SELECTION BIAS - Core concepts

Mabel Carabali

EBOH, McGill University

Updated: 2024-11-19

Expected Competencies

- Knowledge about design mechanisms for Selection Bias.
- Knows what would be the impact of selection bias on study designs

Objectives

- Identify sources and structure of selection bias.
- Differentiate between selection bias and other biases or statistical artifacts.
- Identify alternative to address selection bias.

Selection Bias

- Can arise in both randomized experiments and observational studies, and in both prospective and retrospective studies.

  > *"...the common consequence of selection bias is that the association between exposure and outcome among those selected for analysis differs from the association among those eligible."*

- **Type 1:** people are selected or not-selected into analysis
- **Type 2:** bias introduced by conditioning on a collider, but no one is left out of analysis.

Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25.

- Selection bias as used by H&R What If? does NOT refer to failure to generalize externally.
  - It refers to non-causal (spurious) association that occurs in the data set under analysis.

A Nice Resource

- Directed Acyclic Graphs
- A Taxonomy of Selection Bias
- Failure to Generalize
- Conditioning on a Collider
- Design Considerations
- Analytic Strategies for Addressing Selection Bias
- Sensitivity Analysis

Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias

Smith, L.H. Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias. Curr Epidemiol Rep 7, 179–189 (2020). https://doi.org/10.1007/s40471-020-00241-6

A Nice Resource for causal inference

> *"Selection bias can be further categorized into two broad types:*
>
> - *type 1 selection bias owing to restricting to one or more level(s) of a collider (or a descendant of a collider) and,*
> - *type 2 selection bias owing to restricting to one or more level(s) of an effect measure modifier."*

Toward a Clearer Definition of Selection Bias When Estimating Causal Effects

Lu, H. , Cole, S. , Howe, C. & Westreich, D. (2022). Toward a Clearer Definition of Selection Bias When Estimating Causal Effects. Epidemiology, 33 (5), 699-706. doi: 10.1097/EDE.0000000000001516.

Chapter 8: H&R What if?

> *"We will use the term selection bias to refer to all biases that arise from conditioning on a common effect of two variables, one of which is either the treatment or a cause of treatment, and the other is either the outcome or a cause of the outcome."*

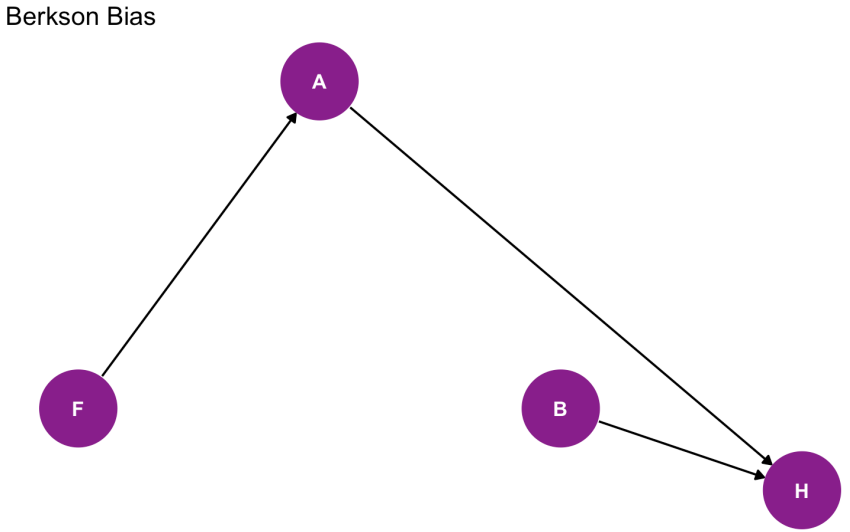
8.2: H&R What if? Practical examples of such structures:

- Differential loss to follow-up (a.k.a informative censoring) - Missing data bias, nonresponse bias
- Healthy worker bias
- Self-selection bias, volunteer bias
- Selection affected by treatment received before study entry

Berkson bias

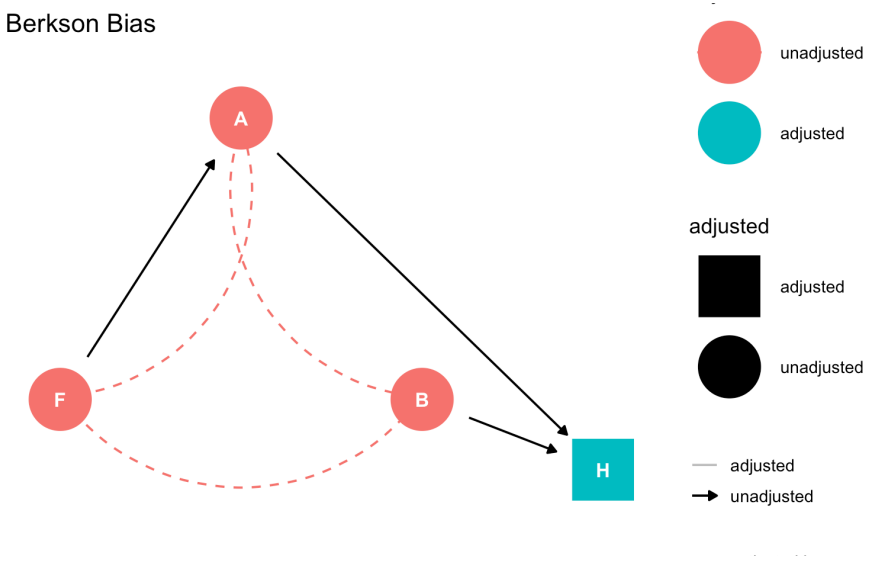
Berkson (1946, 1955) noted that two diseases (A and B) that are not associated in the population could be associated among hospitalized patients when both diseases affect the probability of hospital admission (H).

Berkson's bias can be seen to arise from conditioning on the common effect H of diseases A and B:



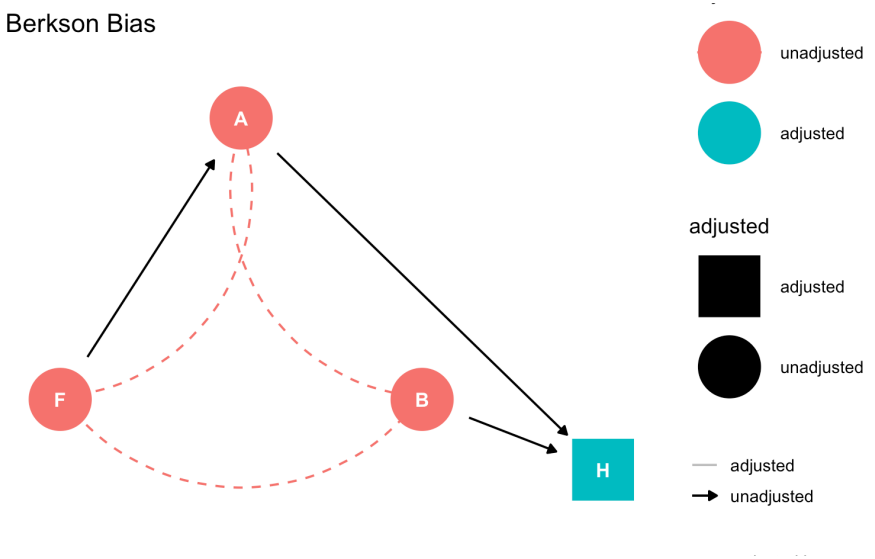
Berkson bias

In a case-control study where cases were hospitalized patients with disease B and controls were hospitalized patients with disease A, an exposure F that causes disease A would appear to be a risk factor for disease B.



That is, $OR_{FB|H=1}$ would differ from 1.0 even if F does not cause B.

Berkson bias



This bias occurs because conditioning on H induces a correlation between A and B, even if these are independent in the population.

Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004 Sep;15(5):615-25.

