Preoperative Atelectasis

Part 8: Statistical Modelling of SpO2 95% vs >95%)

Javier Mancilla Galindo

2024-04-08

Table of contents

Setup

Packages used

```
if (!require("pacman", quietly = TRUE)) {
  install.packages("pacman")
}


pacman::p_load(
  tidyverse, # Used for basic data handling and visualization.
  table1, #Used to add lables to variables.
  CBPS, #Used to calculate non-parametric propensity scores for IPW.
  WeightIt, #Used to calculate weights from propensity scores for IPW.
  mgcv, #Used to model non-linear relationships with a general additive model.
  boot, # Calculate bootstrap confidence intervals.
  gt, #Used to present a summary of the results of regression models.
  flextable, #Used to export tables.
  report #Used to cite packages used in this session.
)
```

Session and package dependencies

```
R version 4.3.3 (2024-02-29 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 11 x64 (build 22631)

Matrix products: default


locale:
[1] LC_COLLATE=Spanish_Mexico.utf8  LC_CTYPE=Spanish_Mexico.utf8
[3] LC_MONETARY=Spanish_Mexico.utf8 LC_NUMERIC=C
[5] LC_TIME=Spanish_Mexico.utf8

time zone: Europe/Berlin
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices datasets  utils     methods   base
```

```
other attached packages:
 [1] report_0.5.8        flextable_0.9.5     gt_0.10.1
 [4] boot_1.3-30         mgcv_1.9-1          nlme_3.1-164
 [7] WeightIt_1.0.0      CBPS_0.23           glmnet_4.1-8
[10] Matrix_1.6-5        numDeriv_2016.8-1.1 nnet_7.3-19
[13] MatchIt_4.5.5       MASS_7.3-60.0.1     table1_1.4.3
[16] lubridate_1.9.3     forcats_1.0.0       stringr_1.5.1
[19] dplyr_1.1.4         purrr_1.0.2         readr_2.1.5
[22] tidyr_1.3.1         tibble_3.2.1        ggplot2_3.5.0
[25] tidyverse_2.0.0     pacman_0.5.1
```

Fractional regression model

Convert SpO2 to fractional values between 0 and 1 to model.

```
data_original <- data_original %>% mutate(spo2_fraction = spo2_VPO/100)
```

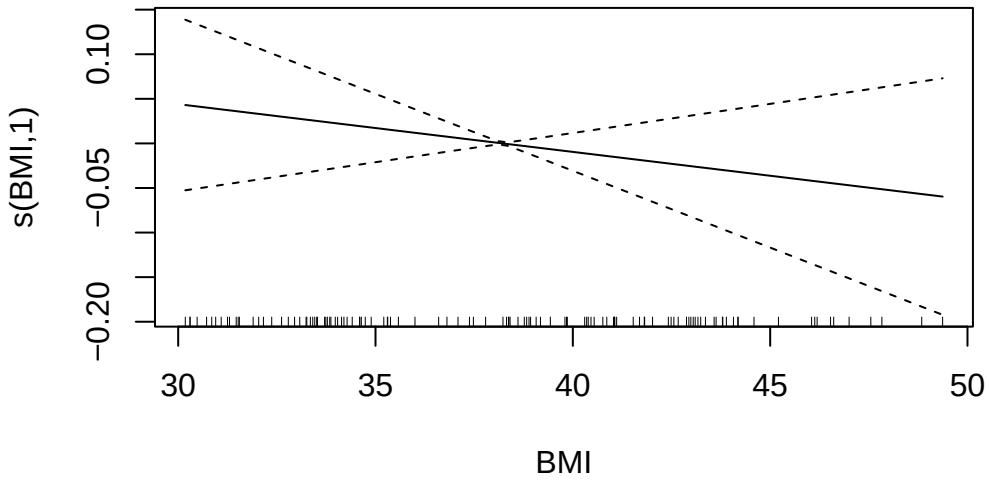
I will model separately by splitting the dataset into participants with SpO2 lower than or equal to 95 vs those with SpO2 higher than 95, according to what was shown in **Part 6**.

I will first reload processed data with original calculated weights and excluded outliers as used in **Part 6**. For the final model estimates, new weights were obtained for a selection of participants with an SpO2 lower than or equal to 95, which will be explained in the corresponding section of this document.

SpO2 high

BMI model

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.54443    0.02986   118.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(BMI) 1.001  1.001 0.809    0.37

