Probability of Causation

Lung cancer due to asbestos exposure

Jenny Deng, Javier Mancilla Galindo, Lutzen Portengen

2024-11-26

Table of contents

Session and package dependencies

```
R version 4.4.0 (2024-04-24 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 22631)

Matrix products: default


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

time zone: Europe/Amsterdam
tzcode source: internal

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

other attached packages:
 [1] report_0.5.9    gt_0.11.0       docstring_1.0.0 gridExtra_2.3
 [5] mixmeta_1.2.0   sjPlot_2.8.16   lme4_1.1-35.3   Matrix_1.7-0
 [9] flextable_0.9.7 table1_1.4.3    lubridate_1.9.3 forcats_1.0.0
[13] stringr_1.5.1   dplyr_1.1.4     purrr_1.0.2     readr_2.1.5
[17] tidyr_1.3.1     tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
[21] pacman_0.5.1
```

Descriptive characteristics

	Total (N=37866)	Controls (N=20965)	Lung cancer cases (N=16901)
Sex			
Female	7810 (20.6%)	4514 (21.5%)	3296 (19.5%)
Male	30056 (79.4%)	16451 (78.5%)	13605 (80.5%)
Age			
Mean (SD)	61.7 (9.63)	61.5 (9.92)	62.0 (9.23)
Asbestos (ff/ml-years)			
Mean (SD)	2.42 (3.01)	2.22 (3.01)	2.60 (3.00)
Median (Q1, Q3)	1.33 (0.550, 3.22)	1.17 (0.476, 2.86)	1.54 (0.622, 3.53)
Min, Max	0.00214, 64.6	0.00229, 64.6	0.00214, 35.4
Never exposed	23114 (61.0%)	13653 (65.1%)	9461 (56.0%)
Exposure duration (years)			
Mean (SD)	18.4 (13.9)	17.9 (13.8)	18.9 (14.0)
Median (Q1, Q3)	15.0 (6.00, 30.0)	14.0 (5.00, 30.0)	15.5 (6.00, 31.0)
Min, Max	1.00, 63.0	1.00, 63.0	1.00, 62.0
Never exposed	23114 (61.0%)	13653 (65.1%)	9461 (56.0%)
Smoking			
Never smoker	8522 (22.5%)	7153 (34.1%)	1369 (8.1%)
Former smoker	13652 (36.1%)	8220 (39.2%)	5432 (32.1%)
Current smoker	15692 (41.4%)	5592 (26.7%)	10100 (59.8%)
Pack-years			
Median (Q1, Q3)	23.3 (1.60, 41.5)	9.75 (0, 29.1)	35.8 (21.0, 51.0)
Time since quitting smoking			
0-7 years	3448 (9.1%)	1422 (6.8%)	2026 (12.0%)
8-15 years	3429 (9.1%)	1898 (9.1%)	1531 (9.1%)
16-25 years	3517 (9.3%)	2346 (11.2%)	1171 (6.9%)
>25 years	3258 (8.6%)	2554 (12.2%)	704 (4.2%)

Exposure to asbestos (binary)

Contingency Table

- Exposure: Asbestos (`ever_asbestos0`). A total of 14752 were ever exposed.
- Outcome: Lung cancer (`status`) occurred in 16901 cases, out of which 13605 (80.5%) were male and 3296 (19.5%) female.

	Lung Cancer	
Exposed Asbestos	1	0
1	7440	7312
0	9461	13653

Odds Ratio (OR) and Attributable Fraction (AF).

```
or <- (7440/7312)/(9461/13653)
```

The crude OR is **1.47**.

```
af <- (or-1)/or
```

The AF is **0.32**.

Here, the attributable fraction refers specifically to an approximation of the *excess fraction*, interpreted as the excess caseload due to exposure.[1]

Armstrong, et al.[2] applied the prior formula to determine the probability of causation (PoC) of bladder cancer at different values of cumulative exposure to aluminium in the workplace, with data from a case-control study. The OR is used here instead of RR in the original formula due to the case-control nature of the underlying data.