8.5. H&R What if? How to adjust for selection bias

- Inverse probability of censoring weights (IPCW) work exactly the same way as inverse of probability of treatment weights (IPTW). In fact, a typical study will have both weights.
- Note that positivity is NOT required for the C=1 stratum, since it is not our target population. Only for the C=0 stratum.

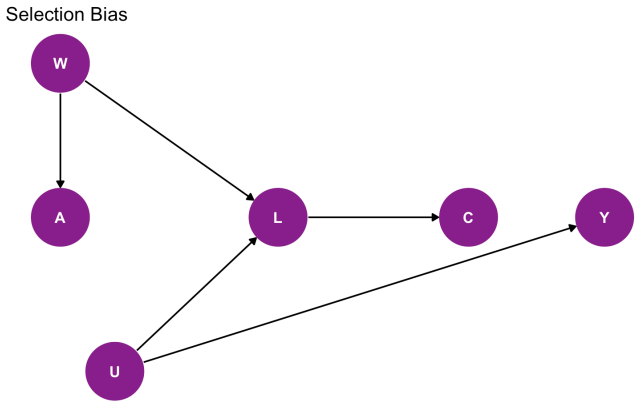
Important point:

For confounding, one had the choice of adjusting for the confounder by stratification or conditioning (conditional effect) or by standardization or g-methods (marginal effect).

- For collapsible effect measures, these are equivalent.

IPCW vs "Adjusting/Conditioning"

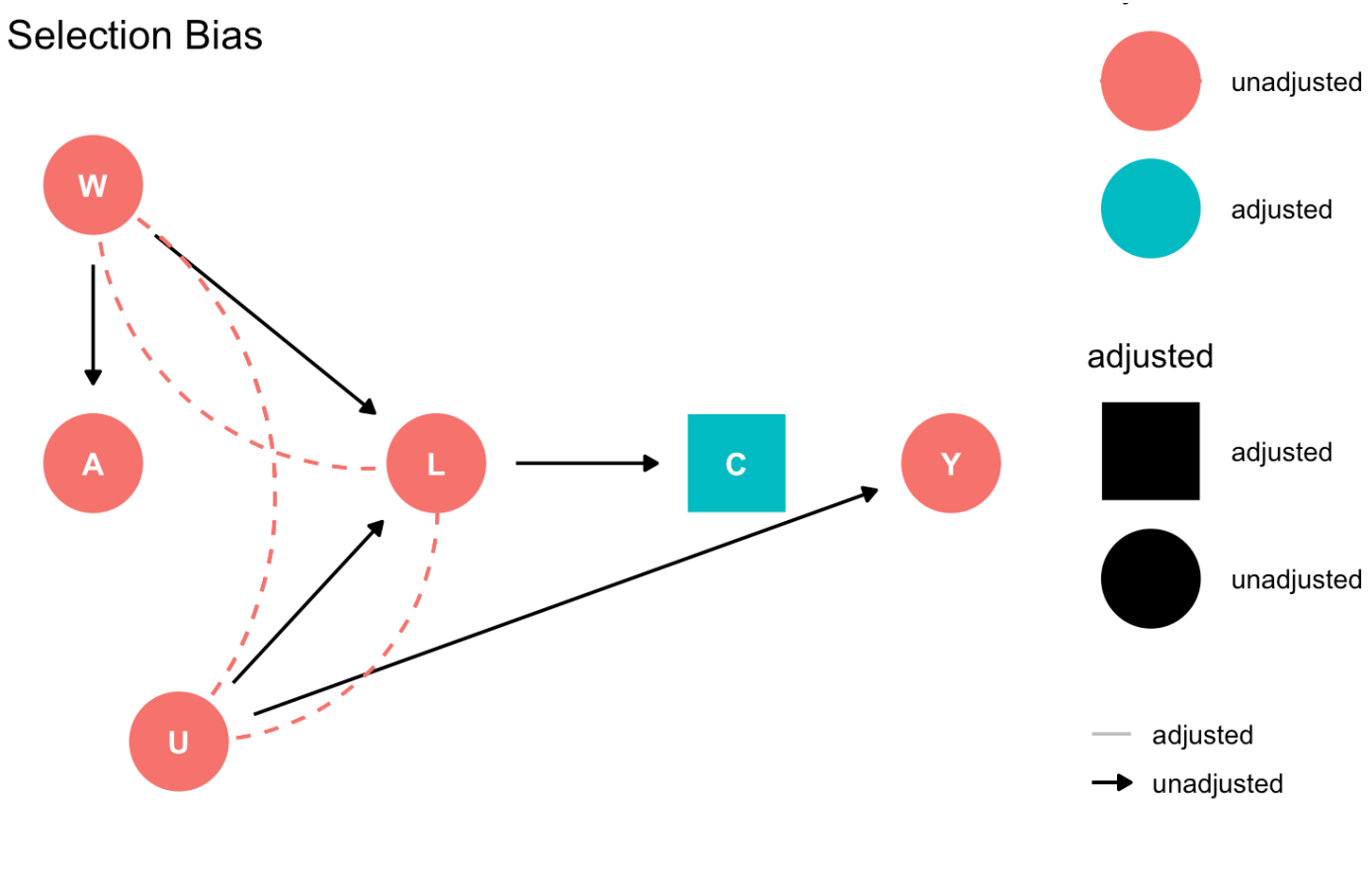
Let's consider this DAG (reproduced from H&R What If? Figure 8.6).



The other c-word, **Censoring**

- Bias due to informative censoring.
- Restricting the analysis to individuals with complete data (C = 0) may result in bias.
- Healthy worker bias
- Self-selection bias, volunteer bias

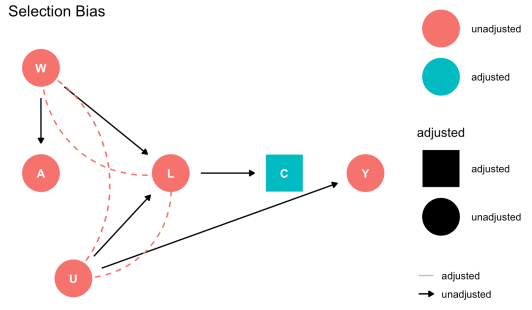
IPCW vs "Adjusting/Conditioning"



IPCW can be used to appropriately adjust for selection bias and conditioning will not work here and similar circumstances.

IPCW vs "Adjusting/Conditioning"

- Because IPCW is not based on estimating effect measures conditional on the measured covariates `L`, but rather on estimating unconditional effect measures after reweighting the subjects according to their treatment `A` and their values of `L`.



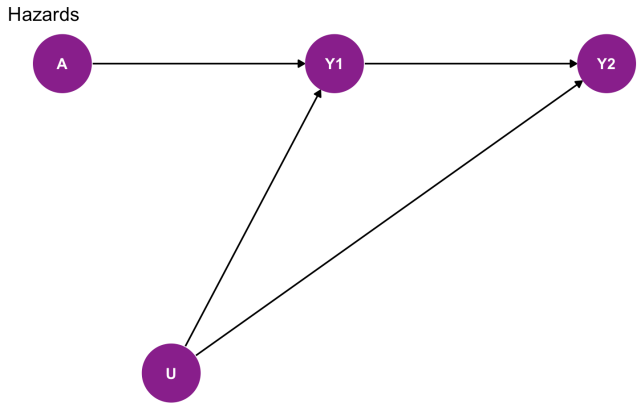
This is the first time `H&R` discuss a situation in which stratification cannot be used to validly compute the causal effect of treatment, even if the three conditions of exchangeability, positivity, and consistency hold. `Modified from Jay Kaufman EPIB-704-2021`

Hazard of the Hazards (Recall?)

Technical Point 8.1: H&R What if? The built-in selection bias of hazard ratios.

