Evaluation of the probability of causation approach for lung cancer: Scoping review

Deduplication of Studies

Javier Mancilla Galindo

2025-08-04

PoC Final Review Deduplication

There are a total of 650 records in Embase, 497 records in PubMed, and 550 records in OpenAlex.

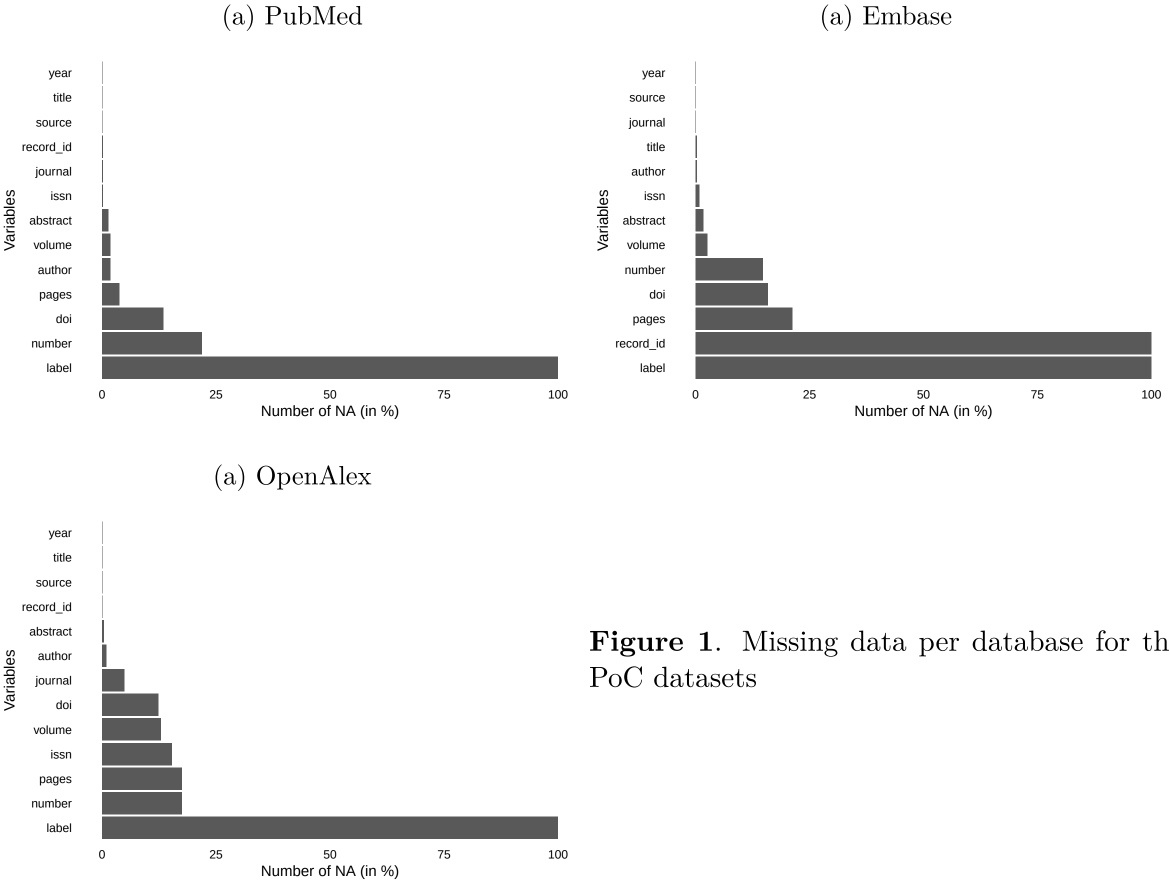
A plot of missing data for corroboration before deduplication is shown in **Figure 1**. Missing data should be lower than 100% for all variables, except for `record_id` and `label`, which are optional. Overall, **PubMed** has the least amount of missing data, so it will be used as the preferred source to keep in the deduplication procedure to minimize the amount of manual corroborations needed.

