Online Appendix

Contents

eAppendix 1: Simulation study

We here provide a small simulation study to support the proposed ascertainment probability weighting (APW) estimator, including an assessment of bias, standard deviation (SD), root mean squared error (RMSE), and confidence interval (CI) coverage for the bootstrap procedure used in the empirical example of the paper, and compare the results with conventional inverse probability weighting (IPW). Code to reproduce the simulation results is available at https://osf.io/8vwsu/.

In this simulation, we use the *daggity* package to simulate binary variables following the same code and directed acyclic graph as in eAppendix 2. From this model, we generate a population of size *N*=1,000,000, from which we draw 1,000 random samples of size *n*=1,000.

Characteristics of the population, including the true values of $\Pr(Y(1) = 1)$, $\Pr(Y(0) = 1)$, $\ln RR$, and RD, can be found in Table A1.

Table A1. Characteristics of the simulated target population (N = 1,000,000).

Characteristic	Value
Data distribution	
$\Pr(Y = 1)$	0.5
$\Pr(X = 1)$	0.5
$\Pr(Z = 1)$	0.5
$\Pr(Y_1^* = 1)$	0.248
$\Pr(Y_2^* = 1)$	0.247
$\Pr(Y_1^* = 1, Y_2^* = 1)$	0.137
$\Pr(Y^* = 1)$	0.358
Target parameters	
$\Pr(Y(1) = 1)$	0.38
$\Pr(Y(0) = 1)$	0.62
$\ln RR$	-0.49
RD	-0.24

The simulation results are presented in Table A2. As expected, conventional IPW performs poorly in terms of bias, RMSE and CI coverage due to under-ascertainment bias. The simulation verifies our theoretical results for the APW estimator, showing that the estimator is consistent in a situation where all its assumptions are met. It also shows that the proposed application of the percentile bootstrap performs well in terms of CI coverage. The SD is consistently larger for APW compared to IPW, which is expected given the extra variability introduced when estimating the ascertainment probabilities.

Table A2. Simulation results comparing conventional inverse probability weighting (IPW) to our ascertainment probability weighting (APW) estimator in a simulated scenario with under-ascertained outcomes and where all assumptions of the APW estimator hold.

Estimate	**Bias**	**SD**	**RMSE**	**95% CI Coverage**
IPW				
$\widehat{\Pr}(Y(1) = 1)$	-0.07	0.02	0.07	0.14
$\widehat{\Pr}(Y(0) = 1)$	-0.21	0.02	0.21	0.00
$\widehat{\ln RR}$	0.22	0.09	0.24	0.27
$\widehat{RD}$	0.14	0.03	0.15	0.01
APW				
$\widehat{\Pr}(Y(1) = 1)$	0.01	0.03	0.04	0.95
$\widehat{\Pr}(Y(0) = 1)$	0.00	0.06	0.06	0.96
$\widehat{\ln RR}$	0.01	0.12	0.12	0.95
$\widehat{RD}$	0.00	0.06	0.06	0.95

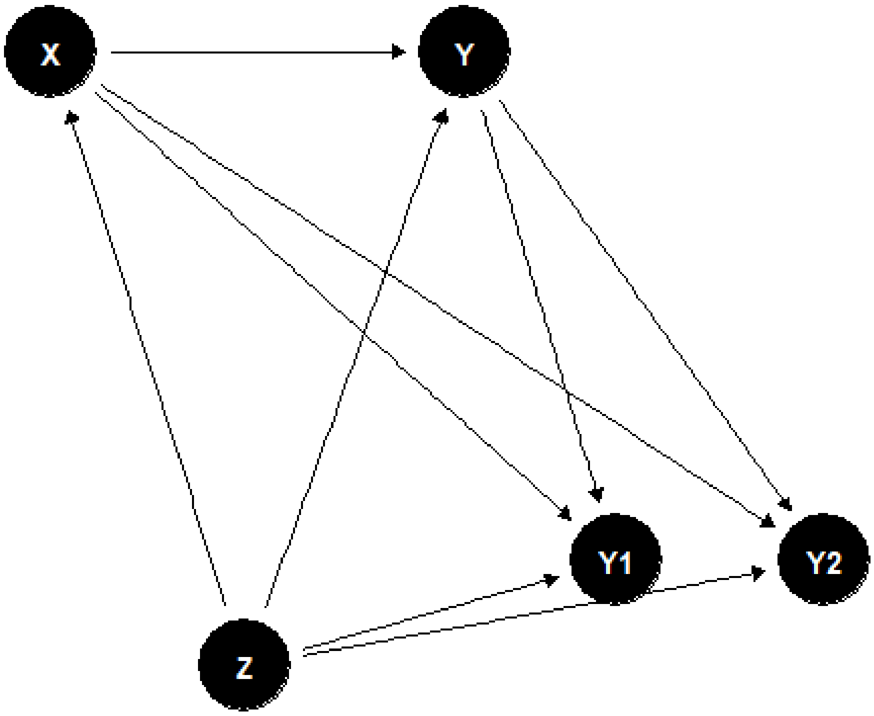
eAppendix 2: Example R code

The following code and output can be used to generate individual-level data similar to the numeric example presented in the main body of the paper. The code and output present below reflects unedited output from R using Quarto.