- In discrete time, the hazard of death at time 1 is the probability of dying at time 1 and thus the associational hazard ratio is the same as aRR_{AY_1}.
- However, the hazard at time 2 is the probability of dying at time 2 among those who survived past time 1.
- Thus, the associational hazard ratio at time 2 is then

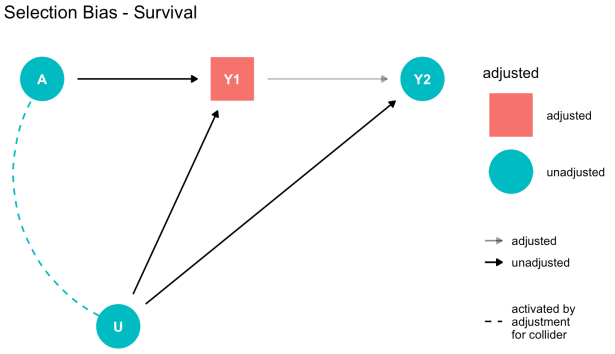
$$aRR_{AY_2|Y_1=0} = \left(\frac{Pr[Y_2=1|A=1,Y_1=0]}{Pr[Y_2=1|A=0,Y_1=0]} \right)$$



Hazard of the Hazards (Recall?)

Survival Produces an Unavoidable Selection Bias.

- Start out with a randomized trial so that all covariates are balanced at time 0.
- Once events occur, if you condition your estimate on having survived to the next time point, every other cause of disease must now be correlated with exposure.



Hazard of the Hazards (Recall?)

This is exactly why the HAZARD RATIO (the parameter estimated by a Cox Proportional Hazards Model) should not be used (unless the outcome is rare):

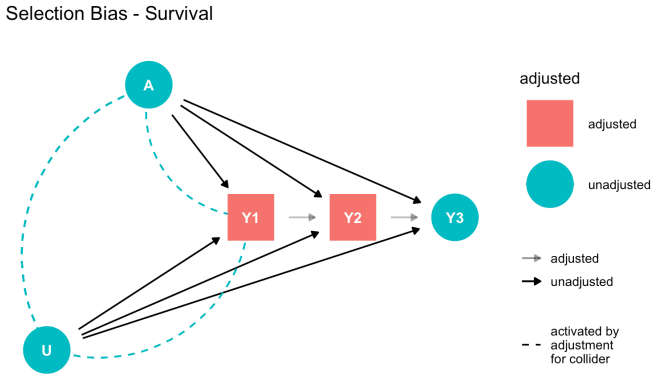
- The hazard of death at t_1 is the probability of dying at t_1.
- But the hazard at t_2 is the probability of dying at t_2 **among those who survived past t_1 .**

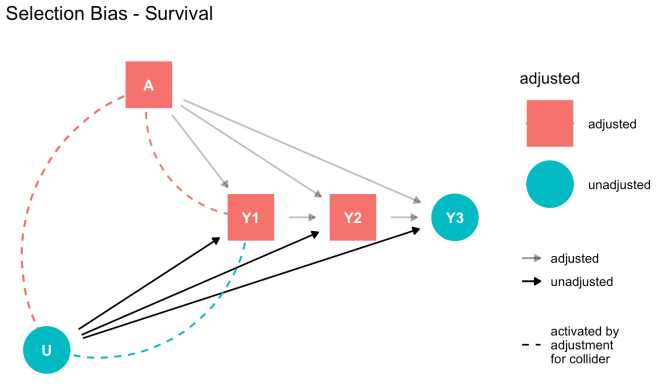
Treated survivors of t_1 differ in their distribution of U compared to untreated survivors of t_1, making this conditional measure confounded by U in a way that a marginal measure is not.

This concern applies to both observational studies and randomized experiments.

[Hernán MA, The hazards of hazard ratios. Epidemiology 2010;21(1):13-5.])
(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3653612/)

Hazard of the Hazards (Recall?)





Selection bias and Confounding

To the extent that confounding and selection bias are due to measured covariates C, these can be handled by inverse weighting (IPTW, IPCW)

- This is especially convenient for longitudinal data in which the confounder C may be effected by previous treatment X_t and may in turn influence the next dose of treatment X_{t+1}.
- It is also helpful in the longitudinal setting where the remaining cohort at each time t becomes increasingly selected.
- Reweighting the cohort by measured characteristics allows remaining subjects to proxy for the ones that are missing.

IPTW vs IPCW

IPTW	IPCW
Pr(Treatment=1)	Pr(Censoring=1)
Weights balance out covars distribution by Exposure (a.k.a Treatment)	Weights "make-up"for the losses in the remaining sample
Exchangeability/ deals with confounding	Selection bias /Censoring
Could be used alone to estimate ATE	Needs other covars and potentially other weights (e.g.,IPTW) to estimate ATE

- Unstable or extreme weights could be problematic in both cases!

The strength and direction of selection bias.

Regardless of the direction of selection bias, another key issue is its magnitude.

- Biases that are not large enough to affect the conclusions of the study may be safely ignored in practice, whether the bias is upwards or downwards.
- Generally speaking, a large selection bias requires strong associations between the collider and both treatment and outcome.
- Greenland (2003) studied the magnitude of selection bias under the null, which he referred to as collider-stratification bias, in several scenarios.

Avoid Selection Bias

- Know your topic
- Know your population(s)
- Use the "best" design you can (remember "built-in" selection bias, e.g., case-control) [1]
  - Case control: Exposure is missing based on outcome: RD is biased
  - Case control: Outcome missing, based on outcome: RD & RR are biased -OR unbiased
- High participation and response rates
- Complete and objective ascertainment
- Complete follow-up

[1] Ch. 7. Case-Control Studies. Epidemiology by Design by Daniel Westreich or https://www.epidemiologybydesign.com/about

If none of these works... try analytic solutions BUT nothing could completely save a poor design

How else we can adjust for selection bias?

(selection) Relative Odds Ratio

If the marginal distribution is known, the estimation of the ratio between the Odds among participants (sub-sample) and the corresponding odds in the source population:

$$ROR = \frac{OR_{SubPop}}{OR_{TotPop}}$$

- $ROR = 1$, no bias; $ROR > 1$, overestimation bias; $ROR < 1$, underestimation bias. Equivalently:

$$OR_{Tot} = ORSub \times ROR$$

- Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. Acta Obstet Gynecol Scand 2018; 97: 407–416. https://doi.org/10.1111/aogs.13319

How else we can adjust for selection bias?

(selection) Relative Odds Ratio

Participation rates	Disease	No disease
Exposed	$\frac{a}{A}$	$\frac{b}{B}$
Not exposed	$\frac{c}{C}$	$\frac{d}{D}$
Small letters:	number of participants in the cell	
Capital letters:	size of source population in the cell	

$$\frac{\frac{a}{A}\cdot\frac{d}{D}}{\frac{b}{B}\cdot\frac{c}{C}} = \frac{a\cdot d}{b\cdot c} \Big/ \frac{A\cdot D}{B\cdot C} = \frac{OR_{part}}{OR_{source}} = ROR$$

The relative odds ratio (ROR) computed as the cross product ratio of the participation rates in the four exposure by outcome categories.

Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. Acta Obstet Gynecol Scand 2018; 97: 407–416. https://doi.org/10.1111/aogs.13319

How else we can to adjust for selection bias?

Correction by Projecting the Exposed Proportion Among Nonparticipants

- Use contingency table
- Depict/Classify participants and non participants by exposure groups
- Obtain the odds of participation by exposure status
- Estimate the OR among non-participants

- Ignore some non-participants characteristics
- Assumes same prevalence of exposure

How else we can to adjust for selection bias?

Correction Using Selection Proportions

Selection probabilities	Exposure = 1	Exposure = 0
Cases	$S_{case,1}$	$S_{case,0}$
Control	$S_{control,1}$	$S_{control,0}$

$$OR_{adj} = OR_{obs} \times \frac{S_{case}, S_{control,1}}{S_{case,1}, S_{control,0}}$$

Applying Quantitative Bias Analysis to Epidemiologic Data: Chapter 4 - Selection Bias Spreadsheet

QUESTIONS?

