Identifying Patients with Lung Cancer from Electronic Health Records: Systematic Evidence Review to Bridge the Gap between Research and Real-World Impact

A. R. Stevens[1], J. R. Malinowski[2], R. T. Levinson[3], M. Chapman[4], S. M. Manemann[5], M. P. Wilson[6], S. J. Bielinski[5], L. V. Rasmussen[7], L. K. Wiley[6]

Introduction

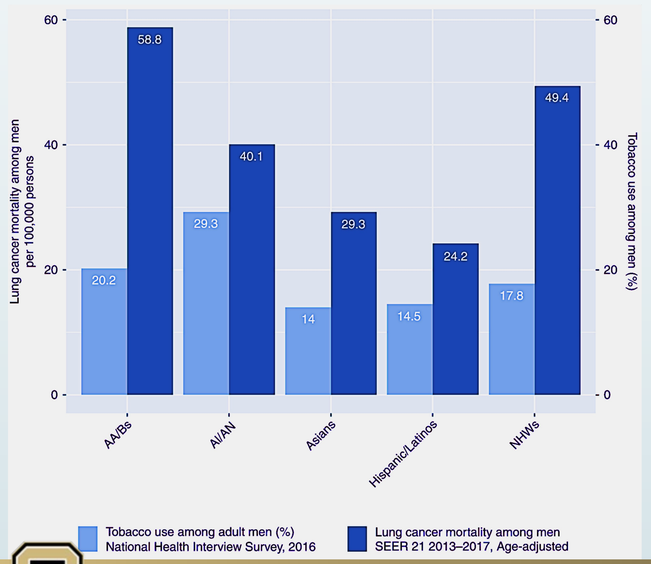
Lung cancer is the leading cause of cancer-related deaths in the United States.

Screening and prevention guidelines are not historically developed on diverse populations.

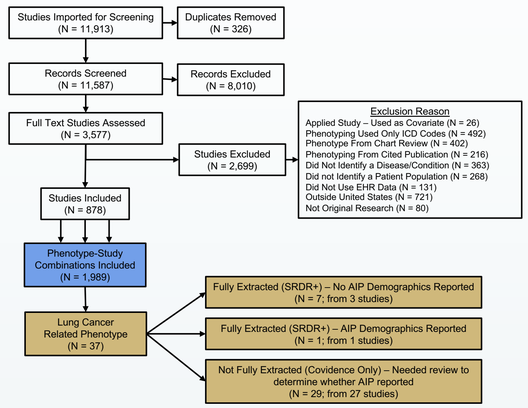
Electronic health record (EHR) data and phenotyping algorithms are increasingly used to identify patients with disease.

It is unclear how well populations from EHR studies align with the known prevalence of disease.

Undetected misrepresentation from algorithm-identified populations may propagate inequities in lung cancer research and policy.



Methods



Searched PubMed for articles mentioning EHRs and terms related to automated cohort identification.

Phenotype-study combinations were identified for each study and filtered to those related to lung cancer.

Extracted demographic variables included: age, sex, gender, race, ethnicity, and ancestry where available.

Results

30 unique phenotype-study pairs were found in EHR studies that identify patients with lung cancer.
Of these, 12 (40%) reported any demographics of their algorithm-identified lung cancer populations.

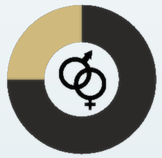
Reporting frequency:

- Sex was the most frequently reported demographic variable (n = 10), followed by age (n = 9), and race/ethnicity (n = 8).
- No algorithms reported gender identity and two algorithms reported genetic ancestry.

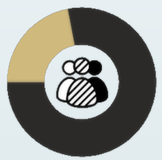
Reporting variability:

- Where reported, race/ethnicity had the most unique data labels (n = 23).
- Age had the greatest variability in reporting techniques (n = 3).

40%
Reported any demographic data



33%
Reported sex



27%
Reported race or ethnicity

7%
Reported genetic ancestry

Conclusions

While many studies acknowledge the importance of demographic data (e.g., age, sex, race), these same features are often omitted when describing the specific populations algorithms identify.

Consequently, current reporting practices make it difficult to understand the generalizability of study results.

These findings prompt a compelling need for standardized demographic reporting, which will amplify research impact through transparency and a greater ability to combat bias in lung cancer research and the clinical guidelines they inform.

References

1. Hripcsak G, Albers D. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association : JAMIA. 2012;20. doi:10.1136/amiajnl-2012-001145
2. Malhotra J, Paddock LE, Lin Y, et al. Racial disparities in follow-up care of early-stage lung cancer survivors. J Cancer Surviv. Published online March 22, 2022. doi:10.1007/s11764-022-01184-1
3. Pinheiro LC, Groner L, Soroka O, et al. Analysis of Eligibility for Lung Cancer Screening by Race After 2021 Changes to US Preventive Services Task Force Screening Guidelines. JAMA Netw Open. 2022;5(9):e2229741. doi:10.1001/jamanetworkopen.2022.29741
4. Rethinking Clinical Trials. Accessed August 16, 2023. https://rethinkingclinicaltrials.org/chapters/conduct/electronic-health-records-based-phenotyping/electronic-health-records-based-phenotyping-introduction/
5. US Preventive Services Task Force. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. JAMA. 2021;325(10):962-970. doi:10.1001/jama.2021.1117
6. USCS Data Visualizations. Accessed August 16, 2023. https://gis.cdc.gov/grasp/USCS/DataViz.html
7. Zavala VA, Bracci PM, Carethers JM, et al. Cancer health disparities in racial/ethnic minorities in the United States. Br J Cancer. 2021;124(2):315-332. doi:10.1038/s41416-020-01038-6

Acknowledgements

Mentorship and funding provided in part by the CUSOM Research Track.
Thank you to Sean Davis, MD PhD for his mentorship and intellectual contributions.

Affiliations:
1. School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States.
2. Write InScite LLC, South Salem, NY, United States.
3. Clinic for General Internal Medicine and Psychosomatics, Heidelberg University Hospital, Heidelberg, Germany.
4. Department of Population Health Sciences, King's College London, London, United Kingdom.
5. Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States.
6. Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, United States.
7. Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, United States.

University of Colorado Anschutz Medical Campus