bias as a percentage; SE = empirical standard error; SP = sensitivity parameter.  is the odds ratio effect of *U* on the outcome. is the odds ratio effect of U on survival. is the ratio of the odds of the outcome following high iron intake between compliant-survivors and always-survivors. True SACE log odds ratio = ln(0.6) = -0.511 | | | | | | | | | | |



Figure S2: Estimates from simulation study scenarios with a null effect of the unmeasured variable on the outcome. Estimated using 10,000 observations simulated 1,200 times for each scenario. The odds ratio effect of the unmeasured variable on the outcome , , was set to 1.0. The black line represents the true exposure effect (on the log odds ratio scale) of -0.51. is the odds ratio of the unmeasured variable, , on survival, .

**Proof for Equation 12**

The notation for the following equations is as defined in the manuscript. Measured exposure-outcome confounders are denoted .

Note, in the illustrative example, the exposure level given by (high iron intake) is considered to be predictive of greater levels of survival and a lower probability of an undesirable outcome (age-related macular degeneration). Whereas the in the paper by Egleston (2007) referred to below, an exposure level equal to one (indicating vision loss) is hypothesised to be associated with lower rates of survival and a greater risk of the undesirable outcome (emotional distress). Therefore, there are differences in the notation between the two papers.

Under the assumptions of ignorable treatment assignment (conditional on measured confounders) and outcome values missing at random:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | (From Egleston 2007, page 537) | (1) |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | (From Egleston 2007, page 537) | (2) |

Under the assumptions of conditional ignorability, and outcome values missing at random:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  | (From Equation 4.1, Egleston 2007, page 533) | (3) |

Under the monotonicity assumption, this value is observable because all survivors with exposure are considered to be always-survivors.

|  |  |  |
| --- | --- | --- |
|  | (From Equation 4.8, Egleston 2007, page 536) | (4) |

when :

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  | (From Egleston 2007, page 536) | (5) |

Under the assumption of monotonicity, this value is also observable because all survivors with exposure will have the same distribution of , regardless of stratum.

|  |  |  |
| --- | --- | --- |
|  | (From Equation 3.1, Egleston 2007, page 530) | (6) |

when :

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  | (Equation 12) | (7) |