R-sq.(adj) =  -0.00175   Deviance explained = 0.647%
GCV = 0.0031306  Scale est. = 0.0029188  n = 120
```



Atelectasis percent

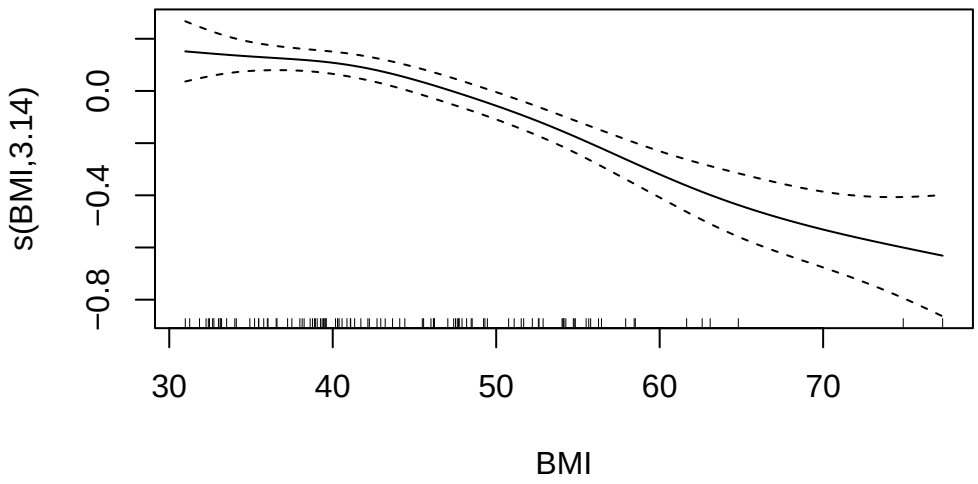
Atelectasis percent	n
0%	120

All patients with SpO2 higher than 95% have 0%. This shows that atelectasis percent and BMI are not relevant variables for SpO2 values above 95%.

SpO2 low

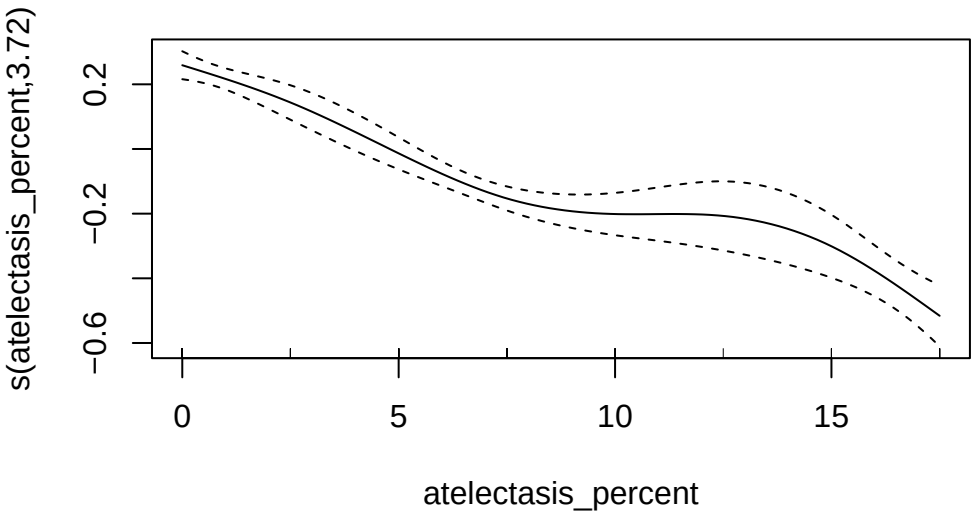
BMI model unadjusted

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.56131    0.02028   126.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(BMI) 3.138  3.854 21.95  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.452   Deviance explained = 44.4%
GCV = 0.0029953  Scale est. = 0.0029245  n = 108
```



Atelectasis percent model unadjusted

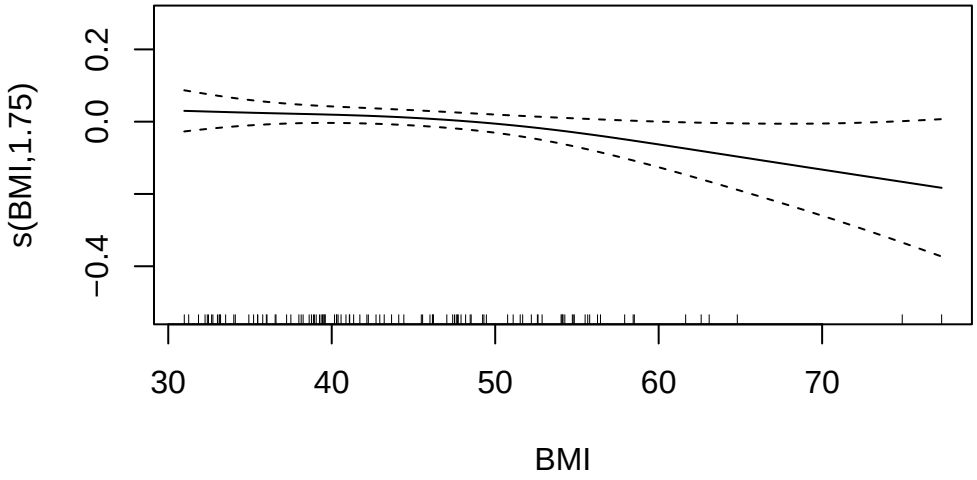
```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(atelectasis_percent, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.57120    0.01457   176.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df     F p-value
s(atelectasis_percent) 3.725  3.954 67.06  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.718   Deviance explained = 72.5%
GCV = 0.0014967  Scale est. = 0.0014811  n = 108
```

