Editorial Commentary

Toward multi-domain lung cancer research databases: opportunities and challenges from MCC-MELD

Yuki Joyama[1]^, Stephanie Tuminello[1,2,3]^, Emanuela Taioli[1,2,3]^

[1]Institute for Translational Epidemiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [2]Department of Thoracic Surgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [3]Tisch Cancer Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Correspondence to: Emanuela Taioli, MD, PhD. Institute for Translational Epidemiology, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1133, New York, NY 10029, USA; Department of Thoracic Surgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA; Tisch Cancer Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA. Email: emanuela.taioli@mountsinai.org.

Comment on: Cui L, Lee J, Miller J, *et al.* Design and Creation of a Racially Diverse Lung Cancer Registry with Detailed Genomic and Environmental Annotation. Cancer Epidemiol Biomarkers Prev 2026;35:49-58.

Keywords: Lung cancer; registry; environmental exposure; electronic health records (EHRs); tumor genomics

Submitted Feb 24, 2026. Accepted for publication Apr 23, 2026. Published online May 27, 2026.
doi: 10.21037/jtd-2026-1-0470
View this article at: https://dx.doi.org/10.21037/jtd-2026-1-0470

Lung cancer remains a leading cause of cancer death worldwide and in the United States (1). While tobacco prevention and control measures have led to declining incidence and mortality rates, the focus of lung cancer research has expanded beyond smoking exposure to include environmental factors such as radon, occupational carcinogens, and fine particulate matter 2.5 (PM2.5), as well as precision medicine encompassing genomic background and treatment responsiveness (2-4). Therefore, the current challenge lies not only in evaluating single interventions but also in establishing an integrated data infrastructure capable of handling prevention, early detection, molecular targeted therapy, while identifying socio-environmental risk factors within a unified analytical framework.

The Meyer Cancer Center Molecularly Enhanced Lung Cancer Database (MCC-MELD), assembled by Cui *et al.*, addresses this challenge by placing electronic health records (EHRs) at its core, integrating tumor registries and in-hospital genomic testing results, and linking patients' residential locations to area-level environmental and social indicators (5). In doing so, MCC-MELD offers a concrete blueprint for the kind of data infrastructure needed to advance lung cancer research. Despite several methodological limitations, including heterogeneity in data capture across institutions and variation in molecular testing and documentation, this database has important strengths. The plan to repeat external validation, improve outcome completeness, incorporate events that occurred outside the health system, and refine spatial analytic approaches would further enhance the reliability and robustness of this resource. Conceptually, MCC-MELD represents a deliberate effort to overcome the familiar trade-offs of existing databases, which are either large-scale but clinically and molecularly shallow, clinically rich but genomically sparse, or genomics-heavy but drawn from relatively homogeneous populations. Very few such efforts are available, and are summarized in *Table 1* to provide additional context on their respective strength and limitations (5-13).

MCC-MELD accomplishes this by utilizing newer methodologies, specifically by applying natural language processing (NLP) to clinical notes and other unstructured text to abstract tumor molecular data, including epidermal growth factor receptor (*EGFR*) mutation status. While authors were able to report high accuracy, should this approach be scaled for a multi-institutional database, several limitations become more salient. NLP performance can vary across institutions because documentation practices

^ ORCID: Yuki Joyama, 0009-0004-0700-2225; Stephanie Tuminello, 0000-0002-4011-768X; Emanuela Taioli, 0000-0001-5734-4806.