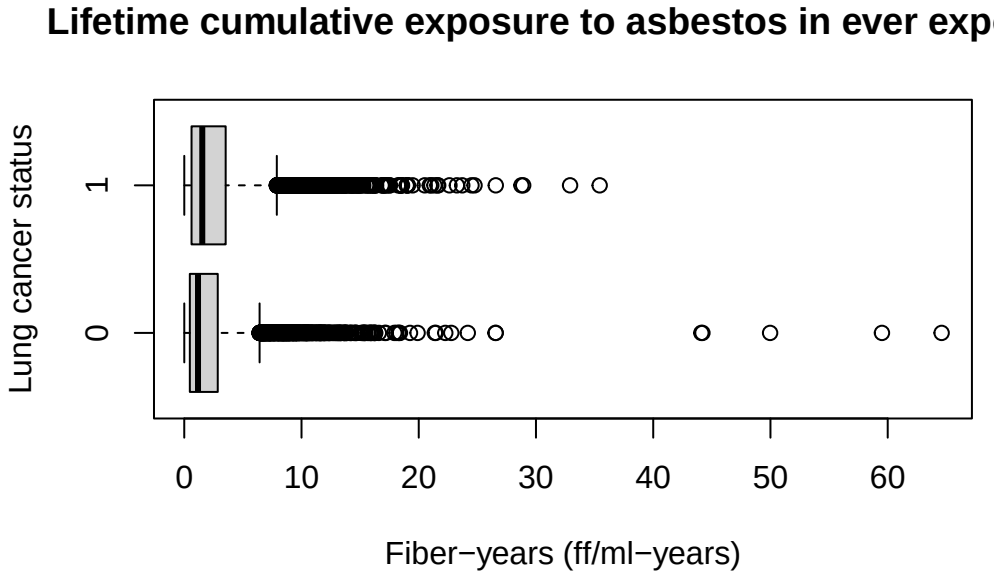
```
PoCfun <- function(logor) {
  OR <- exp(logor)
  pmax((OR-1)/OR,0)
}
```

Exposure to asbestos as continuous variable

Two exposure measures were used for modelling the continuous exposure to asbestos.

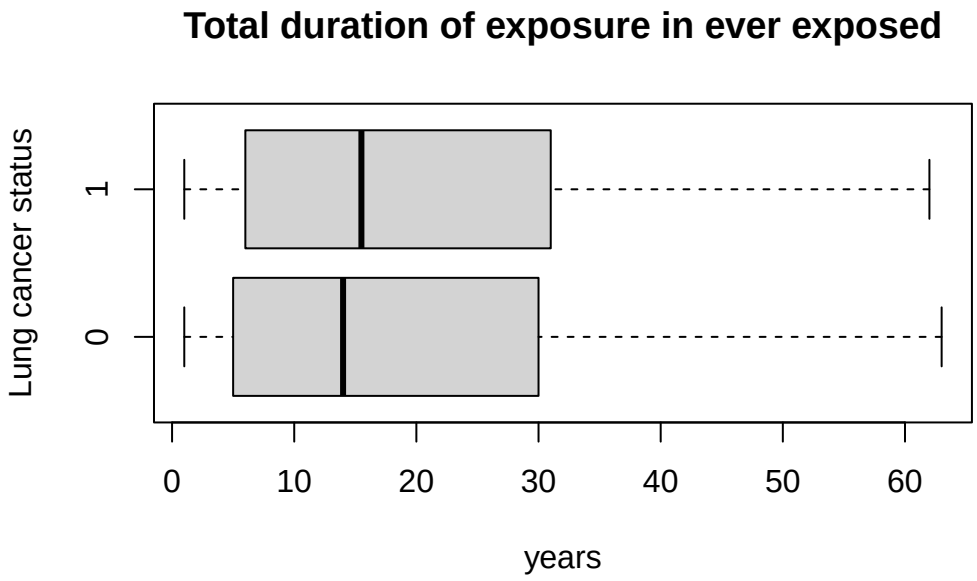
Lifetime cumulative exposure to asbestos (`asbestos_cum0`)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.0000  0.0000  0.0000  0.9409  0.8267 64.5996 
```



Total duration of exposure in years (`asbestos_dur0`)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.000   0.000   7.161   9.000  63.000 
```



Probability of Causation (PoC)

Calculations of PoC with different statistical models and incorporating error terms for misclassification are presented in the following subsections.