There are a total of 1697 records. These will be deduplicated using the Automated Systematic Search Deduplicator (ASySD).[1]

Search for remaining duplicates by doi:

```
# A tibble: 1 x 1
      n
  <int>
1   100
```

After deduplication, there are a total of 1129 studies.



(a) PubMed

(a) Embase

(a) OpenAlex

Figure 1. Missing data per database for the PoC datasets

ASReview labels

Pilot Search Deduplication

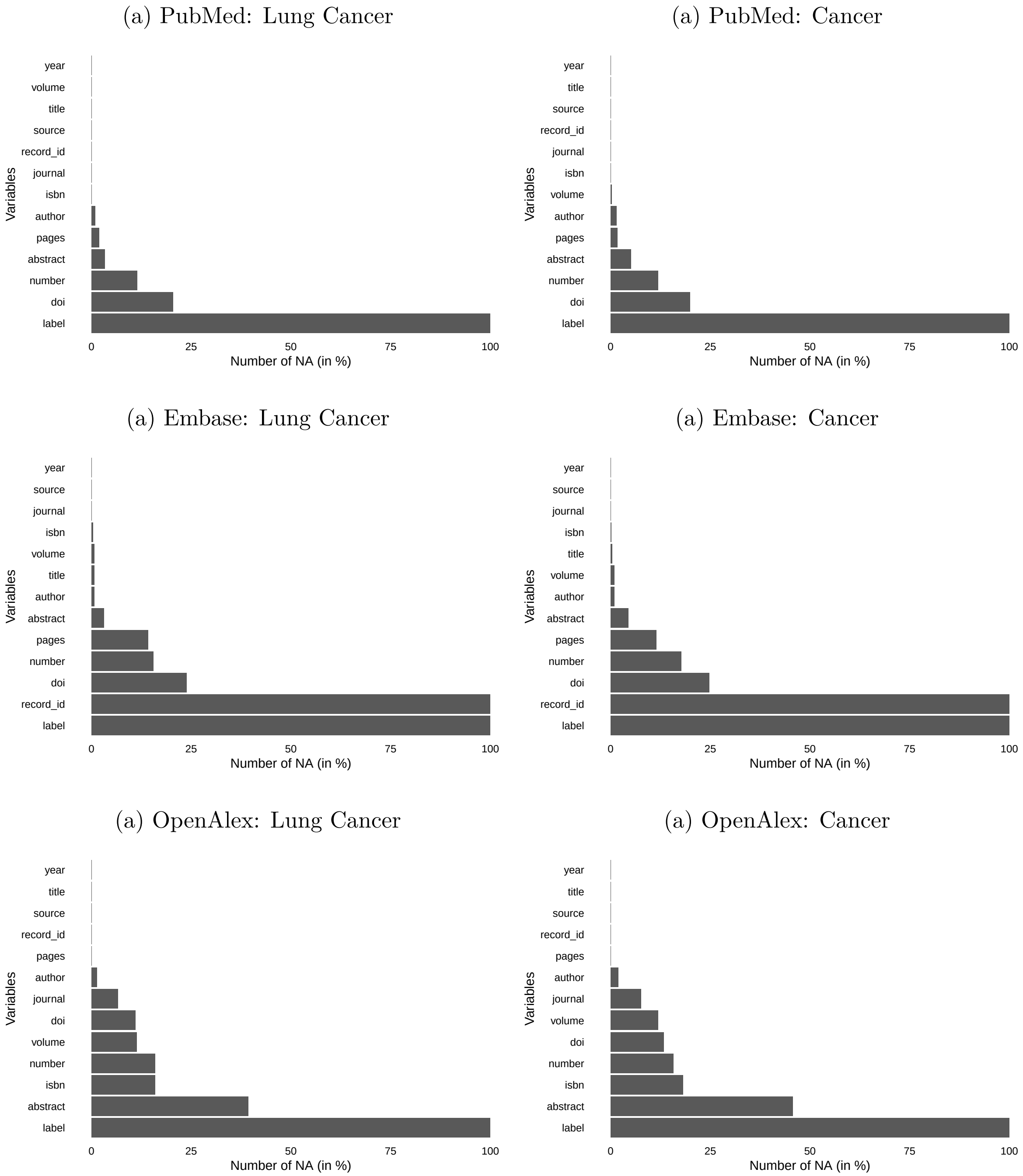
The pilot search strategy was conducted on 2 December 2025 in Embase, PubMed, and OpenAlex. Documentation of search strings is available in the research protocol (protocol-scoping-review.qmd)

There are a total of 289 records in Embase for lung cancer and 559 records for all types of cancer; 210 records in PubMed for lung cancer and 412 records for all types of cancer; and 300 records in OpenAlex for lung cancer and 629 records for all types of cancer.