Table 1 Comparison of major data sources for cancer research

Parameters	MCC-MELD (5)	FMI-CGDB (6,7)	MSK-CHORD (8)	AACR-GENIE (9,10)	NCI-CRDC (11)	STS GTSD (12,13)
Number of lung cancer patients	9,573	14,230	7,674	30,801	50,270 (GDC)	103,407
Population	3 New York Presbyterian—affiliated sites in New York City	~280 US cancer clinics (80% community, 20% academic)	A single academic center, Memorial Sloan Kettering	20 academic medical centers (including 5 international sites)	Multiple NCI-funded programs (e.g., TCGA, CPTAC)	300+ participating hospitals (across North America)
Race/ethnic breakdown	White 49.3%, Black or African American 13.7%, Asian 19.8%, Hispanic 7.4%, other 9.9%	White 68.1%, Black or African American 6.0%, Asian 2.9%, other 14.6%, unknown 8.3%	White 79.2%, Black or African American 5.4%, Asian 9.7%, other 2.1%, unknown 3.7%	White 61.1%, Black or African American 6.5%, Asian 6.7%, other 15.4%, unknown 10.4% (lung cancer cohort)	White 20.3%, Black or African American 2.4%, Asian 2.8%, other <0.1%, unknown 74.6% (lung cancer cohort from GDC)	White 85.6%, Black or African American 8.6%, Asian 3.9%, other 2.2%
Clinical data	EHR, institutional tumor registry (ONCOLog), manual research. Includes unstructured notes	EHR; unstructured documents curated by technology-enabled abstraction	Demographics and treatment from EHR; NLP annotations of radiology, pathology, and clinician notes	Limited clinical data mapped to OncoTree ontology (sex, race, age, cancer type, sample type)	Multiple commons; most commonly through the GDC and CTDC	Manually abstracted from hospital records; includes detailed comorbidities and surgery variables
Tumor genomics	In-house (cBioPortal). Targeted panels (Oncomine, TSO500). Focus on somatic mutations (e.g., EGFR). NLP used to extract status from notes where structured data is missing	Comprehensive genomic profiling (>300 genes). Somatic mutations, TMB, MSI status	MSK-IMPACT. Targeted sequencing (341–468 genes)	CLIA/ISO-certified NGS data (panels vary by center, 50-500 genes). Somatic mutations, copy-number, structural rearrangements	GDC; available data includes WGS, WXS, RNA-Seq, methylation array, ATAC-Seq, etc. Germline variant available for some projects (e.g., TCGA)	Available for cases where molecular testing was performed for clinical purposes.
Environmental exposures	NLP-extracted smoking history. Geocoded addresses linked to air pollution (PM2.5), ADI, and census data	EHR social history (e.g., smoking status) available via abstraction	NLP-derived smoking status. Other exposure information is limited	None	Dataset specific (e.g., smoking status, pack-years, asbestos exposure available on GDC and PDC)	Detailed smoking history; lacks environmental data
Follow-up	Available for participants followed at one of the participating sites	Available for participants followed within the Flatiron system	Available for participants followed at MSK	Longitudinal data are not collected in a standardized way	Varies (e.g., clinical trials *vs.* cross-sectional molecular studies)	30-day post-operative outcomes. Long-term survival available through NDI linkages
Outcomes	Death within hospital systems. Recurrence and other endpoints planned for future curation	Mortality (EHR + external data), real-world progression, response to therapy	Overall survival, time-to-treatment failure, time-to-metastasis for specific organs (CNS, liver, bone)	Overall survival, disease-free survival	Mortality, disease-specific survival (available via CDA search if present in source study)	30-day mortality, post-operative complications, re-operation rates, length of hospital stay
Key strength	Geocoding enables fine-scale environmental linkage; NLP captures missing genomic and smoking data	Represents community oncology settings as opposed to academic cohorts	Large single-institution cohort with uniform sequencing; NLP automated extraction of metastatic sites	Large sample size (>100k goal); covers rare cancers; real-world but from expert tertiary centers	Connects genomics, proteomics, and imaging; cloud-based infrastructure	Standardized definitions for surgical quality assessment
Limitations	Generalizability limited to NYC; cross-site EHR heterogeneity due to different adoption times	Proprietary; limited to sequenced patients (selection bias)	Potential referral bias (tertiary center); cohort entry defined by sequencing date (immortal time bias)	Variable gene panels across institutions; manual curation required for detailed outcomes	Diverse data models/ontologies make harmonization difficult	Only surgical patients. Selection bias due to voluntary participation. Data missingness varies across participating institutions

AACR-GENIE, American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange; ADI, area deprivation index; ATAC-Seq, assay for transposase-accessible chromatin with high-throughput sequencing; CDA, Cancer Data Aggregator; CLIA/ISO, Clinical Laboratory Improvement Amendments/International Organization for Standardization; CNS, central nervous system; CPTAC, Clinical Proteomic Tumor Analysis Consortium; CTDC, Clinical Trial Data Commons; EGFR, epidermal growth factor receptor; EHR, electronic health record; FMI-CGDB, Flatiron Health and Foundation Medicine Clinico-Genomic Database; GDC, Genomic Data Commons; IMPACT, Integrated Mutation Profiling of Actionable Cancer Targets; MCC-MELD, Meyer Cancer Center Molecularly Enhanced Lung Cancer Database; MSI, microsatellite instability; MSK-CHORD, Memorial Sloan Kettering Cancer Center Clinico-Genomic Harmonized Oncologic Real-World Dataset; NCI-CRDC, National Cancer Institute Cancer Research Data Commons; NDI, National Death Index; NGS, next-generation sequencing; NLP, natural language processing; PDC, Proteomic Data Commons; PM2.5, particulate matter 2.5; RNA-Seq, RNA sequencing; STS GTSD, Society of Thoracic Surgeons General Thoracic Surgery Database; TCGA, The Cancer Genome Atlas; TMB, tumor mutation burden; WGS, whole genome sequencing; WXS, whole exome sequencing.

