Evaluation of the probability of causation for lung cancer workers' compensation

Screening

Javier Mancilla Galindo

2025-06-05

The machine learning assisted learning tool ASReview[1] (v.1.6.3) was used for the pilot screening phase, whereas v.2.0.2 will be used in the screening stage of the full-length review. The SAFE procedure will be followed as the stopping heuristic for screening.[2]

Full-length review screening

Three percent of the records were randomly screened to determine the number of relevant records expected in the whole dataset, according to the SAFE procedure. The size of the training set (t) is:

$$t = 1129 \times 0.03$$

Thus, **34 records** were randomly reviewed.

There was 1 relevant record (RR_t) found. Thus, the number of relevant records (RR_T) expected in the whole dataset (T) is:

$$RR_T = \frac{RR_t}{t} \times T$$

A total of **34 relevant records** are expected. These will later be used as the prior knowledge in the full dataset used in the scoping review.

According to the SAFE procedure, active learning with an initial lightweight model (ELAS u4, default configuration) will be used according to the following stopping heuristics:

- At least twice the (RR_T) should be screened: **68 records**

- A minimum of 10% of the records should be screened: **113 records**
- No extra relevant records have been identified in the last 25 records.

When these criteria are met, the model will be changed to the ELAS h3 deep learning model (default configuration), to allow the identification of more complex semantic context that can uncover difficult to find records.

The stopping heuristic will be:

- No extra relevant records have been identified in the last 25 records.

Pilot Screening

For the pilot screening phase, 5% of the records from the lung cancer search were be randomly screened to determine the number of relevant records expected in the whole dataset, according to the SAFE procedure. The size of the training set (t) is:

$$t = 453 \times 0.05$$

Thus, **23 records** were randomly reviewed.

There was 1 relevant record (RR_t) found. Thus, the number of relevant records (RR_T) expected in the whole dataset (T) is:

$$RR_T = \frac{RR_t}{t} \times T$$

A total of **20 relevant records** were expected, but only 10 were selected chosen for pilot data extraction. These records were later incorporated as prior knowledge in the full dataset used in the scoping review.

References

1. Van De Schoot R, De Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*. 2021;3(2):125-133. doi:10.1038/s42256-020-00287-7

2. Boetje J, Van De Schoot R. The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic Reviews*. 2024;13(1):81. doi:10.1186/s13643-024-02502-7

Session Information

```
R version 4.4.0 (2024-04-24 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 26100)

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] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] report_0.5.9     gt_0.11.0        overviewR_0.0.13 lubridate_1.9.3
 [5] forcats_1.0.0    stringr_1.5.1    dplyr_1.1.4      purrr_1.0.2
 [9] readr_2.1.5      tidyr_1.3.1      tibble_3.2.1     ggplot2_3.5.1
[13] tidyverse_2.0.0  pacman_0.5.1
```

Package References

- 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.
- 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/.
- Meyer C, Hammerschmidt D (2023). *overviewR: Easily Extracting Information About Your Data*. R package version 0.0.13, https://CRAN.R-project.org/package=overviewR.
- 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/.
- Rinker TW, Kurkiewicz D (2018). *pacman: Package Management for R*. version 0.5.0, http://github.com/trinker/pacman.
- 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.