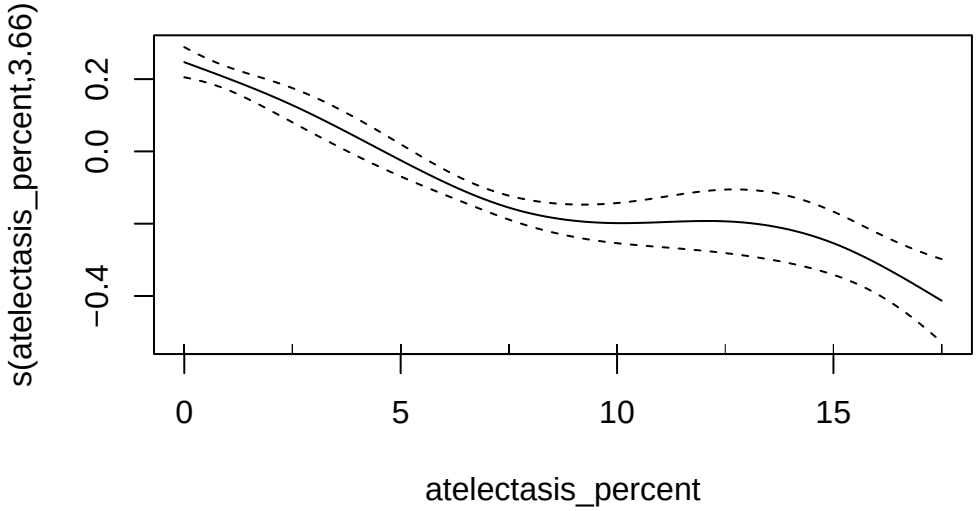


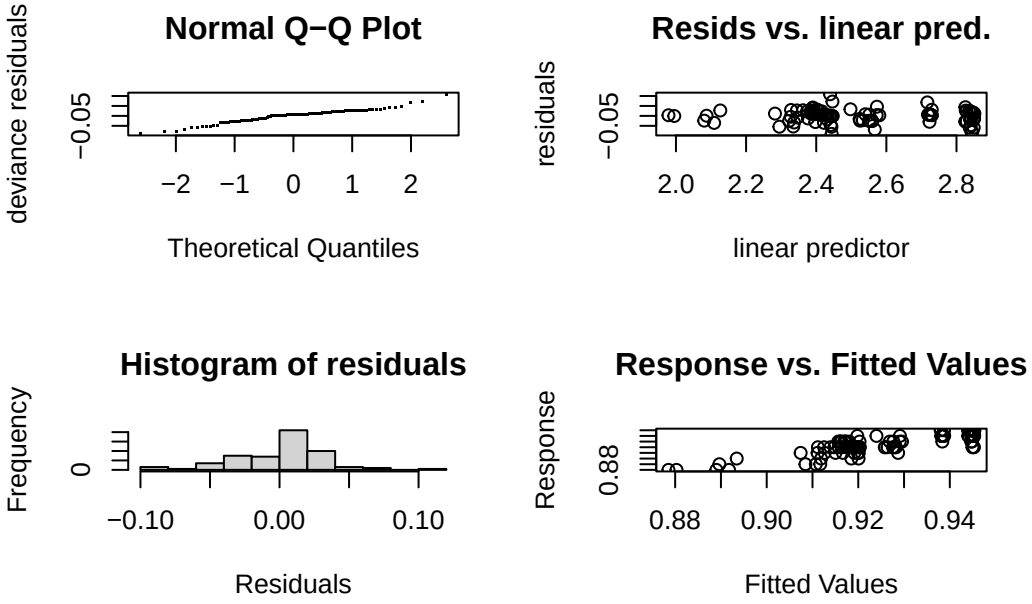
IPW model

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, k = 8) + s(atelectasis_percent, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.57462    0.01342   191.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df      F p-value
s(BMI)                 1.748  2.214  1.999   0.132
s(atelectasis_percent) 3.657  3.926 47.739  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.741   Deviance explained =   75%
```

```
GCV = 0.0011247  Scale est. = 0.0010753  n = 108
```





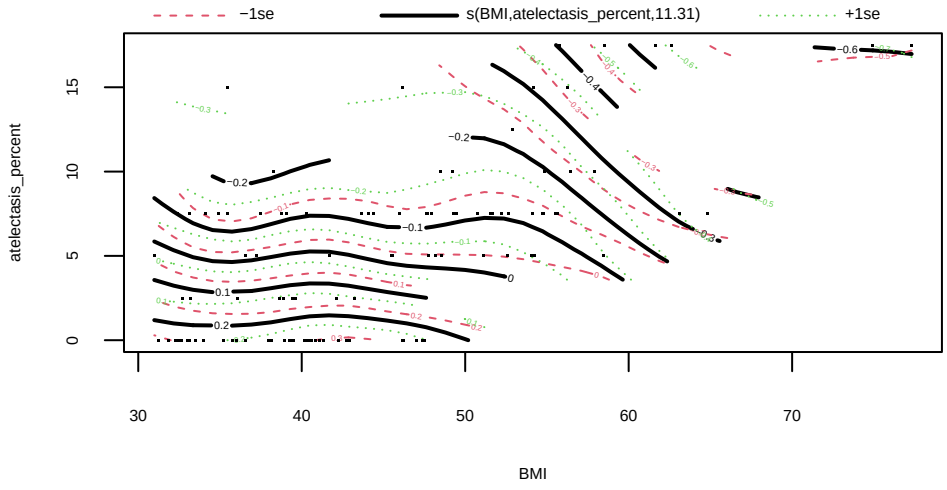



```
Method: GCV   Optimizer: outer newton
full convergence after 4 iterations.
Gradient range [4.947534e-11,4.571352e-10]
(score 0.00112472 & scale 0.001075342).
Hessian positive definite, eigenvalue range [4.183644e-06,5.475084e-06].
Model rank =  12 / 12

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                          k'  edf k-index p-value
s(BMI)                 7.00 1.75    0.94    0.23
s(atelectasis_percent) 4.00 3.66    1.17    0.96
```

Test for interaction

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ s(BMI, atelectasis_percent)
```