Spline Model

This model incorporates the shape of the exposure-response relationship from nonlinear meta-regression models as determined by van der Bij, et al.[3] An additional 5 studies have been incorporated into the meta-analysis to update the exposure-response relationship. The underlying data and results of the meta-analysis are contained within the object `MOD`. Tailored-made functions are sourced into this analysis, namely:

- `predict.nsplin`: Generates predictions using natural spline transformations based on the provided model object.
- `getPoC`: Calculate probability of causation (PoC) and risk ratio (RR) based on a specified (linear/spline) model that derives the exposure-response relationship.

```
MOD <- readRDS(file.path(inputfolder,"MOD.rds"))
source("scripts/predict.nsplin.R")
source("scripts/getPoC.R")
```

There are 4 different models possible:

1. linear model, assumes difference in background rate of outcome
2. linear model, assumes no difference in background rate of outcome
3. spline model, assumes difference in background rate of outcome
4. spline model, assumes no difference in background rate of outcome

Shape of the exposure-response curve under a wide range of values:

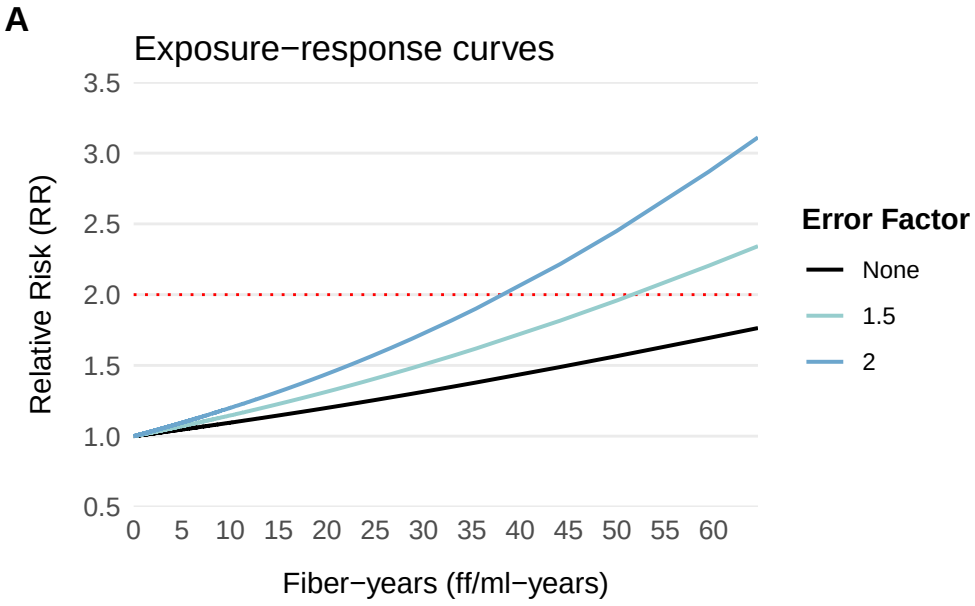
This figure shows that a doubling in the risk of lung cancer (RR = 2) is not expected to be seen with the values of exposure observed in the SYNERGY study (shaded area).

Using modelling strategy number 4, and the actual exposure data in SYNERGY, the PoC and prediction intervals as a measure of uncertainty from study heterogeneity can be obtained following Higgins, et al. method.[4]

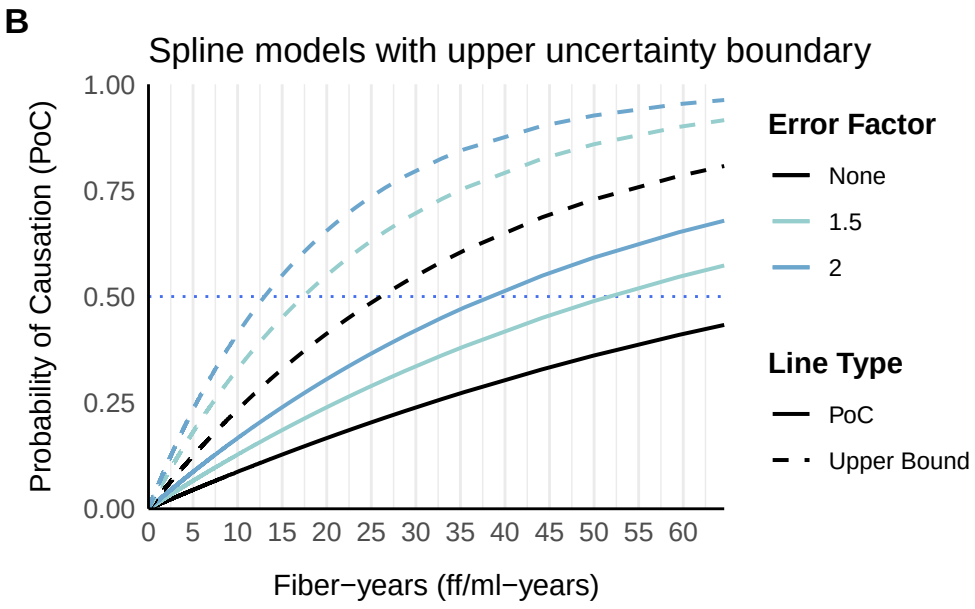
Participants with outcome and a PoC greater than or equal to 50%:

