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

Deduplication of Studies

Javier Mancilla Galindo

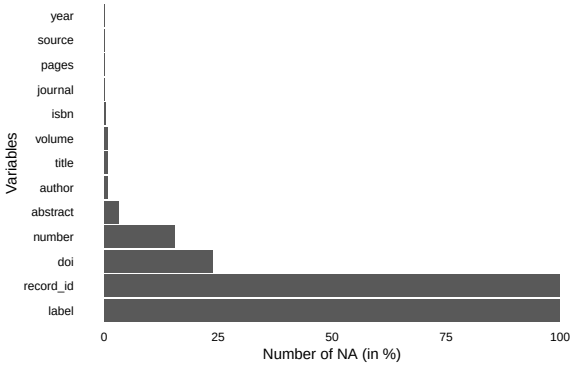
2024-12-03

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.

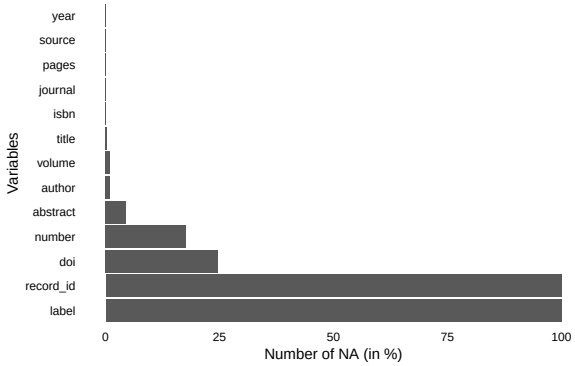
A plot of missing data for corroboration before deduplication is shown in the Figure. Missing data should be lower than 100% for all variables, except for `record_id` and `label`, which are optional.

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]

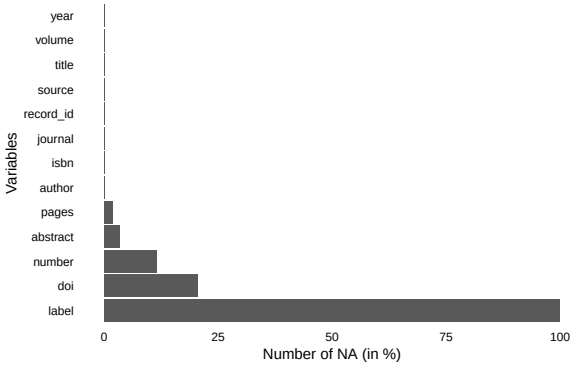
After deduplication, there are a total of 609 studies for the lung cancer search and 1223 records for all types of cancer. However, there are remaining potentially duplicated items to review manually: 36 repeated doi for lung cancer and 68 repeated doi for all types of cancer. I will add a label for manual review



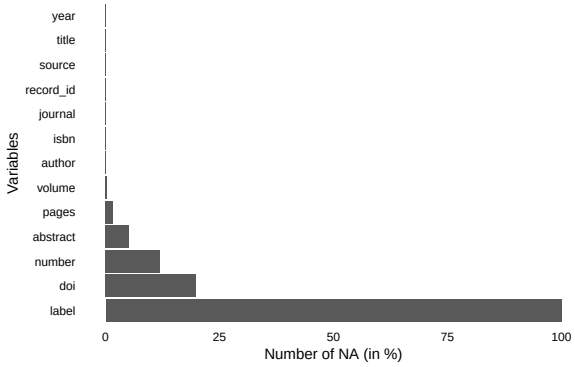
(a) EMBASE: Lung Cancer



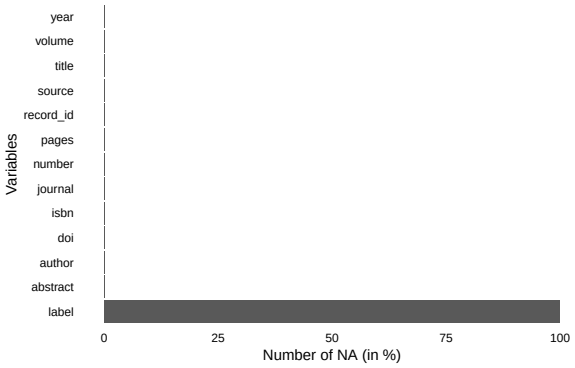
(a) EMBASE: Cancer



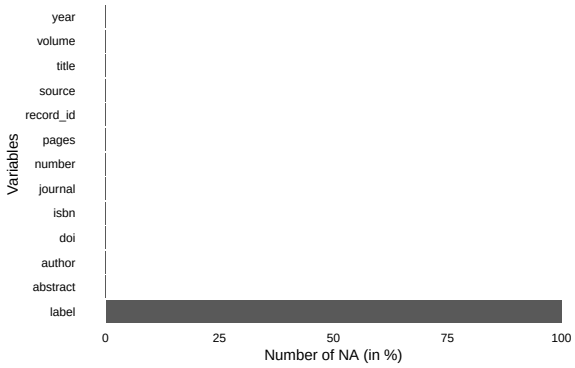
(a) PubMed: Lung Cancer



(a) PubMed: Cancer



(a) OAlex: Lung Cancer



(a) OAlex: Cancer

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 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] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] ASySD_0.4.1      report_0.5.9     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.