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.57405    0.01413   182.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                              edf Ref.df     F p-value
s(BMI,atelectasis_percent) 11.31  15.18 20.79  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.755   Deviance explained = 77.3%
GCV = 0.0011512  Scale est. = 0.001035  n = 108
```



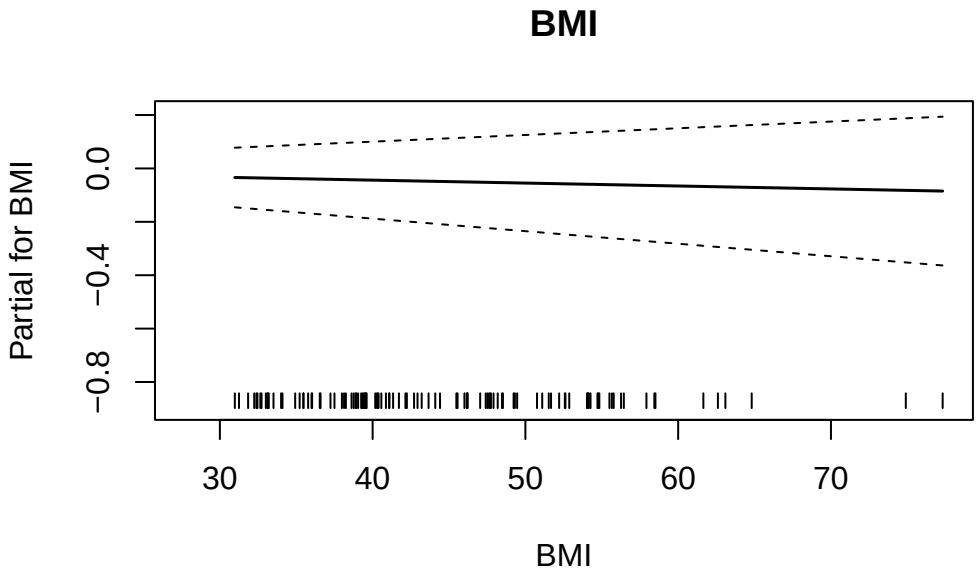
Effect mediation analysis assuming linearity

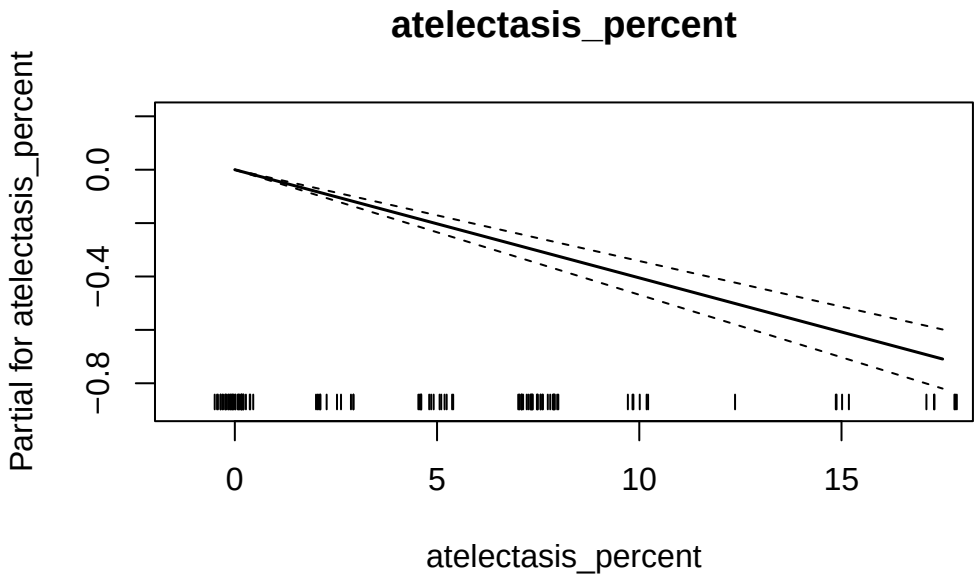
If we would assume a linear relationship, this would allow to calculate the proportion mediated. Since the relationships in prior models did not deviate seriously from linear, I will model with linear terms and check distribution of residuals. If this suggests that assuming linearity results in good enough models in this subset of participants with SpO2 lower than or equal to 95%,

I will calculate the proportion mediated to have an idea of how much of the effect of BMI on SpO2 is mediated by atelectasis.

Direct and indirect effects

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ BMI + atelectasis_percent

Parametric coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.837504   0.072771  38.992   <2e-16 ***
BMI                 -0.001097   0.001802  -0.609    0.544
atelectasis_percent -0.040524   0.003162 -12.816   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


R-sq.(adj) =  0.679   Deviance explained = 68.8%
GCV = 0.0013148  Scale est. = 0.0012932  n = 108
```





Total effect

```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ BMI

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.359829   0.105595  31.818  < 2e-16 ***
BMI         -0.017853   0.002252  -7.928  2.4e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


R-sq.(adj) =  0.386   Deviance explained = 37.1%
GCV = 0.0031584  Scale est. = 0.0031604  n = 108
```

Proportion mediated

```
  BMI
93.86
```

This model, however, could be biased due to selection (filtering has been done by conditioning on SpO2, which is a descendant and a collider). Therefore, reweighting after selection could provide a better estimate. Thus, I will obtain new weights for the pseudopopulation of participants with SpO2 lower than 95%. Since selection on a descendant (SpO2) likely introduced novel backdoor pathways, I will include all the ancestor variables of interest in the propensity score models, contrary to what I had done before. Despite this, it should be noted that additional novel backdoor pathways with other (un)measured confounders could still be latent, reason why the proportion mediated estimate shown here should be taken with some level of skepticism and interpreted as an approximate number of the proportion of the effect of BMI mediated by atelectasis percent, which could be biased.