Bound	Cases	Cases per 10,000	Min cum exp
No Error Factor			
PoC	0	0	Inf
Upper Bound	5	3	26.57110
Error Factor: 1.5			
PoC	0	0	Inf
Upper Bound	30	18	17.48189
Error Factor: 2			
PoC	0	0	Inf
Upper Bound	94	56	13.05413

This is the plot of lung cancer risk ratio with the observed exposure values in this study:



And the probability of causation, with the upper prediction interval:



Logistic model - cum0

A multiple logistic regression model is used to estimate the PoC, with `asbestos_cum0` as the main explanatory variable, and adjusted for:

- The study source of participants (`study_name`)
- Age category (`agegroup`). The age groups are: <45, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, and > 74 years.
- Sex (`sex`), female and male.
- Smoking (`packyrs`), cigarette pack-years.
- Time since smoking cessation (`time_quit`). The categories are: current smokers; stopping smoking 2–7, 8–15, 16–25, and >= 26 years before interview/diagnosis; and never-smokers.

```
logistic_model <- glm(
  status ~ asbestos_cum0 +
    study_name + agegroup + sex + packyrs + time_quit,
  data = df,
  family=binomial(link="logit")
  )
```

The following values will be extracted from the model and/or assigned in order to be used for the probability of causation calculations:

```
## Coefficient for exposure
b <- coef(summary(logistic_model))["asbestos_cum0", "Estimate"]

## Upper bound of 95% confidence interval
confint <- confint(logistic_model)["asbestos_cum0", ]
UB <- confint[2]

## Underestimation of the exposure-response relationship due to
## misclassification error, with a factor of 1.5 (A) and 2 (B):
A <- 1.5
B <- 2
```

The coefficient for exposure (b) is 0.0371 and the upper bound of the 95% CI (UB) is 0.0485, which correspond to an increase in lung cancer risk of:

- b: There is an increase of 3.78% in the risk of lung cancer per every additional fibre-years.
- UB: There is an increase of 4.97% in the risk of lung cancer per every additional fibre-years.

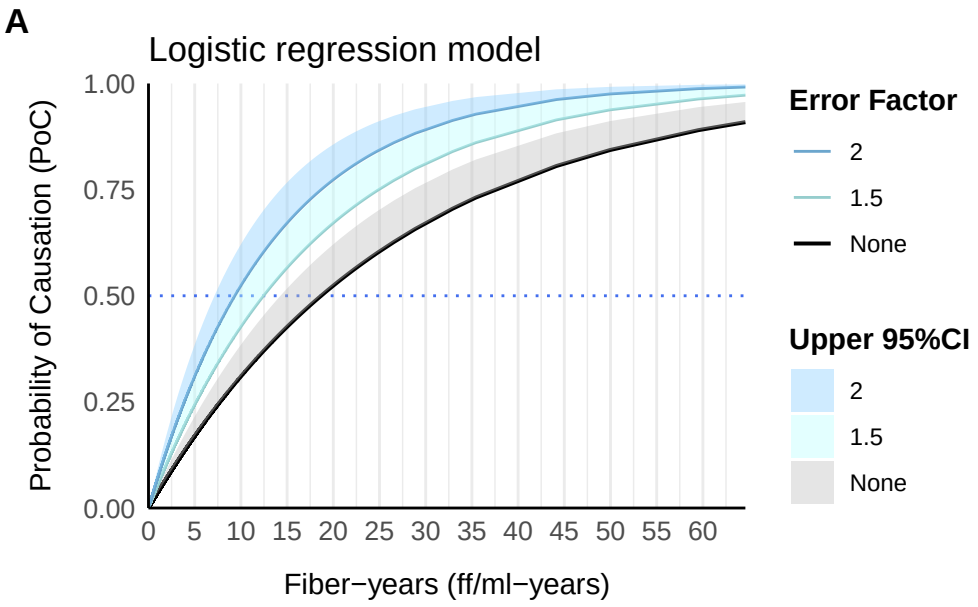
Calculation of PoC:

```
## Select subset of variables of interest and apply PoC function:
df.cum0 <- df %>%
  select(subjctid, status, sex, agegroup, study_name, packyrs, time_quit,
         list_a, ever_asbestos0, asbestos_cum0, asbestos_dur0) %>%
  mutate(PoC = PoCfun(b * asbestos_cum0),
         PoC_UB = PoCfun(UB * asbestos_cum0),
         PoC_scenario_b1.5 = PoCfun(b * A * asbestos_cum0),
         PoC_scenario_b2 = PoCfun(b * B * asbestos_cum0),
         PoC_scenario_UB1.5 = PoCfun(UB * A * asbestos_cum0),
         PoC_scenario_UB2 = PoCfun(UB * B * asbestos_cum0)
         )
```