COMMENTS?

RECOMMENDATIONS?

Example from a vector-borne disease n=839 individuals followed for 3 months

Summary of covars distribution

Characteristic	N = 839[1]
agecat_c	
0	76 / 839 (9.1%)
1	81 / 839 (9.7%)
2	215 / 839 (26%)
3	343 / 839 (41%)
4	124 / 839 (15%)
sex	422 / 839 (50%)
ipd	72 / 839 (8.6%)
insurance2	
0	465 / 839 (55%)
1	344 / 839 (41%)
2	30 / 839 (3.6%)

[1] n / N (%)

Characteristic	N = 839[1]
comorb	109 / 784 (14%)
Unknown	55
overfever	
Median (Q1, Q3)	7.0 (5.0, 9.0)
myalgiadef	771 / 832 (93%)
Unknown	7
arthralgiadef	722 / 831 (87%)
Unknown	8
abdominalpain	326 / 834 (39%)
Unknown	5
a_leucopenia	179 / 710 (25%)
Unknown	129
outc	230 / 686 (34%)
Unknown	153

[1] n / N (%)

Example from a vector-borne disease: A descriptive exercise, no "treatment" or "intervention", just modeling the outcome as function of covariates (n=520 complete cases (`na.rm=T`), vs n=634 where some vars are imputed), `dataset without missingness, n=839`

```
mod0or<- glm(outc ~  factor(agecat_c) + sex +  ipd+ insurance2 + comorb + overfever +
               myalgiadef+ arthralgiadef + abdominalpain + leucopenia + lowplatel,
             data=sb.data1a, family= binomial)
```

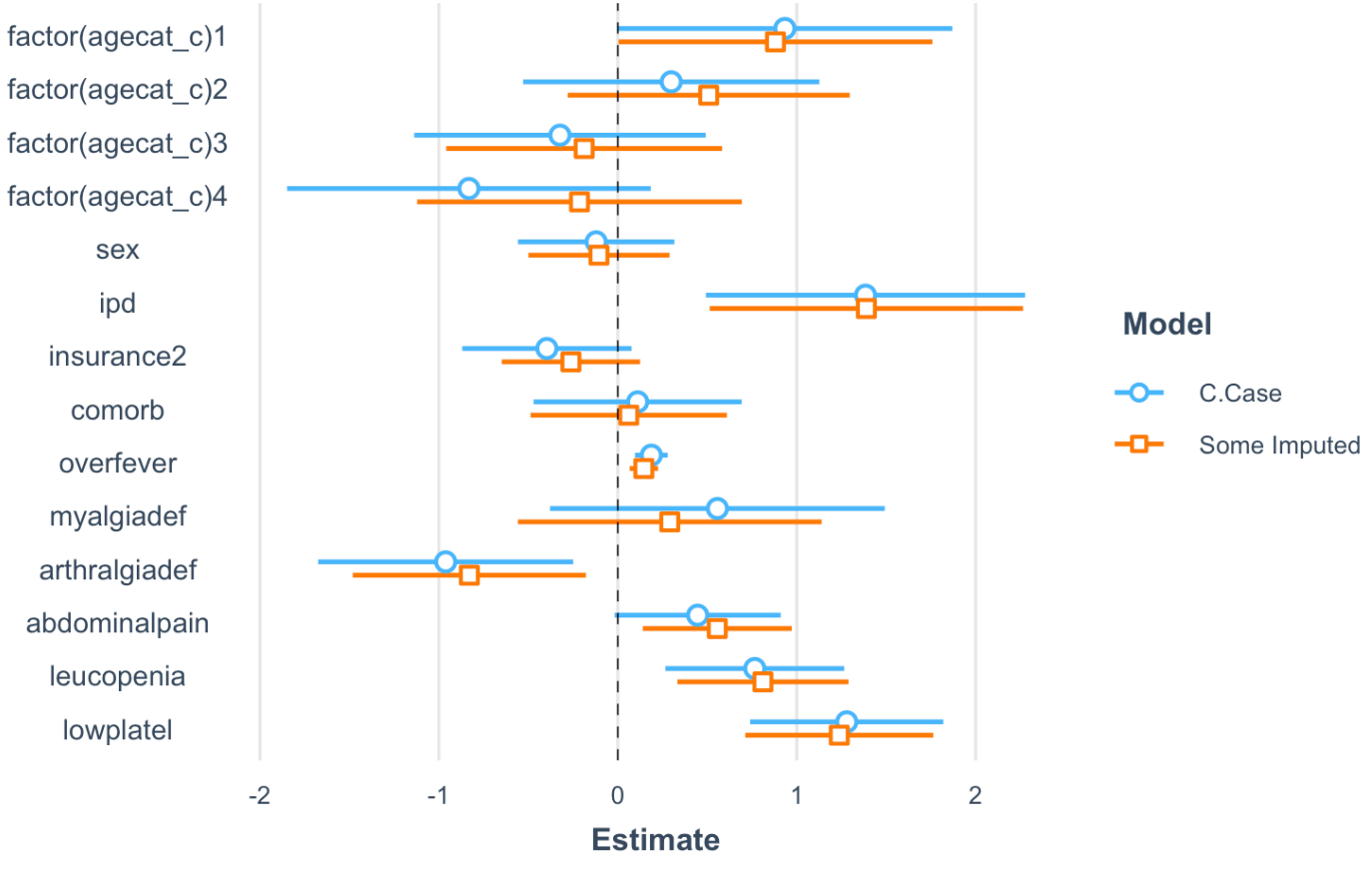
Complete Case Analysis

```
##                       OR 2.5 % 97.5 %
## (Intercept)         0.10 -3.64  -1.05
## factor(agecat_c)1 2.54  0.01   1.89
## factor(agecat_c)2 1.35 -0.51   1.15
## factor(agecat_c)3 0.72 -1.13   0.51
## factor(agecat_c)4 0.44 -1.86   0.18
## sex                 0.89 -0.56   0.32
## ipd                 3.99  0.53   2.33
## insurance2          0.67 -0.87   0.07
## comorb              1.12 -0.48   0.69
## overfever           1.21  0.10   0.28
## myalgiadef          1.74 -0.35   1.53
## arthralgiadef       0.38 -1.69  -0.26
## abdominalpain       1.56 -0.02   0.91
## leucopenia          2.15  0.26   1.27
## lowplatel           3.59  0.74   1.82
```

Analysis with Imputed data

```
##                       OR 2.5 % 97.5 %
## (Intercept)         0.12 -3.33  -1.03
## factor(agecat_c)1 2.41  0.02   1.78
## factor(agecat_c)2 1.66 -0.26   1.32
## factor(agecat_c)3 0.83 -0.94   0.61
## factor(agecat_c)4 0.81 -1.12   0.70
## sex                 0.90 -0.50   0.29
## ipd                 4.01  0.55   2.32
## insurance2          0.77 -0.65   0.12
## comorb              1.06 -0.50   0.60
## overfever           1.16  0.07   0.23
## myalgiadef          1.34 -0.53   1.17
## arthralgiadef       0.44 -1.49  -0.18
## abdominalpain       1.74  0.14   0.97
## leucopenia          2.25  0.33   1.29
## lowplatel           3.45  0.71   1.77
```