Propensity scores

Weights for exposure (BMI):

Weights for mediator (atelectasis percent):

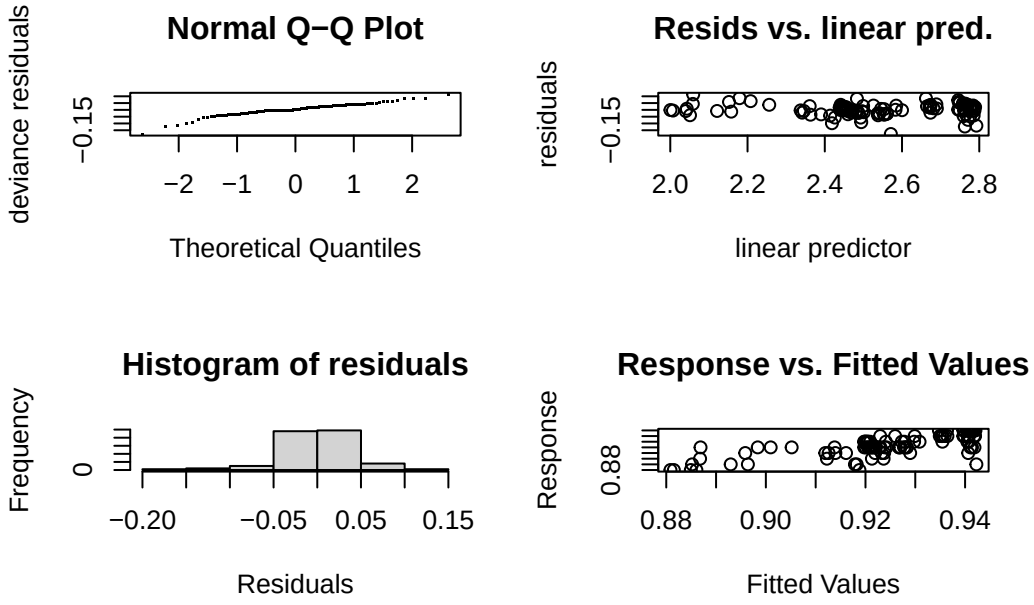
Overall weight:

