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

Snowball citation retrieval

Javier Mancilla Galindo

2025-08-12

Table of contents

# OpenAlex snowball citation retrieval

The full dataset screened with ASReview includes n = 1128 records, out of which n = 23 were labeled as relevant.

After fetching metadata from OpenAlex for the 23 relevant records using their DOIs, a total n = 22 records with complete OpenAlex metadata were identified. This means that one record[1](#ref-Choudat2003) from the relevant dataset was not included for snowball citation retrieval.

The openalexR package was used to perform citation snowballing to identify additional relevant studies by searching both cited references and citing papers for the relevant records and converted snowball search results into a structured dataframe format for further processing. The initial snowball dataset contains n = 1058 records.

Retracted papers were excluded and only articles and preprints were kept, resulting in the exclusion of 1 review paper from the initial set of relevant papers,[2](#ref-guidotti2002) for which nonetheless all its relevant citations had already been obtained. Therefore, the size of the snowball dataset was n = 840 records, out of which:

* seed papers (relevant in ASReview dataset) n = 21,
* citing papers (forward citations) n = 445,
* referenced papers (backward citations) n = 366, and
* papers connected in both directions n = 8.

The original OpenAlex records from the PoC search (n = 550) was loaded to deduplicate previously identified records.

Those records in the original OpenAlex dataset screened with ASReview were removed from the new snowball citation dataset, resulting in (n = 813) remaining records.

# Deduplication

The combined set of studies (PubMed + Embase + OpenAlex) screened with ASReview (n = 1128) with the snowball citations (n = 813) contains **n = 1941 records**. These will be deduplicated using the Automated Systematic Search Deduplicator (ASySD).[3](#ref-ASySD2023)

Automatic deduplication resulted in n = 1927 unique records and n = 12 potential duplicates requiring manual review.

After manual deduplication, the dataset contained n = 1917 records.

Remaining duplicate records sharing the same DOI (n = 6) were identified.

Hierarchical deduplication by DOI, title, and abstract with source priority (PubMed > Embase > OpenAlex) was appliead and HTML tags from abstracts were cleaned, followed by filtering out records with malformed abstracts. Subsequently, new records from snowball search (those with URL-based IDs) were identified and processed, resulting in the removal of those already present in the ASReview dataset, followed by assigning new sequential duplicate IDs. As a result, there were n = 588 new snowball records to be added.

# Final dataset

The original dataset was combined with the retrieved snowball records to create the final dataset. The final combined dataset has **n = 1716 records** (original: n = 1128, new: n = 588).

# References

1. Choudat D. [Risk, etiologic fraction and causation probability in case of multiple exposures]. *Arch Mal Prof Med Trav*. 2003;64(6):279-285.

2. Guidotti TL. Apportionment in Asbestos-Related Disease for Purposes of Compensation. *Industrial Health*. 2002;40(4):295-311. doi:[10.2486/indhealth.40.295](https://doi.org/10.2486/indhealth.40.295)

3. 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](https://doi.org/10.1186/s12915-023-01686-z)

# Session

R version 4.4.2 (2024-10-31 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\_0.6.1 gt\_1.0.0 ASySD\_0.4.5.9000 openalexR\_1.4.0   
 [5] readxl\_1.4.5 lubridate\_1.9.4 forcats\_1.0.0 stringr\_1.5.1   
 [9] dplyr\_1.1.4 purrr\_1.1.0 readr\_2.1.5 tidyr\_1.3.1   
[13] tibble\_3.3.0 ggplot2\_3.5.2 tidyverse\_2.0.0 devtools\_2.4.5   
[17] usethis\_3.1.0 pak\_0.9.0

# Package References

* Csárdi G, Hester J (2025). *pak: Another Approach to Package Installation*. R package version 0.9.0, <https://CRAN.R-project.org/package=pak>.
* 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 (2025). *gt: Easily Create Presentation-Ready Display Tables*. R package version 1.0.0, <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/>.
* Massimo A, Le Trang, Corrado C, Alessandra B, June C (2024). “openalexR: An R-Tool for Collecting Bibliometric Data from OpenAlex.” *The R Journal*, *15*, 167-180. ISSN 2073-4859, doi:10.32614/RJ-2023-089 <https://doi.org/10.32614/RJ-2023-089>.
* Müller K, Wickham H (2025). *tibble: Simple Data Frames*. R package version 3.3.0, <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/>.
* 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 (2025). *readxl: Read Excel Files*. R package version 1.4.5, <https://CRAN.R-project.org/package=readxl>.
* Wickham H, Bryan J, Barrett M, Teucher A (2024). *usethis: Automate Package and Project Setup*. R package version 3.1.0, <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 (2025). *purrr: Functional Programming Tools*. R package version 1.1.0, <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>.