differ; molecular results are reported with great variability, which ranges from structured fields to scanned documents and external laboratory formats. Smoking history, often missing or incompletely reported in existing databases, adds another layer of complexity because it is often self-reported, inconsistently recorded, and occasionally inaccurate. For these reasons, the planned external validation at Thomas Jefferson University and the University of Pennsylvania is essential. It will support generalizability and help determine whether NLP-derived variables can function as a consistent approach for multi-center clinical research.

Even if NLP or similar methods could be optimized, we must acknowledge that EHR data is inherently limited for clinical research purposes. For example, in the MCC-MELD database, more than 64% of cases had unknown *EGFR* status, which likely reflects variation in testing uptake (14). Testing follows care pathways, so institution, insurance type, and facility access can all influence who does or does not receive testing. Moreover, molecular testing became standardized relatively recently, so earlier cohorts tend to have higher missingness (15). This represents a substantial challenge to EHR-based research infrastructure, solutions to which must be rooted in enhanced data collection during clinical care, rather than on the backend during database curation.

Similarly, data on occupational and environmental exposures are limited at best or completely missing at worst from the EHR. Most often, residential addresses are used to estimate geography-based (i.e., zip code-level) environmental exposures, but rarely are individual-level exposures captured (16). Geography-based indicators such as PM2.5 and area deprivation index (ADI) are ecological measures that do not capture an individual's specific exposure, nor their residential mobility, commuting exposure, or occupational exposure. MCC-MELD attempted to visualize geographic disparities by linking residential location data to census tract-level environmental and social indicators. While this approach frames lung cancer risk within the broader context of place, where exposure and social conditions may shape both disease risk and downstream outcomes, careful framing of this effort is needed to avoid overinterpreting these maps as direct evidence of causality. MCC-MELD is a facility-based database, so the spatial distribution of cases may reflect referral patterns. Differences in healthcare access across neighborhoods, including variation in insurance coverage, healthcare utilization, and structural barriers to care, may also influence which patients are captured in the registry. As a consequence, in the mapped rates, the numerator represents cases diagnosed or treated within the participating health system, whereas the denominator reflects the underlying general population within each census tract. Together, these factors may lead to over- or underestimation of disease burden. Regional differences in age structure can further distort crude rates, underscoring the importance of age standardization.

The authors outlined a set of planned expansions that would shift MCC-MELD from a single health system cohort toward a more comprehensive research resource in an effort to capture longitudinal outcomes, such as complications, recurrence, or death, which may be happening across multiple institutions. This matters because lung cancer trajectories vary across molecular subtypes, and many clinically meaningful endpoints are poorly represented in routine data. Examples include time to next treatment, reasons for treatment discontinuation, treatment-related toxicity, and reasons for loss to follow-up. If MCC-MELD is intended to represent real-world practice patterns, outcome completeness and traceability should be treated as primary performance targets.

There are several additional extensions beyond those proposed by the authors that would strengthen MCC-MELD and other lung cancer research databases. These include: (I) integration of germline genetic data to study inherited susceptibility and gene-environment interactions; (II) strengthening cause-specific outcome evaluation through additional data linkages, such as cancer registries and the national death index; (III) integrating radiological imaging along with tumor mutations and environmental exposures, to explore whether exposures modify or mediate the relationship between genotype and imaging findings.

EHR-based databases are needed as they capture real-world patient populations, especially older adults, where treatment decisions are affected by age, comorbidities, frailty, and the availability of social support; MCC-MELD and other EHR-based databases can support more systematic pre-trial screening and can help externally validate findings from other clinical trials. Consequently, this can narrow the gap between clinical trial cohorts and the populations treated in practice and can facilitate real-world observational studies that complement randomized clinical trials.