Example from a vector-borne disease Comparative plots



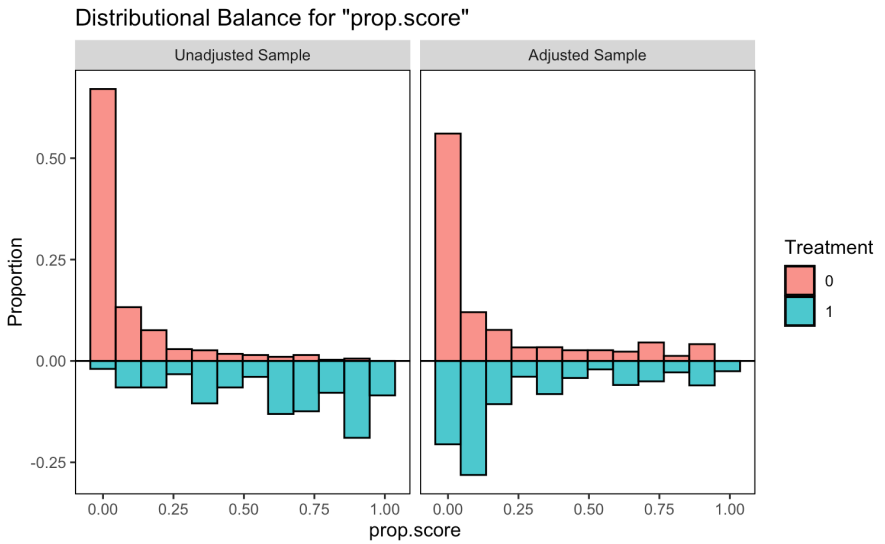
Example from a vector-borne disease, Using (IPCW) to correct for the missingness

Using the weightIt package we can upweigth the population to make up for the entire sample: (n=634 + IPCW), `dataset without missingness, n=839`

```
censor.form= "censor ~  factor(agecat_c) + sex + ipd+ insurance2 + comorb + overfever +
myalgiadef+  arthralgiadef + abdominalpain + leucopenia + lowplatel"

 CW1 <- weightit(as.formula(censor.form), data = sb.data2, method = "ps",# PScore weighting
                 estimand = "ATE")  # ATE estimand (not ATT)
 #summary(CW1)
```

as in PSTW and IPTW, we check the balance:



Example from a vector-borne disease (IPCW)

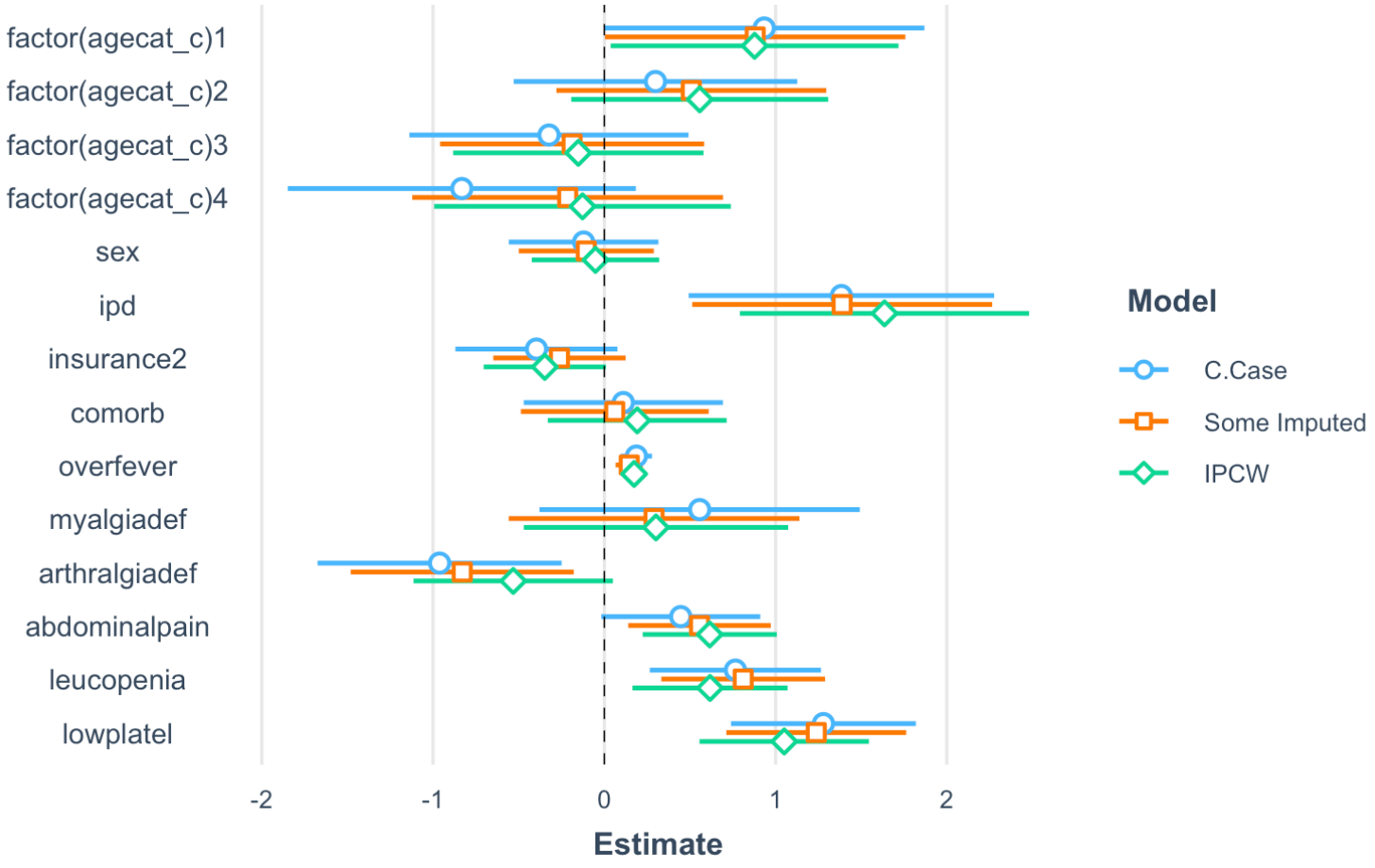
Here we use the weights of the censoring in the regression model:

```
 mod2or<- glm(outc ~  factor(agecat_c) + sex + ipd+ insurance2 + comorb + overfever +
                myalgiadef+  arthralgiadef + abdominalpain + leucopenia + lowplatel,
              weights = CW1$weights, data=sb.data1, family= binomial)
round(jtools::summ(mod2or, digits=2)$"coeftable", 2)  #;summary(mod2or);
```

```
##                     Est. S.E. z val.    p
## (Intercept)       -2.68 0.53  -5.07 0.00
## factor(agecat_c)1  0.88 0.43   2.05 0.04
## factor(agecat_c)2  0.56 0.38   1.45 0.15
## factor(agecat_c)3 -0.15 0.37  -0.41 0.68
## factor(agecat_c)4 -0.13 0.44  -0.29 0.77
## sex               -0.05 0.19  -0.27 0.79
## ipd                1.64 0.43   3.80 0.00
## insurance2        -0.35 0.18  -1.91 0.06
## comorb             0.19 0.27   0.72 0.47
## overfever          0.17 0.04   4.91 0.00
## myalgiadef         0.30 0.39   0.77 0.44
## arthralgiadef     -0.53 0.30  -1.79 0.07
## abdominalpain      0.62 0.20   3.08 0.00
## leucopenia         0.62 0.23   2.67 0.01
## lowplatel          1.05 0.25   4.16 0.00
```

Example from a vector-borne disease (IPCW)

```
plot_summs(mod0or, mod1or, mod2or, robust = list(FALSE,FALSE, FALSE),
           model.names = c("C.Case", "Some Imputed", "IPCW"))
```



Example from a vector-borne disease (IPCW) Estimating the PS of censoring (n=634 + IPCW), `dataset without missingness, n=839` **There are several ways to obtain weights as well**

```
modwt<- glm(censor ~ factor(agecat_c) + sex + ipd+ insurance2 + comorb + overfever +
              myalgiadef+  arthralgiadef + abdominalpain + leucopenia + lowplatel,
            data=sb.data2, family= binomial)
sb.data2$prob_wt <- predict(modwt, type = "response")
#summary(sb.data2$prob_wt)
sb.data2$cen_wt0<- 1/sb.data2$prob_wt
#summary(sb.data2$cen_wt0) #unstable weight
sb.data2$cen_wt1<- (mean(sb.data2$prob_wt))/ mean(1-sb.data2$prob_wt)/(1-sb.data2$prob_wt)
#summary(sb.data2$cen_wt1) #stabilize weight opt1
sb.data2$cen_wt2<- ifelse(sb.data2$censor==1, sb.data2$cen_wt0, 1/(1-sb.data2$prob_wt))
summary(sb.data2$cen_wt2)#stabilize weight opt2
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.002   1.030   1.618   1.227  44.950
```