Load libraries

```
library(dagitty)

## Warning: package 'dagitty' was built under R version 4.2.3

library(ggdag)

## Warning: package 'ggdag' was built under R version 4.2.3

##
## Attaching package: 'ggdag'

## The following object is masked from 'package:stats':
##
##     filter

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Simulate logistic data according to the following directed acyclic graph



```r
simdag <- dagitty('dag {
  bb="0,0,1,1"
  X [pos="0.190,0.459"]
  Y [pos="0.329,0.459"]
  Y1 [pos="0.419,0.461"]
  Y2 [pos="0.474,0.461"]
  Z [pos="0.502,0.320"]
  Z -> X [beta=-0.5]
  Z -> Y [beta=0.3]
  X -> Y [beta=0.5]
  X -> Y1 [beta=-0.3]
  X -> Y2 [beta=-0.4]
  Z -> Y1 [beta=0.5]
  Z -> Y2 [beta=0.6]
}
')

#Simulate data
set.seed(12908)
simdat <- simulateLogistic(
  simdag,
  N = 100000
)

#Recode to 0 and 1
simdf <- data.frame(apply(simdat,2,function(x) ifelse(x=="-1",0,1)))

#Flip the X variable to fit example
```

```
simdf <- simdf %>% mutate(X = 1-X)

#Create observed (under-ascertained) outcomes
simdf <- simdf %>% mutate(Y1=Y*Y1,Y2=Y*Y2,
                          ystar_both = ifelse(Y1==1&Y2==1,1,0),
                          ystar = ifelse(Y1==1|Y2==1,1,0))
```

IPW, true values based on unobserved Y

```
# Estimate exposure propensity scores
emod <- glm(X~Z,simdf,family="binomial")
eprob <- predict(emod,type="response")

## IPW on known Y for reference (true target)
Pr_Y_x1 = weighted.mean(simdf[simdf$X==1,]$Y,w=1/eprob[simdf$X==1])
Pr_Y_x0 = weighted.mean(simdf[simdf$X==0,]$Y,w=1/(1-eprob)[simdf$X==0])
True_RR = Pr_Y_x1/Pr_Y_x0
True_RD = Pr_Y_x1-Pr_Y_x0
data.frame(Pr_Y_x1,Pr_Y_x0,True_RR,True_RD)

##     Pr_Y_x1   Pr_Y_x0   True_RR    True_RD
## 1 0.3793058 0.6212119 0.6105901 -0.2419061
```

IPW, biased values based on Y*

```
# IPW on Y* for reference
Pr_Ystar_x1 = weighted.mean(simdf[simdf$X==1,]$ystar,w=1/eprob[simdf$X==1])
Pr_Ystar_x0 = weighted.mean(simdf[simdf$X==0,]$ystar,w=1/(1-eprob)[simdf$X==0])
Biased_RR = Pr_Ystar_x1/Pr_Ystar_x0
Biased_RD = Pr_Ystar_x1-Pr_Ystar_x0
data.frame(Pr_Ystar_x1,Pr_Ystar_x0,Biased_RR,Biased_RD)

##   Pr_Ystar_x1 Pr_Ystar_x0 Biased_RR   Biased_RD
## 1   0.3133615   0.4119739 0.7606343 -0.09861241
```

Apply APW

```
# Ascertainment in both (j=1 and j=2)

j12_mod <- glm(ystar_both~X+Z,simdf%>%filter(ystar==1),family="binomial")
p12 <- predict(j12_mod,newdata=simdf,type="response") #Predicted probabilities

# Ascertainment in j=1

j1_mod <- glm(Y1~X+Z,simdf%>%filter(ystar==1),family="binomial")
p1 <- predict(j1_mod,newdata=simdf,type="response") #Predicted probabilities

# Ascertainment in j=2

j2_mod <- glm(Y2~X+Z,simdf%>%filter(ystar==1),family="binomial")
p2 <- predict(j2_mod,newdata=simdf,type="response") #Predicted probabilities

# Ascertainment probability estimates
```

```
aprob <- p12/(p1*p2)

# APW results

y_x1_apw <- mean(ifelse(simdf$ystar==1&simdf$X==1,1,0)/(aprob*eprob)) #PO with X=1
, APW
y_x0_apw <- mean(ifelse(simdf$ystar==1&simdf$X==0,1,0)/(aprob*(1-eprob))) #PO with
X=0, APW
APW_RR = y_x1_apw/y_x0_apw
APW_RD = y_x1_apw-y_x0_apw
data.frame(y_x1_apw,y_x0_apw,APW_RR,APW_RD)