Out of those ever exposed to asbestos, the summary of PoC with this approach is as follows:

Scenario	N	mean	sd	min	q25	median	q75	max
PoC	14752	0.08	0.09	0	0.02	0.05	0.11	0.91
PoC_UB	14752	0.10	0.11	0	0.03	0.06	0.14	0.96
PoC_scenario_b1.5	14752	0.12	0.12	0	0.03	0.07	0.16	0.97
PoC_scenario_b2	14752	0.15	0.15	0	0.04	0.09	0.21	0.99
PoC_scenario_UB1.5	14752	0.14	0.14	0	0.04	0.09	0.21	0.99
PoC_scenario_UB2	14752	0.18	0.17	0	0.05	0.12	0.27	1.00

Plot of PoC values with the upper bound of the 95% CI under the different scenarios of misclassification error, according to cumulative asbestos exposure values.



Participants with outcome and a PoC greater than or equal to 50% under different scenarios:

Overall

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	21	12	18.97
PoC_UB	66	39	14.29
PoC_scenario_b1.5	105	62	12.48
PoC_scenario_b2	275	163	9.35
PoC_scenario_UB1.5	266	157	9.54
PoC_scenario_UB2	586	347	7.14

Males

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	21	15	18.97
PoC_UB	62	46	14.29
PoC_scenario_b1.5	101	74	12.48
PoC_scenario_b2	265	195	9.35
PoC_scenario_UB1.5	256	188	9.54
PoC_scenario_UB2	570	419	7.14

Female:

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	0	0	NA
PoC_UB	4	12	15.06
PoC_scenario_b1.5	4	12	15.06
PoC_scenario_b2	10	30	10.32
PoC_scenario_UB1.5	10	30	10.32
PoC_scenario_UB2	16	49	7.16

Mixed effects model - cum0

Adding random effects with a random intercept for each study source (`study_name`) and random slopes for the exposure (`asbestos_cum0`) within each study source.

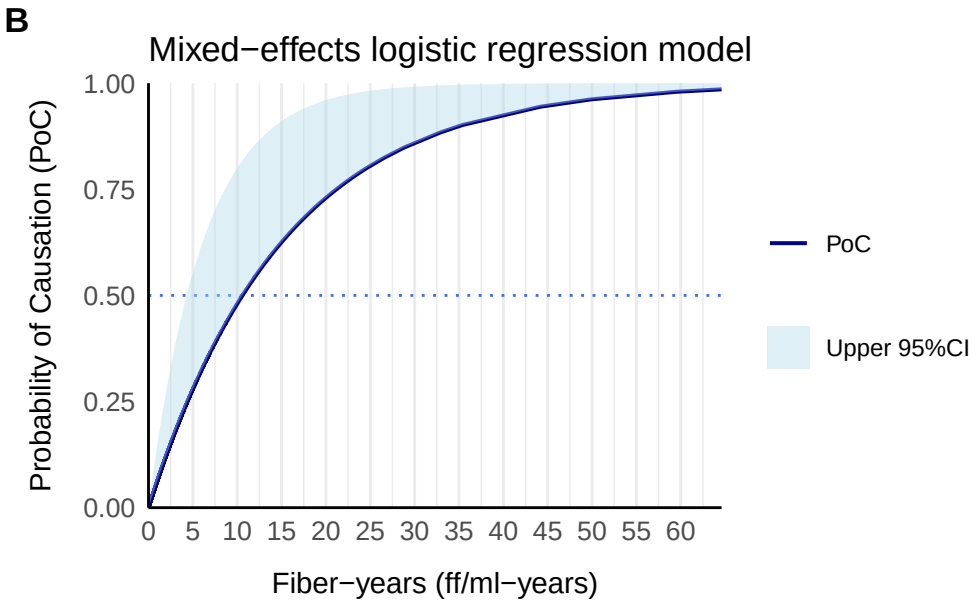
```
mixed_model_cum0 <- glmer(
  status ~ asbestos_cum0 + agegroup + sex + packyrs + time_quit +
    (1 + asbestos_cum0 | study_name),
  data = df,
  family = binomial,
  control = glmerControl(optimizer = "bobyqa"),
  nAGQ = 0
  )
```