Example from a vector-borne disease (IPCW)

Estimating the PS of censoring (n=634 + IPCW), dataset without missingness, n=839

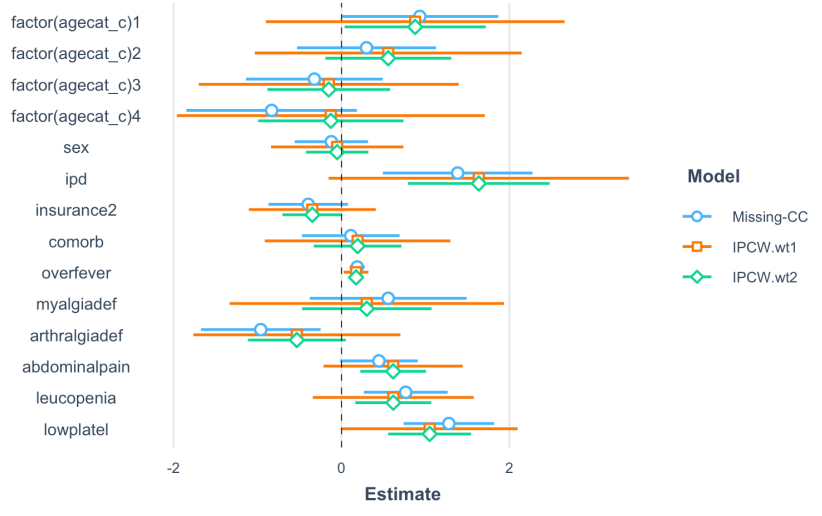
```
mod4or<- glm(outc ~  factor(agecat_c) + sex + ipd + insurance2 + comorb + overfever +
               myalgiadef+  arthralgiadef + abdominalpain + leucopenia + lowplatel,
             weights = sb.data2$cen_wt1, data=sb.data1, family= binomial)
mod4ort<-jtools::summ(mod4or)
round(mod4ort$coeftable,2) #;summary(mod4or);
```

```
##                    Est. S.E. z val.    p
## (Intercept)       -2.68 1.12  -2.39 0.02
## factor(agecat_c)1  0.88 0.91   0.97 0.33
## factor(agecat_c)2  0.56 0.81   0.69 0.49
## factor(agecat_c)3 -0.15 0.79  -0.19 0.85
## factor(agecat_c)4 -0.13 0.94  -0.14 0.89
## sex               -0.05 0.40  -0.13 0.90
## ipd                1.64 0.91   1.79 0.07
## insurance2        -0.35 0.39  -0.90 0.37
## comorb             0.19 0.56   0.34 0.73
## overfever          0.17 0.08   2.32 0.02
## myalgiadef         0.30 0.83   0.36 0.72
## arthralgiadef     -0.53 0.63  -0.85 0.40
## abdominalpain      0.62 0.42   1.46 0.15
## leucopenia         0.62 0.49   1.26 0.21
## lowplatel          1.05 0.53   1.96 0.05
```

Example from a vector-borne disease (IPCW)

```
##                    Est. S.E. z val.    p
## (Intercept)       -2.68 0.53  -5.07 0.00
## factor(agecat_c)1  0.88 0.43   2.05 0.04
## factor(agecat_c)2  0.56 0.38   1.45 0.15
## factor(agecat_c)3 -0.15 0.37  -0.41 0.68
## factor(agecat_c)4 -0.13 0.44  -0.29 0.77
## sex               -0.05 0.19  -0.27 0.79
## ipd                1.64 0.43   3.80 0.00
## insurance2        -0.35 0.18  -1.91 0.06
## comorb             0.19 0.27   0.72 0.47
## overfever          0.17 0.04   4.91 0.00
## myalgiadef         0.30 0.39   0.77 0.44
## arthralgiadef     -0.53 0.30  -1.79 0.07
## abdominalpain      0.62 0.20   3.08 0.00
## leucopenia         0.62 0.23   2.67 0.01
## lowplatel          1.05 0.25   4.16 0.00
```

Comparison Plots



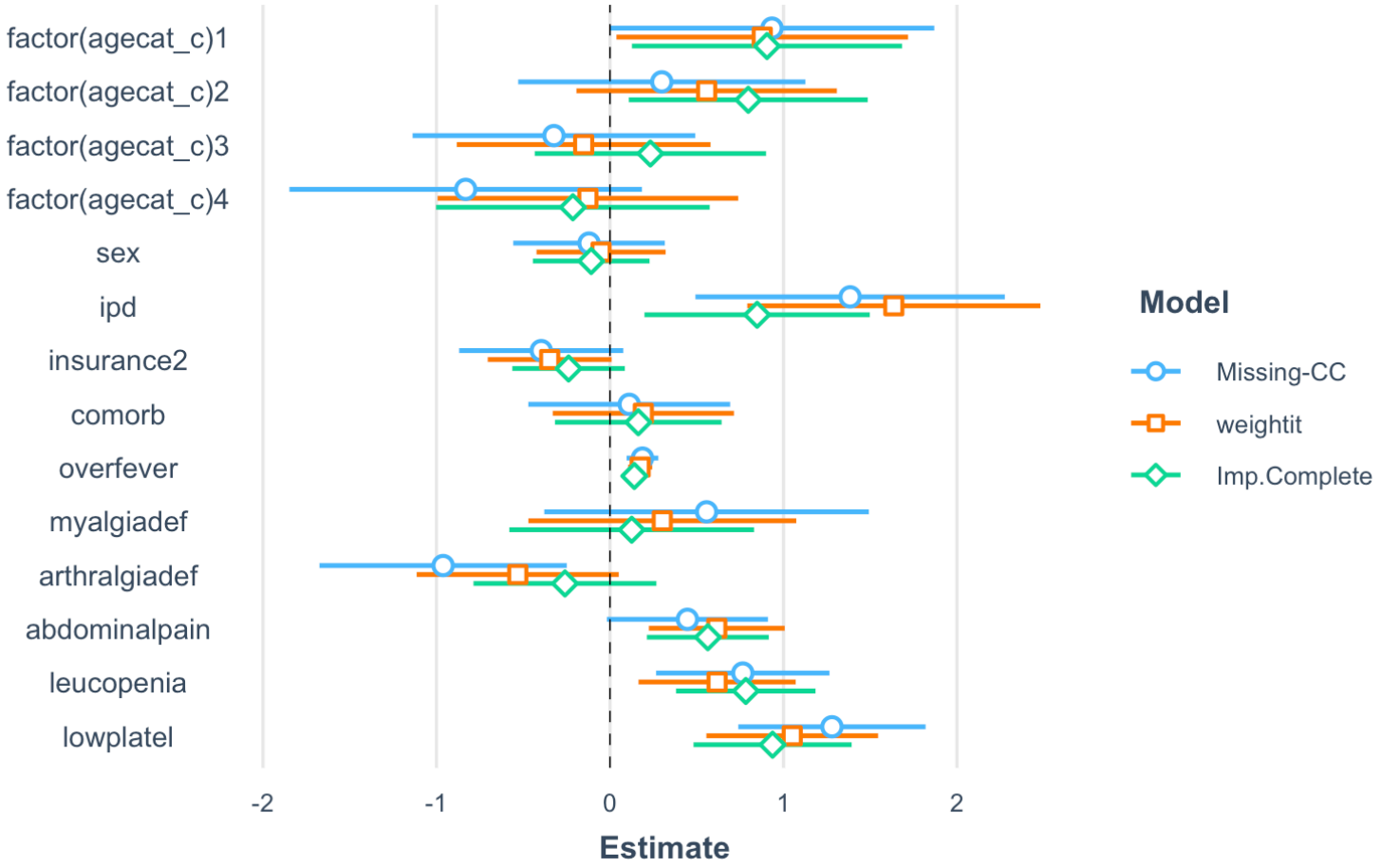
Example from a vector-borne disease (IPCW)

