Hematology and Oncology
Abstract

IDENTIFYING PATIENTS WITH LUNG CANCER FROM ELECTRONIC HEALTH RECORDS: A SYSTEMATIC EVIDENCE REVIEW TO BRIDGE THE GAP BETWEEN RESEARCH AND REAL-WORLD IMPACT

AUTHORS: 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]

INSTITUTIONS:
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.

Purpose of Study: Lung cancer is the leading cause of cancer-related deaths in the United States. Nevertheless, current lung cancer guidelines are not always developed on diverse populations. Researchers increasingly use electronic health record (EHR) data to inform strategies for disease prevention, treatment, and monitoring. These studies require phenotyping algorithms to identify patients with lung cancer, but it is unclear how well these algorithm-identified populations align with the known prevalence of the disease. Our study aims to address this knowledge gap, which currently poses a potential source of bias and misrepresentation of lung cancer identification that may propagate healthcare inequities in research and policy.

Methods Used: We conducted a sub-analysis of all lung cancer phenotyping algorithms identified by a systematic evidence review (SER) of US-based EHR phenotyping algorithms. The larger study searched PubMed for articles published through August 10, 2022 that mentioned both EHRs and one or more terms for automated cohort identification methods. All studies were reviewed in duplicate with a standardized protocol to assess inclusion/exclusion and extract key variables related to algorithm quality. In this study, a single reviewer performed additional extraction of demographic variables for lung cancer algorithms including: age, sex/gender, and race/ethnicity/ancestry.

Summary of Results: A total of 11,913 studies were identified with 856 studies included after review. This analysis extracted data from 29 studies that identified one or more lung cancer related phenotypes, representing 30 distinct phenotype-study pairs. Of these, 12 (40%) reported any demographics of their algorithm-identified lung cancer population. 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. Where reported, race/ethnicity had the most unique data labels (n = 23), while age had the greatest variability in reporting techniques. We are in the process of analyzing whether each algorithm included additional study specific filters that may influence the representativeness of the algorithm- identified population.

Conclusions: This study provides insights into the current state of demographic reporting for algorithm-identified lung cancer populations. 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.