##    y_x1_apw  y_x0_apw  APW_RR     APW_RD
## 1 0.3833061 0.6202666 0.61797 -0.2369604

# Compare to truth

error_unadj_x1 = Pr_Y_x1-Pr_Ystar_x1
error_unadj_x0 = Pr_Y_x0-Pr_Ystar_x0
error_unadj_RR = True_RR-Biased_RR
error_unadj_RD = True_RD-Biased_RD

error_APW_x1 = Pr_Y_x1-y_x1_apw
error_APW_x0 = Pr_Y_x0-y_x0_apw
error_APW_RR = True_RR-APW_RR
error_APW_RD = True_RD-APW_RD

data.frame(error_unadj_x1,error_unadj_x0,error_unadj_RR,error_unadj_RD)

##   error_unadj_x1 error_unadj_x0 error_unadj_RR error_unadj_RD
## 1     0.06594427      0.2092379     -0.1500443     -0.1432937

data.frame(error_APW_x1,error_APW_x0,error_APW_RR,error_APW_RD)

##   error_APW_x1 error_APW_x0 error_APW_RR error_APW_RD
## 1 -0.004000337 0.0009453005 -0.007379937 -0.004945637
```

Session information

```
## ─ Session info ───────────────────────────────────────────────────────────────
##  setting  value
##  version  R version 4.2.2 (2022-10-31 ucrt)
##  os       Windows 10 x64 (build 19045)
##  system   x86_64, mingw32
##  ui       RTerm
##  language (EN)
##  collate  Swedish_Sweden.utf8
##  ctype    Swedish_Sweden.utf8
##  tz       Europe/Berlin
##  date     2023-06-16
##  pandoc   2.19.2 @ C:/Program Files/RStudio/resources/app/bin/quarto/bin/tools/
(via rmarkdown)
##
```

```
## ─ Packages ───────────────────────────────────────────────────────────────────
##  package     * version date (UTC) lib source
##  cli           3.6.1   2023-03-23 [1] CRAN (R 4.2.3)
##  sessioninfo   1.2.2   2021-12-06 [1] CRAN (R 4.2.3)
##
##  [1] C:/Program Files/R/R-4.2.2/library
##
## ──────────────────────────────────────────────────────────────────────────────
```

Online Appendix for: A capture-recapture-based ascertainment probability weighting method for effect estimation with under-ascertained outcomes

eTables

eTable 1. Coding and categorization of the occupational category variable by The Swedish Standard Classification of Occupations 2012 (SSYK2012) codes and relation to our exposure variable in the empirical example.

Occupational category	**SSYK2012 codes**	**Exposure group (X_i)**
Teacher	'2330','2341','2342','2343','5311','5312','2351'	0
Social care	'5342','5343'	0
Service sector	'5221','5222','5223','5225','5226','5227','5230','7611', '9411','9412','9413','5131','5132'	0
Postal/delivery	'4420'	0
Transport services	'8321','8331'	0
Policy/security	'3360','5412','5413'	0
Cleaner	'9111'	0
Healthcare	'2211','2212','2213','2219','2260','2221','2222','2223', '2224','2226','2227','2228','2231','2232','2235','2239', '2271','2272','2273','2289','3250','5350','2284','5321', '5322','5323','5324','5325','5326','5330'	1

eTable 2. Variable definitions register data sources and period for each covariate in the empirical example.

Variable	**Categorization**	**Data source (register)**	**Period**
Age	Five-year age groups	Total Population Register	2019 (Dec 31)
Sex	Binary indicator for women	Total Population Register	N/A
Income	Disposable income categorized into national quartiles (Q1-Q4)	The longitudinal integrated database for health insurance and labour market studies (LISA)	2019
Birth country	Sweden or abroad, where abroad is further classified into lower income, lower-middle income, upper-middle income, and high income countries on the World Bank's classification in 2020	Total Population Register	N/A
Educational attainment	Primary education (max 9 years of schooling) Secondary education (max 12 years of schooling), Tertiary education (university-level education)	The longitudinal integrated database for health insurance and labour market studies (LISA)	2019
Marital status	Married Unmarried (incl. divorced, widowed)	The longitudinal integrated database for health insurance and labour market studies (LISA)	2019
Children in household	Categorized into 0, 1, 2, 3, 4+	Sum of all variables in LISA containing the number of children individual *i* has who live at home in various age groups	2019
Household type	Single parent (family type codes 31, 32, 41, or 42) Living with partner (family type codes 11, 12, 13, 21, 22, 23) Living alone without children (NOT any of the above codes)	Derived from the variable Family type (FamTyp) in LISA	2019
Pre-pandemic comorbidities	Scored one or more on the Charlson Comorbidity Index	Derived from ICD-10 codes in the National Patient Register	2019
Municipality of residence	Categorized into the 49 municipalities in Västra Götaland	Total Population Register	2019