Data set without missing data, `complete dataset without missingness, n=839`

```
##                     Est. S.E. z val.    p
## (Intercept)       -2.65 0.51  -5.22 0.00
## factor(agecat_c)1  0.90 0.40   2.28 0.02
## factor(agecat_c)2  0.80 0.35   2.27 0.02
## factor(agecat_c)3  0.23 0.34   0.68 0.49
## factor(agecat_c)4 -0.21 0.40  -0.53 0.59
## sex               -0.11 0.17  -0.63 0.53
## ipd                0.85 0.33   2.56 0.01
## insurance2        -0.24 0.17  -1.44 0.15
## comorb             0.16 0.25   0.67 0.51
## overfever          0.14 0.03   4.59 0.00
## myalgiadef         0.13 0.36   0.35 0.73
## arthralgiadef     -0.26 0.27  -0.97 0.33
## abdominalpain      0.56 0.18   3.14 0.00
## leucopenia         0.78 0.21   3.82 0.00
## lowplatel          0.94 0.23   4.03 0.00
```

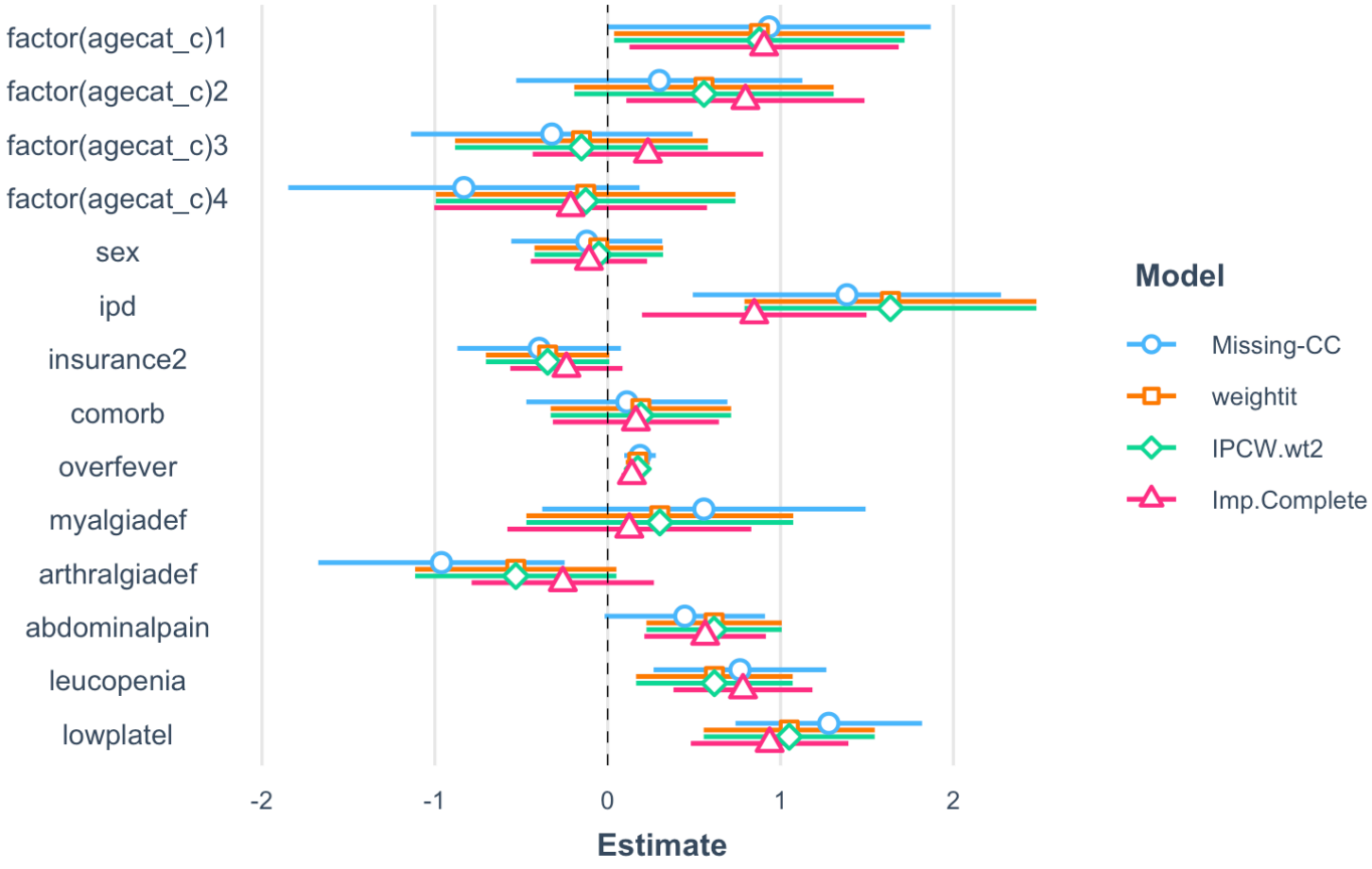
Example from a vector-borne disease (IPCW)

Comparison Plots



Example from a vector-borne disease (IPCW)

Comparison Plots



The Bayesian way

```
mod0<- stan_glm(outc ~factor(agecat_c) +
                  sex + sgss + comorb + overf
                  myalgiadef+  arthralgiadef
                  abdominalpain + a_leucopeni
                  a_lowplatel, data=sb.data1,
                  family= binomial(link = "lo
                  refresh=0)
#print(mod0, digits=2)#;summary(mod0)
mod1<- stan_glm(outc ~ factor(agecat_c) + s
                ipd + insurance2 + comorb + o
                myalgiadef+ arthralgiadef +
                abdominalpain + leucopenia +
                lowplatel, weights = cen_wt2,
                data=sb.data2,
                family= binomial(link = "log"
                refresh=0)
```

```
print(mod1, digits=2, detail = F)#;summary(
```