The coefficient for exposure (b) is 0.0654 which corresponds to an increase in lung cancer risk of 6.76% per every additional fibre-years.

Out of those ever exposed to asbestos, the summary of PoC with this approach is as follows:

Scenario	N	mean	sd	min	q25	median	q75	max
PoC	14752	0.13	0.13	0	0.04	0.08	0.19	0.99
PoC_LB	14752	0.00	0.00	0	0.00	0.00	0.00	0.00
PoC_UB	14752	0.27	0.23	0	0.08	0.19	0.40	1.00

Plot of PoC values with the upper bound of the 95% prediction interval, according to cumulative asbestos exposure values.



Participants with outcome and a PoC greater than or equal to 50% under different scenarios:

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	193	114	10.59
PoC_LB	0	0	NA
PoC_UB	1429	846	4.31

Males

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	184	135	10.59
PoC_LB	0	0	NA
PoC_UB	1392	1023	4.31

Female:

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	9	27	10.89
PoC_LB	0	0	NA
PoC_UB	37	112	4.31

Logistic model - dur0

```
logistic_model <- glm(
  status ~ asbestos_dur0 +
    study_name + agegroup + sex + packyrs + time_quit,
  data = df,
  family=binomial(link="logit")
  )
```

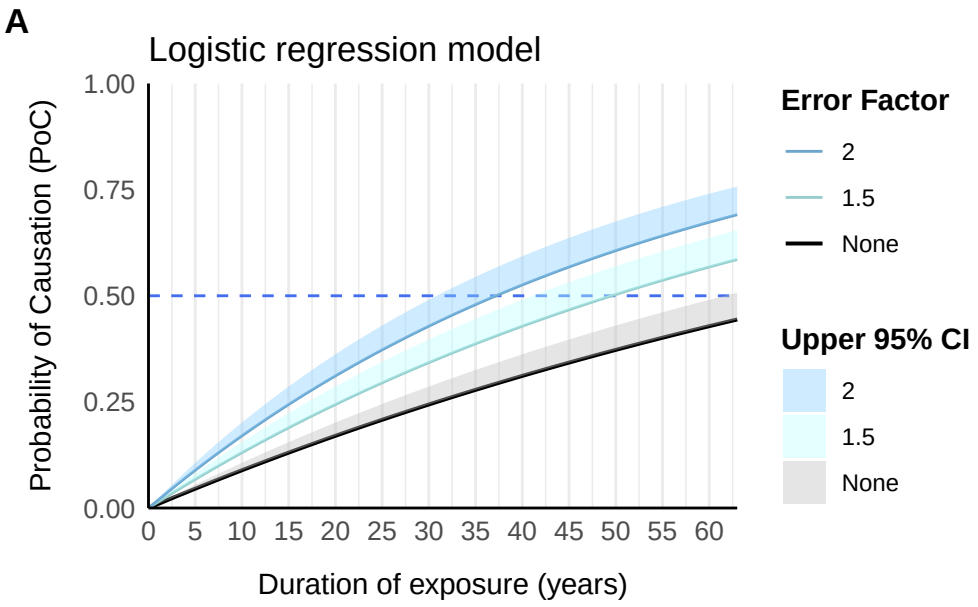
The coefficient for exposure (b) is 0.0093 and the upper bound of the 95% CI (UB) is 0.0112, which correspond to an increase in lung cancer risk of:

- b: There is an increase of 0.94% in the risk of lung cancer per additional year exposed.
- UB: There is an increase of 1.13% in the risk of lung cancer per additional year exposed.