In summary, MCC-MELD represents the future of lung cancer data, which will increasingly rely on EHR infrastructure. Examining MCC-MELD as an example offers insights into future directions that could create more comprehensive, robust data resources across

multiple institutions. The goal is to create databases capable of supporting actionable evidence for lung cancer epidemiology and precision medicine in a single analytic framework, from prevention and early detection through treatment and survivorship.

Acknowledgments

None.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Journal of Thoracic Disease*. The article has undergone external peer review.

Peer Review File: Available at https://jtd.amegroups.com/article/view/10.21037/jtd-2026-1-0470/prf

Funding: None.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jtd.amegroups.com/article/view/10.21037/jtd-2026-1-0470/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Siegel RL, Kratzer TB, Giaquinto AN, et al. Cancer statistics, 2025. CA Cancer J Clin 2025;75:10-45.
2. Liu Y, Fu C, Li Y, et al. Uncovering hidden dangers in urban housing: Sources of indoor radon and associated health risks. J Environ Manage 2025;387:125899.
3. Olsson A, Bouaoun L, Schüz J, et al. Lung Cancer Risks Associated with Occupational Exposure to Pairs of Five Lung Carcinogens: Results from a Pooled Analysis of Case-Control Studies (SYNERGY). Environ Health Perspect 2024;132:17005.
4. Hamanaka RB, Mutlu GM. Particulate matter air pollution: effects on the respiratory system. J Clin Invest 2025;135:e194312.
5. Cui L, Lee J, Miller J, et al. Design and Creation of a Racially Diverse Lung Cancer Registry with Detailed Genomic and Environmental Annotation. Cancer Epidemiol Biomarkers Prev 2026;35:49-58.
6. Agarwala V, Khozin S, Singal G, et al. Real-World Evidence In Support Of Precision Medicine: Clinico-Genomic Cancer Data As A Case Study. Health Aff (Millwood) 2018;37:765-72.
7. Snow T, Snider J, Comment L, et al. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron health-Foundation Medicine Clinico-Genomic Databases, Flatiron Health Research Databases, and the National Cancer Institute SEER population-based cancer registry. medRxiv. 2023.01.03.22283682. Available online: http://dx.doi.org/10.1101/2023.01.03.22283682
8. Jee J, Fong C, Pichotta K, et al. Automated real-world data integration improves cancer outcome prediction. Nature 2024;636:728-36.
9. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov 2017;7:818-31.
10. News and updates. American Association for Cancer Research (AACR). 2022 [cited 2026 Feb 19]. Available online: https://www.aacr.org/professionals/research/aacr-project-genie/news-updates/
11. Brady A, Charbonneau A, Grossman RL, et al. NCI Cancer Research Data Commons: Core Standards and Services. Cancer Res 2024;84:1384-7.
12. Alex GC, Engelhardt K, Rajaram R, et al. The Society of Thoracic Surgeons General Thoracic Surgery Database: 2025 Annual Update. Ann Thorac Surg 2026;121:521-31.
13. Watkins AA, Parsons N, Shehata DG, et al. Race, Ethnicity, and Payor status Impact the Surgical Approach and Postoperative Outcomes of Lung Cancer Amongst Patients within The Society of Thoracic Surgeons General Thoracic Surgery Database. Ann Surg 2025. [Epub ahead of print]. doi: 10.1097/SLA.0000000000006670.
14. Tuminello S, Turner WM, Untalan M, et al. Racial and socioeconomic disparities in non-small cell lung

cancer molecular diagnostics uptake. J Natl Cancer Inst 2025;117:112-9.
15. Lindeman NI, Cagle PT, Aisner DL, et al. Updated Molecular Testing Guideline for the Selection of Lung Cancer Patients for Treatment With Targeted Tyrosine Kinase Inhibitors: Guideline From the College of American Pathologists, the International Association for the Study of Lung Cancer, and the Association for Molecular Pathology. J Mol Diagn 2018;20:129-59.
16. Craver A, Luo J, Kibriya MG, et al. Air quality and cancer risk in the All of Us Research Program. Cancer Causes Control 2024;35:749-60.

Cite this article as: Joyama Y, Tuminello S, Taioli E. Toward multi-domain lung cancer research databases: opportunities and challenges from MCC-MELD. J Thorac Dis 2026;18(5):555. doi: 10.21037/jtd-2026-1-0470