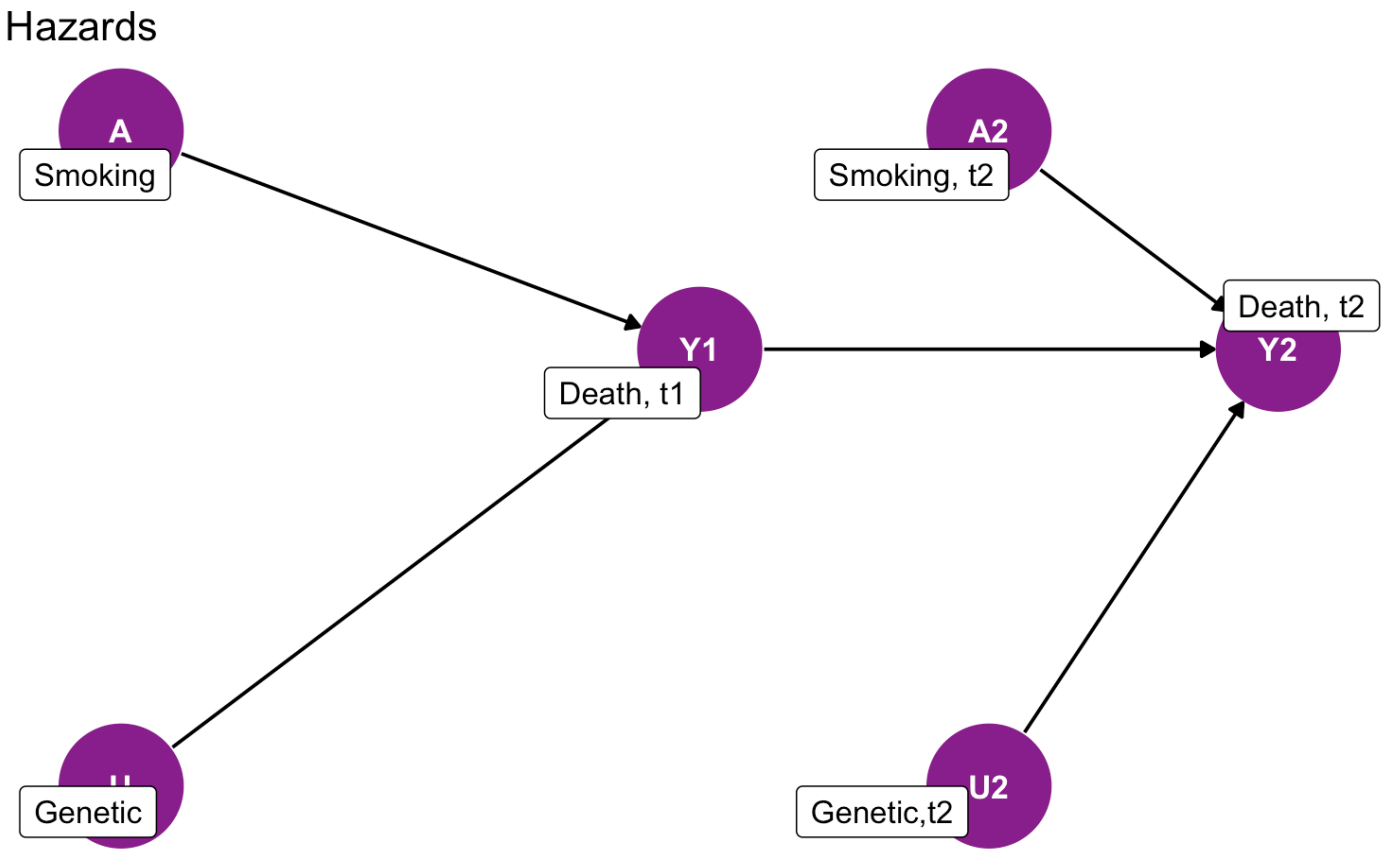
```
##                     Median MAD_SD
## (Intercept)         -1.85   0.28
## factor(agecat_c)1    0.37   0.21
## factor(agecat_c)2    0.16   0.20
## factor(agecat_c)3   -0.01   0.21
## factor(agecat_c)4   -0.12   0.24
## sex                 -0.02   0.08
## ipd                  0.17   0.11
## insurance2          -0.11   0.07
## comorb               0.15   0.10
## overfever            0.03   0.01
## myalgiadef           0.08   0.20
## arthralgiadef       -0.14   0.09
## abdominalpain        0.26   0.11
## leucopenia           0.13   0.08
## lowplatel            0.65   0.13
```

The Bayesian way

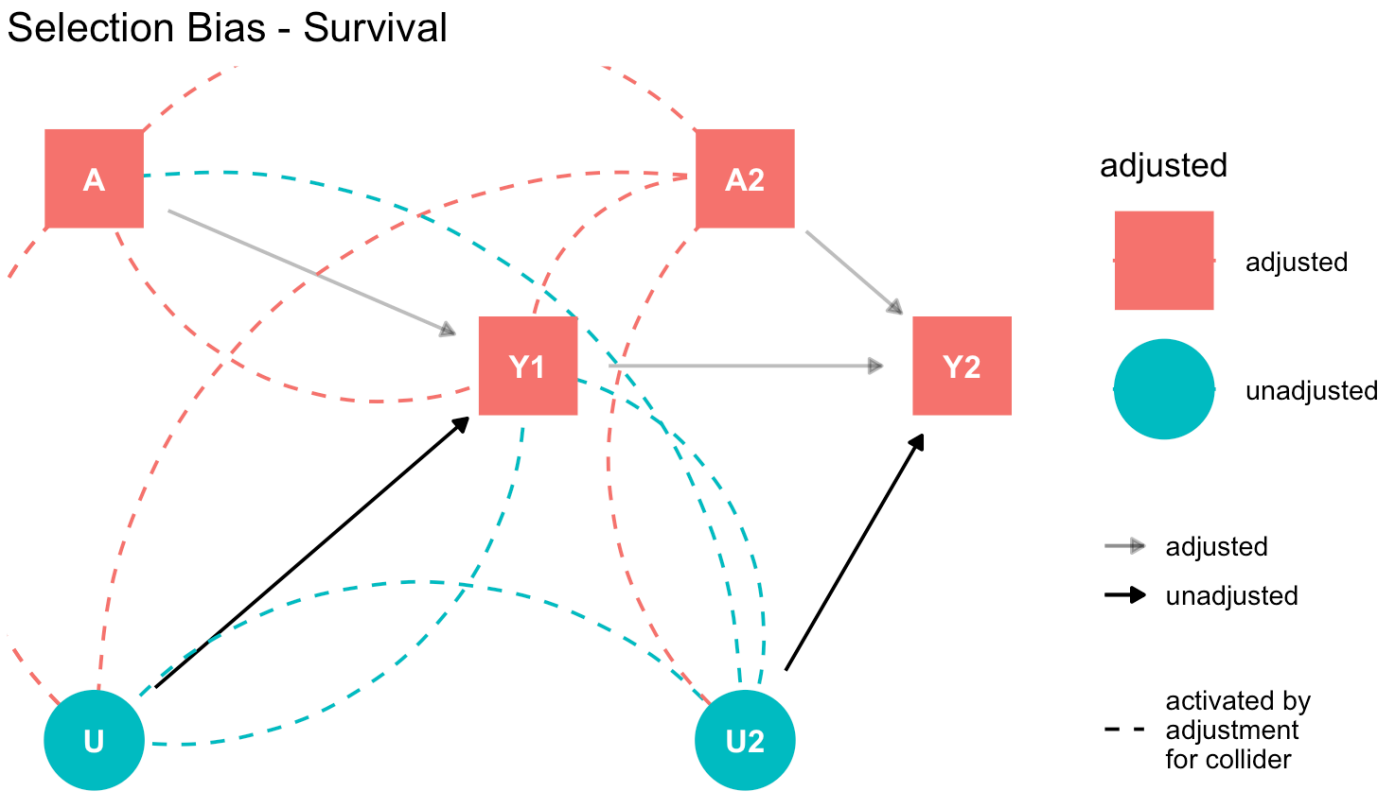
```
mod3RR<- stan_glm(DENV_DICHOT ~  factor(agecat_c) + sex + ipd +insurance2 + comorb +
                    overfever + myalgiadef+  arthralgiadef + abdominalpain + leucopenia +
                    lowplatel,  data=sb.data2, family= binomial(link = "log"), refresh=0)
print(mod3RR, digits=2, detail = F)
```

```
##                    Median MAD_SD
## (Intercept)        -1.89   0.30
## factor(agecat_c)1   0.49   0.25
## factor(agecat_c)2   0.39   0.23
## factor(agecat_c)3   0.20   0.22
## factor(agecat_c)4  -0.11   0.27
## sex                -0.05   0.08
## ipd                 0.07   0.10
## insurance2         -0.11   0.08
## comorb              0.14   0.10
## overfever           0.03   0.01
## myalgiadef          0.05   0.19
## arthralgiadef      -0.14   0.10
## abdominalpain       0.18   0.10
## leucopenia          0.23   0.10
## lowplatel           0.55   0.13
```

Extra DAGs


Hazards
A
Smoking
A2
Smoking, t2
Death, t2
Y1
Death, t1
Y2
Genetic
U2
Genetic,t2

Extra DAGs


Selection Bias - Survival
A
A2
Y1
Y2
U
U2
adjusted
adjusted
unadjusted
adjusted
unadjusted
activated by
adjustment
for collider