IPW linear model

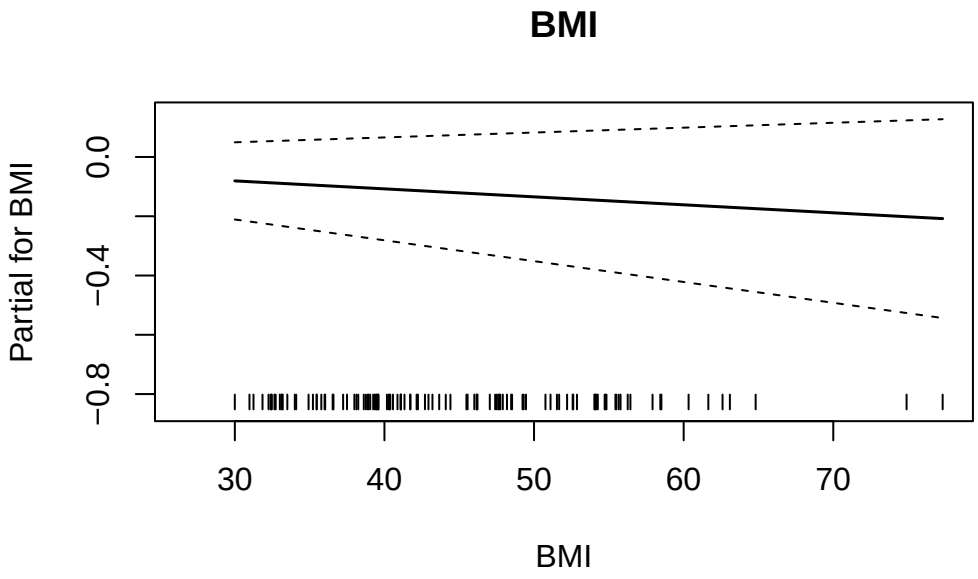
Direct and indirect effects

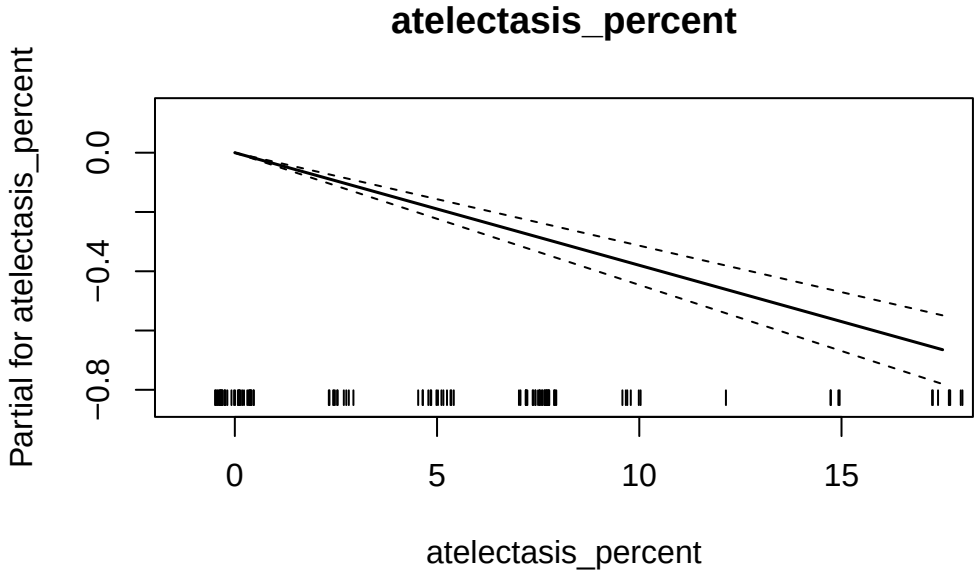
```
Family: quasibinomial
Link function: logit

Formula:
spo2_fraction ~ BMI + atelectasis_percent

Parametric coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.873277   0.089707  32.029   <2e-16 ***
BMI                 -0.002691   0.002169  -1.241    0.217
atelectasis_percent -0.037986   0.003311 -11.473   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


R-sq.(adj) =  0.661   Deviance explained = 66.5%
```

```
GCV = 0.0019088  Scale est. = 0.0019185  n = 114
```




```
Method: GCV   Optimizer: outer newton
Model required no smoothing parameter selectionModel rank =  3 / 3
```

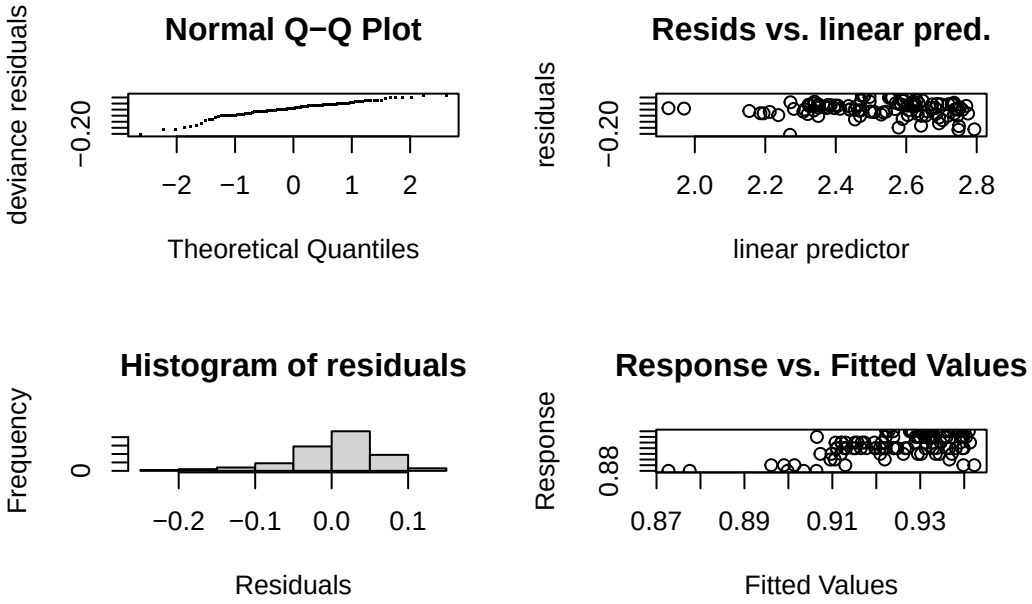




Total effect

```
Family: quasibinomial
```

```
Link function: logit

Formula:
spo2_fraction ~ BMI

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.343978   0.112361  29.761  < 2e-16 ***
BMI         -0.018351   0.002397  -7.657 7.29e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


R-sq.(adj) =  0.367   Deviance explained = 34.7%
GCV = 0.0037505  Scale est. = 0.003848  n = 114
```




```
Method: GCV   Optimizer: outer newton
Model required no smoothing parameter selectionModel rank =  2 / 2
```

Proportion mediated

```
  BMI
85.34
```

Outliers

```
data_spo2_low %>%
  mutate(
  cooksd = cooks.distance(model_linear_BMI_atelectasis),
  outlier = ifelse(cooksd < 4/nrow(data_spo2_low), "keep","delete")
) %>%
  filter(outlier=="delete") %>%
  dplyr::select(ID,BMI,spo2_VPO,cooksd,outlier) %>%
  arrange(desc(cooksd)) %>%
  gt()
```