A plot of missing data for corroboration before deduplication is shown in **Figure 2**. Missing data should be lower than 100% for all variables, except for `record_id` and `label`, which are optional. Overall, **PubMed** has the least amount of missing data, so it will be used as the preferred source to keep in the deduplication procedure to minimize the amount of manual corroborations needed.

There are a total of 799 records for lung cancer and 1600 records for all types of cancer. These will be deduplicated using the Automated Systematic Search Deduplicator (ASySD).[1]

After deduplication, there are a total of 609 studies for the lung cancer search and 1223 records for all types of cancer. However, there are remaining potentially duplicated items to review manually: 36 repeated doi for lung cancer and 67 repeated doi for all types of cancer. I will add a label for manual review to these records.



(a) PubMed: Lung Cancer

(a) PubMed: Cancer

(a) Embase: Lung Cancer

(a) Embase: Cancer

(a) OpenAlex: Lung Cancer

(a) OpenAlex: Cancer

Figure 2. Missing data per database

Manual deduplication

Manual deduplication was carried out in excel by first examining the repeated DOIs and then comparing the records for potential duplicates. Subsequently, empty abstract fields were searched in Google Scholar, WorldCat, and journal websites and added to the records when found. A summary of deduplicated records is as follows:

Field	Lung_Cancer
DOI removed	44
DOI non-duplicates	2
DOI missing	64
Title duplicates	21
Removed (other)	63
Abstract added	50
Abstract not available	21
Total Complete	453

Other records removed corresponded to:

- bibliography lists (1793, 1794, 1796, 1798);
- journal decision letters (1787, 1788, 1789);
- news highlights (1777, 1779, 1784)
- duplicated publication in different journals (1453 kept - 1451 removed; 1421 kept - 1568 removed);
- a record (1079 kept) presented multiple times in different conferences (1092, 1098); and
- uncurated abstracts (n = 49) that would preclude the usefulness of the training set for assisted learning and keyword identification with `litsearchr`.

A total of **453** manually duplicated records for lung cancer will be used for pilot screening.

References

1. Hair K, Bahor Z, Macleod M, Liao J, Sena ES. The automated systematic search deduplicator (ASySD): A rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. *BMC Biology*. 2023;21(1):189. doi:10.1186/s12915-023-01686-z

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] ASySD_0.4.1      report_0.6.1     gt_0.11.0        overviewR_0.0.13
 [5] lubridate_1.9.3  forcats_1.0.0    stringr_1.5.1    dplyr_1.1.4
 [9] purrr_1.0.2      readr_2.1.5      tidyr_1.3.1      tibble_3.2.1
[13] ggplot2_3.5.1    tidyverse_2.0.0  devtools_2.4.5   usethis_3.0.0
[17] 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/.
- Hair K, Bahor Z, Macleod M, Liao J, Sena ES (2021). "The Automated Systematic Search Deduplicator (ASySD): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews." *bioRxiv*. doi:10.1101/2021.05.04.442412 https://doi.org/10.1101/2021.05.04.442412.
- 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, Bryan J, Barrett M, Teucher A (2024). *usethis: Automate Package and Project Setup*. R package version 3.0.0, https://CRAN.R-project.org/package=usethis.
- 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, Hester J, Chang W, Bryan J (2022). *devtools: Tools to Make Developing R Packages Easier*. R package version 2.4.5, https://CRAN.R-project.org/package=devtools.
- Wickham H, Vaughan D, Girlich M (2024). *tidyr: Tidy Messy Data*. R package version 1.3.1, https://CRAN.R-project.org/package=tidyr.