Out of those ever exposed to asbestos, the summary of PoC with this approach is as follows:

Scenario	N	mean	sd	min	q25	median	q75	max
PoC	14752	0.15	0.11	0.01	0.05	0.13	0.24	0.44
PoC_UB	14752	0.18	0.12	0.01	0.07	0.16	0.29	0.51
PoC_scenario_b1.5	14752	0.21	0.14	0.01	0.08	0.19	0.34	0.59
PoC_scenario_b2	14752	0.27	0.18	0.02	0.11	0.24	0.43	0.69
PoC_scenario_UB1.5	14752	0.25	0.17	0.02	0.10	0.22	0.40	0.65
PoC_scenario_UB2	14752	0.31	0.20	0.02	0.13	0.29	0.49	0.76

Plot of PoC values with the upper bound of the 95% CI under the different scenarios of misclassification error, according to duration of asbestos exposure.



Participants with outcome and a PoC greater than or equal to 50% under different scenarios:

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	0	0	NA
PoC_UB	1	1	62
PoC_scenario_b1.5	58	34	50
PoC_scenario_b2	1013	599	38
PoC_scenario_UB1.5	517	306	42
PoC_scenario_UB2	1913	1132	31

Males

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	0	0	NA
PoC_UB	1	1	62
PoC_scenario_b1.5	57	42	50
PoC_scenario_b2	1005	739	38
PoC_scenario_UB1.5	511	376	42
PoC_scenario_UB2	1885	1386	31

Female:

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	0	0	NA
PoC_UB	0	0	NA
PoC_scenario_b1.5	1	3	52
PoC_scenario_b2	8	24	38
PoC_scenario_UB1.5	6	18	42
PoC_scenario_UB2	28	85	31

Mixed effects model - dur0

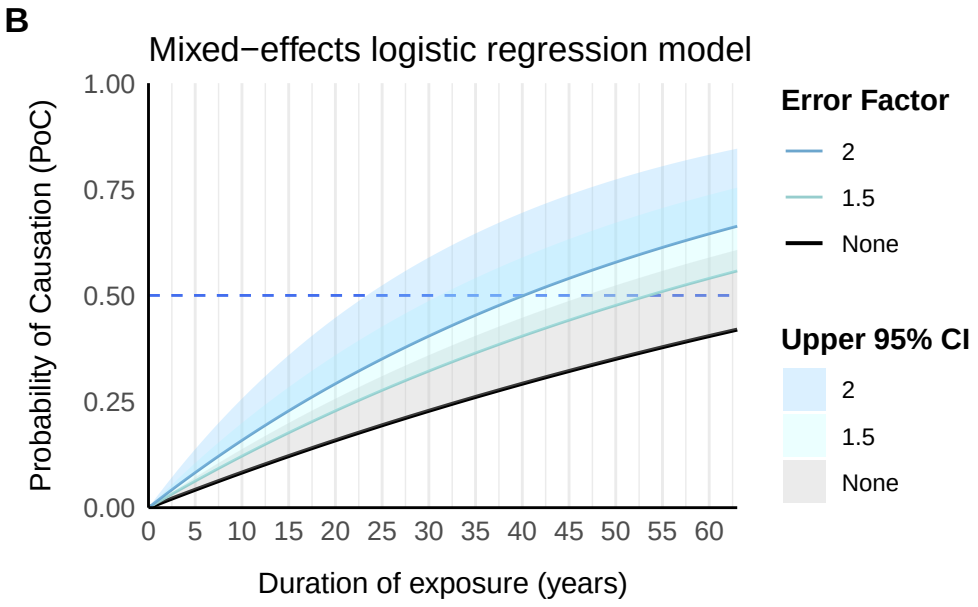
```
mixed_model_dur0 <- glmer(
  status ~ asbestos_dur0 + agegroup + sex + packyrs + time_quit +
    (1 + asbestos_dur0 | study_name),
  data = df,
  family = binomial,
  control = glmerControl(optimizer = "bobyqa"),
  nAGQ = 0
  )
```

The coefficient for exposure (b) is 0.0086 which corresponds to an increase in lung cancer risk of 0.87% per every additional fibre-years.