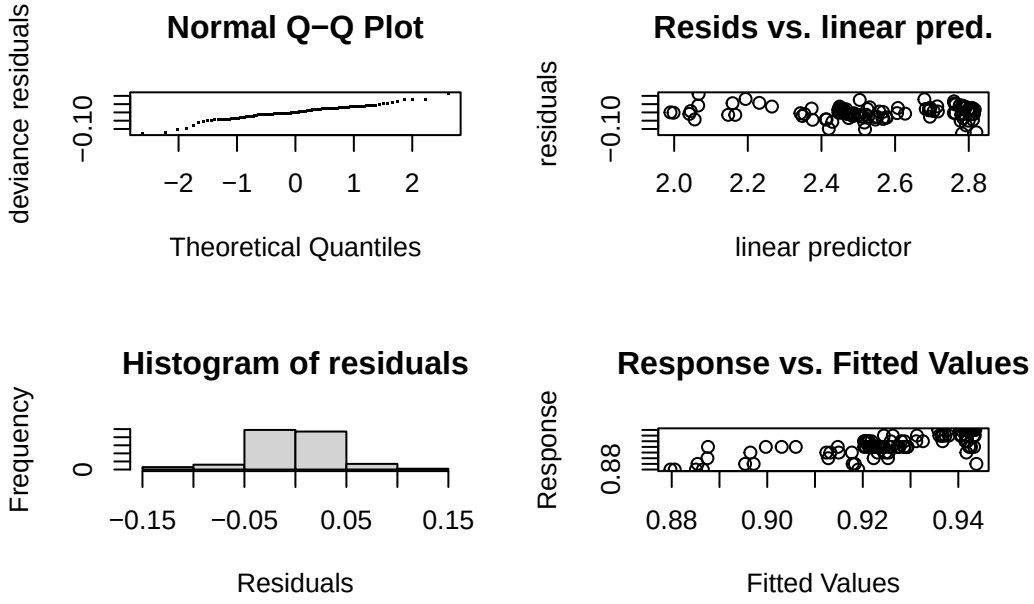
ID	BMI	spo2_VPO	cooksd	outlier
163	41.72	91	0.40316741	delete
165	34.93	91	0.19160504	delete
205	55.44	92	0.14837208	delete
228	58.51	88	0.14068422	delete
117	63.09	89	0.12581363	delete
170	46.16	92	0.08682621	delete
208	35.47	92	0.08170736	delete
70	47.40	95	0.06084543	delete
114	41.10	91	0.05127373	delete
209	56.26	91	0.03803523	delete

I will remove this very influential outlier (ID = 163)

Direct and indirect effects

```
Call:
glm(formula = spo2_fraction ~ BMI + atelectasis_percent, family = quasibinomial(link = logit)
    data = data_spo2_low_linear, weights = weight)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.925568   0.083903  34.868   <2e-16 ***
BMI                 -0.003495   0.002018  -1.732    0.086 .
atelectasis_percent -0.038021   0.003071 -12.381   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.00162844)

    Null deviance: 0.59268  on 112  degrees of freedom
Residual deviance: 0.17251  on 110  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

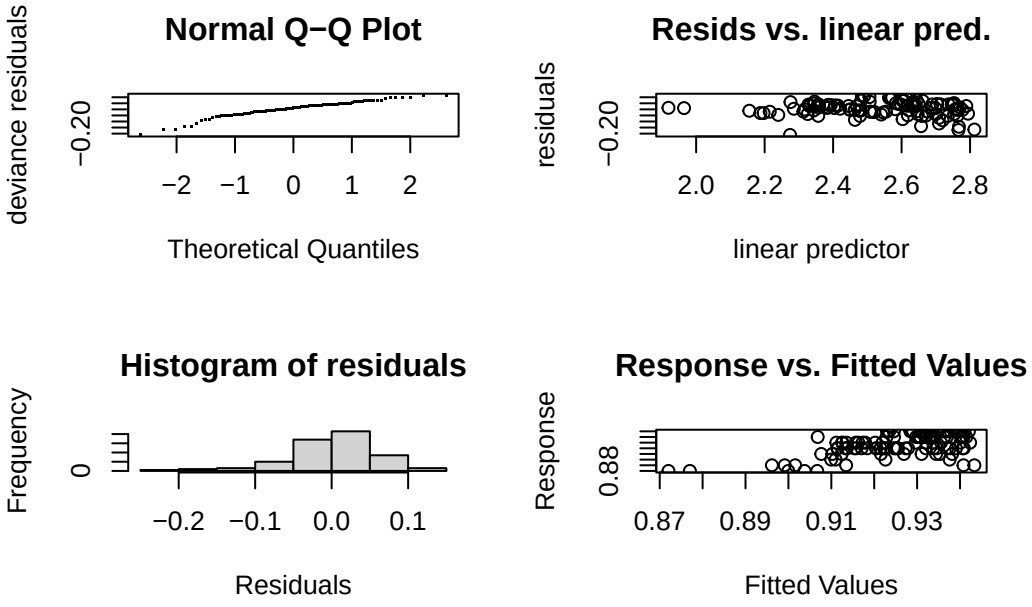



```
'gamm' based fit - care required with interpretation.
Checks based on working residuals may be misleading.
```

Total effect

```
Call:
glm(formula = spo2_fraction ~ BMI, family = quasibinomial(link = logit),
    data = data_spo2_low_linear, weights = weight1)

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.37916    0.11066  30.537  < 2e-16 ***
BMI         -0.01888    0.00235  -8.038 1.08e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.003646863)

    Null deviance: 0.61869  on 112  degrees of freedom
Residual deviance: 0.38967  on 111  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```




```
'gamm' based fit - care required with interpretation.
Checks based on working residuals may be misleading.
```

Proportion mediated

```
  BMI
81.49
```

As it can be seen from the models, the proportion mediated estimate is quite sensible to decisions in analysis. (i.e., removing outliers, obtaining new weights, etc). Nonetheless, the overall message remains the same: the proportion mediated is rather high, in the magnitude of 80 to 92%.

I will obtain the confidence intervals for the proportion mediated of this last estimate of 81.49%:

The confidence intervals for the proportion mediated werre calculated with the sourced script ***confidence_intervals_proportion_mediated.R***.

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_results, type = "all")

Intervals :
Level      Normal              Basic
95%   ( 0.5520,  1.0776 )   ( 0.5333,  1.0746 )

Level     Percentile            BCa
95%   ( 0.5552,  1.0965 )   ( 0.5596,  1.1100 )
Calculations and Intervals on Original Scale
```

Note that the proportion mediated should not include values higher than 1. Therefore, I will truncate the upper boundary value of the confidence interval for the reporting. Fortunately, this confidence interval somehow reflects how decisions in analysis can lead to such different estimates, which are contained in this confidence interval.

Confidence intervals for the coefficients.

I will calculate confidence intervals with bootstrapping since confidence intervals from the weighted model would be incorrectly narrow due to weights.

Confidence intervals and OR calculated with the accompanying sourced script ***confidence_intervals_mediation.R***.

Supplementary Table

Characteristic	OR	95%CI
Total effect of BMI		
BMI	0.98	0.97—0.99
Direct and indirect effects of BMI		
BMI	1	0.99—1
Atelectasis percent	0.96	0.96—0.97

Package References

- Angelo Canty, B. D. Ripley (2024). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-30. A. C. Davison, D. V. Hinkley (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge. ISBN 0-521-57391-2, doi:10.1017/CBO9780511802843.
- Bates D, Maechler M, Jagan M (2024). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.6-5, https://CRAN.R-project.org/package=Matrix.
- Fong C, Ratkovic M, Imai K (2022). *CBPS: Covariate Balancing Propensity Score*. R package version 0.23, https://CRAN.R-project.org/package=CBPS.
- Friedman J, Tibshirani R, Hastie T (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, *33*(1), 1-22. doi:10.18637/jss.v033.i01 https://doi.org/10.18637/jss.v033.i01. Simon N, Friedman J, Tibshirani R, Hastie T (2011). "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software*, *39*(5), 1-13. doi:10.18637/jss.v039.i05 https://doi.org/10.18637/jss.v039.i05. Tay JK, Narasimhan B, Hastie T (2023). "Elastic Net Regularization Paths for All Generalized Linear Models." *Journal of Statistical Software*, *106*(1), 1-31. doi:10.18637/jss.v106.i01 https://doi.org/10.18637/jss.v106.i01.
- Gilbert P, Varadhan R (2019). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.1, https://CRAN.R-project.org/package=numDeriv.
- Gohel D, Skintzos P (2024). *flextable: Functions for Tabular Reporting*. R package version 0.9.5, https://CRAN.R-project.org/package=flextable.
- Greifer N (2024). *WeightIt: Weighting for Covariate Balance in Observational Studies*. R package version 1.0.0, https://CRAN.R-project.org/package=WeightIt.
- Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate." *Journal of Statistical Software*, *40*(3), 1-25. https://www.jstatsoft.org/v40/i03/.
- Ho D, Imai K, King G, Stuart E (2011). "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software*, *42*(8), 1-28. doi:10.18637/jss.v042.i08 https://doi.org/10.18637/jss.v042.i08.
- Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J (2024). *gt: Easily Create Presentation-Ready Display Tables*. R package version 0.10.1, https://CRAN.R-project.org/package=gt.
- Makowski D, Lüdecke D, Patil I, Thériault R, Ben-Shachar M, Wiernik B (2023). "Automated Results Reporting as a Practical Tool to Improve Reproducibility and Methodological Best Practices Adoption." *CRAN*. https://easystats.github.io/report/.
- Müller K, Wickham H (2023). *tibble: Simple Data Frames*. R package version 3.2.1, https://CRAN.R-project.org/package=tibble.
- Pinheiro J, Bates D, R Core Team (2023). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-164, https://CRAN.R-project.org/package=nlme. Pinheiro JC, Bates DM (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York. doi:10.1007/b98882 https://doi.org/10.1007/b98882.

- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
- Rich B (2023). *table1: Tables of Descriptive Statistics in HTML*. R package version 1.4.3, https://CRAN.R-project.org/package=table1.
- Rinker TW, Kurkiewicz D (2018). *pacman: Package Management for R*. version 0.5.0, http://github.com/trinker/pacman.
- Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.
- Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
- Wickham H (2023). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 1.0.0, https://CRAN.R-project.org/package=forcats.
- Wickham H (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.1, https://CRAN.R-project.org/package=stringr.
- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.
- Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4, https://CRAN.R-project.org/package=dplyr.
- Wickham H, Henry L (2023). *purrr: Functional Programming Tools*. R package version 1.0.2, https://CRAN.R-project.org/package=purrr.
- Wickham H, Hester J, Bryan J (2024). *readr: Read Rectangular Text Data*. R package version 2.1.5, https://CRAN.R-project.org/package=readr.
- Wickham H, Vaughan D, Girlich M (2024). *tidyr: Tidy Messy Data*. R package version 1.3.1, https://CRAN.R-project.org/package=tidyr.
- Wood SN (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." *Journal of the Royal Statistical Society (B)*, *73*(1), 3-36. Wood S, N., Pya, S"afken B (2016). "Smoothing parameter and model selection for general smooth models (with discussion)." *Journal of the American Statistical Association*, *111*, 1548-1575. Wood SN (2004). "Stable and efficient multiple smoothing parameter estimation for generalized additive models." *Journal of the American Statistical Association*, *99*(467), 673-686. Wood S (2017). *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC. Wood SN (2003). "Thin-plate regression splines." *Journal of the Royal Statistical Society (B)*, *65*(1), 95-114.