Out of those ever exposed to asbestos, the summary of PoC with this approach is as follows:

Scenario	N	mean	sd	min	q25	median	q75	max
PoC	14752	0.14	0.10	0.01	0.05	0.12	0.23	0.42
PoC_UB	14752	0.22	0.15	0.01	0.09	0.20	0.36	0.61
PoC_scenario_b1.5	14752	0.20	0.14	0.01	0.07	0.18	0.32	0.56
PoC_scenario_b2	14752	0.25	0.17	0.02	0.10	0.23	0.40	0.66
PoC_scenario_UB1.5	14752	0.31	0.20	0.02	0.12	0.28	0.49	0.75
PoC_scenario_UB2	14752	0.37	0.23	0.03	0.16	0.36	0.59	0.85

Plot of PoC values with the upper bound of the 95% prediction interval, according to duration of asbestos exposure.



Participants with outcome and a PoC greater than or equal to 50% under different scenarios:

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	0	0	NA
PoC_UB	139	82	47
PoC_scenario_b1.5	11	7	54
PoC_scenario_b2	633	375	41
PoC_scenario_UB1.5	1765	1044	32
PoC_scenario_UB2	2739	1621	24

Males

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	0	0	NA
PoC_UB	138	101	47
PoC_scenario_b1.5	11	8	54
PoC_scenario_b2	627	461	41
PoC_scenario_UB1.5	1743	1281	32
PoC_scenario_UB2	2698	1983	24

Female:

Scenario	Cases	Cases per 10,000	Min cum exp
PoC	0	0	NA
PoC_UB	1	3	52
PoC_scenario_b1.5	0	0	NA
PoC_scenario_b2	6	18	42
PoC_scenario_UB1.5	22	67	32
PoC_scenario_UB2	41	124	24

Summary of coefficients

Increase in lung cancer odds per exposure unit

Error Factor	Increase in Odds (%)	Upper Bound Increase (%)
Logistic: fiber-years		
None	3.78	4.97
1.5	5.72	7.55
2	7.70	10.19
Mixed: fiber-years		
None	6.76	17.46
Logistic: years		
None	0.94	1.13
1.5	1.41	1.70
2	1.88	2.27
Mixed: years		
None	0.87	1.49
1.5	1.30	2.25
2	1.74	3.01

References

1 Rothman KJ, VanderWeele TJ, Lash TL. Measures of Effect and Measures of Association. In: Lash TL, VanderWeele TJ, Haneuse S, *et al.*, eds. *Modern epidemiology*. Philadelphia: Wolters Kluwer 2021:79–104.

2 Armstrong B, Tremblay C, Theriault G. Compensating Bladder Cancer Victims Employed in Aluminum Reduction Plants. *Journal of Occupational and Environmental Medicine*. 1988;30:771–5. doi: 10.1097/00043764-198810000-00004

3 Van Der Bij S, Koffijberg H, Lenters V, *et al.* Lung cancer risk at low cumulative asbestos exposure: Meta-regression of the exposure–response relationship. *Cancer Causes & Control*. 2013;24:1–12. doi: 10.1007/s10552-012-0107-7

4 Higgins JPT, Thompson SG, Spiegelhalter DJ. A Re-Evaluation of Random-Effects Meta-Analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*. 2009;172:137–59. doi: 10.1111/j.1467-985X.2008.00552.x

Package References

- Auguie B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3, https://CRAN.R-project.org/package=gridExtra.
- Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01 https://doi.org/10.18637/jss.v067.i01.
- Bates D, Maechler M, Jagan M (2024). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.7-0, https://CRAN.R-project.org/package=Matrix.
- Gohel D, Skintzos P (2024). *flextable: Functions for Tabular Reporting*. R package version 0.9.7, https://CRAN.R-project.org/package=flextable.
- 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/.
- Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J, Brevoort K, Roy O (2024). *gt: Easily Create Presentation-Ready Display Tables*. R package version 0.11.0, https://CRAN.R-project.org/package=gt.
- Kurkiewicz D (2017). *docstring: Provides Docstring Capabilities to R Functions*. R package version 1.0.0, https://CRAN.R-project.org/package=docstring.
- Lüdecke D (2024). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.16, https://CRAN.R-project.org/package=sjPlot.
- 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.
- 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.
- Sera F, Armstrong B, Blangiardo M, Gasparrini A (2019). "An extended mixed-effects framework for meta-analysis." *Statistics in Medicine*, DOI: 10.1002/sim.8362. http://www.ag-myresearch.com/2019_sera_statmed.html.
- 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.