Cancer Phenotyping Pitfalls in EHR: The case of Non-Small Cell Lung Cancer

Asieh Golozar, Martin Lavallee, Adam Black, Darya Kosareva, Michael Gurley, Christian Reich

Background

The goal of phenotyping is to identify patients with certain conditions of interest from disparate EHR data with high validity and reliability. Inherent nuances of data and differences in data types, settings and capture conventions, however, can impact the validity of the phenotypes and bias study results. These limitations are:

- Lack of adequate standard cancer vocabularies or coding schemes.
- Lack of precision for generic condition vocabularies or coding schemes. For example, ICD-10 does not provide any details about the histology of the disease. Only affected organs and their topography are captured.
- A changing disease over time (going through remission, progression, changing histology and malignancy) requiring updating of the diagnosis.
- Use of "working diagnosis" during diagnostic workup. Such diagnosis may or not may hold up eventually after all diagnostic tests have been completed.

To overcome this, researchers use a mixture of clinical logic, tacit postulations about the availability of the data, and the certainty of a record to define a specific condition. The extent to which any of these decisions can impact the phenotype performance is unknown, but depends on the phenotype of interest, the level of details required for accurate ascertainment of the events, how they are used (exposure, outcome, or study population), the research question, and the source of data.

Cancer is a special case where the performance of rule-based phenotypes might be more heavily impacted by these factors compared to other parts of medicine. Cancer diagnoses are explicitly defined through a set of attributes: histology, anatomic site of origin (topology), stage, grade, and cancer-specific biomarkers. This level of detail, however, is not readily available in structured electronic health record (EHR) data and needs to be either extracted from the patient records or augmented through linkage with other sources of information such as tumor registries. Cancer conditions are reported at a higher level without detailed information on the histology or other behaviors of the tumor that are needed for precise identification of the phenotype of interest.

The workarounds researchers apply to overcome these issues are risky; they often use a combination of drugs indicated for the treatment of the specific cancer subtype, evidence of targeted diagnostic and/or prognostic tests, and lack of information on other potential clinical events to define cancer phenotypes. This approach is based on several assumptions that are not necessarily true:

- Assumption 1: *Cancer patients receive treatment.* Despite the availability of different treatment guidelines, a substantial proportion of patients with cancer do not receive any treatment. This figure varies by tumor type and tumor stage. The untreated have shown to differ in age, comorbidities, access to care and severity of the disease. This group represents a specific sub-population who are not represented clinical trials and understudied in terms of their characteristics and outcomes. A phenotyping approach based on treatment would consequently lack generalizability and cannot provide comprehensive evidence on the entire cancer population.

- Assumption 2: *Cancer patients are being treated according to the guidelines.* Several studies have shown significant variation in adherence to cancer treatment guidelines. Patient and provider factors, treatment modalities, and introduction of new treatments are among the factors influencing adherence. At the same time, there is significant overlap in the treatment approaches for multiple cancers complicating the use of treatments in defining cancer phenotypes.
- Assumption 3: *Information on all potential treatment strategies for cancer is reliably available in all data sources.* Different data sources provide different and unique snapshots of the patient journey with cancer patients treated in different facilities. This leads to the availability of fragmented information on the patient journey across data sources. For example, a newly diagnosed prostate cancer is usually managed by urologist through surgery and other procedures such as radiotherapy. These patients would only be treated by oncologists, and show up in their EHR, when the disease is advanced and requires systemic anti-neoplastic treatment.

In this work, we aim to assess the potential impact of the currently used rule-base definitions on the performance of cancer phenotypes in different data source using non-small cell lung cancer (NSCLC) as a use case. Non-small cell lung cancer (NSCLC) is the second most commonly diagnosed cancer worldwide and is the most common type of lung cancer in the US (1). Small cell lung cancer (SCLC) only represents 15% of all lung cancers (ref). Accurate identification of this condition requires information on tumor histology (adenocarcinoma, squamous cell carcinoma, and large cell carcinoma) in addition to the location of the tumor (lung).

Several attempts have been made to define NSCLC patient cohorts using a combination of lung cancer diagnosis and presence of NSCLC specific treatment or absence of SCLC specific treatment (2-4). The performance of these approaches, however, remains unclear. We therefore measured the performance of these definitions data sources that also provides the actual histological diagnosis and estimated the rate of misclassification.

Methods

Data from IQVIA-OncoEMR, a US-wide oncology database with available information on tumor histology, was used to assess the performance of the following definitions for NSCLC:

- C1: Lung cancer (LC)
- C2: LC with confirmed diagnosis of NSCLC
- C3: LC with confirmed diagnosis of SCLC
- C4: LC + treatment recommended for NSCLC
- C5: LC + treatment recommended for SCLC
- C6: LC + treatment recommended for NSCLC only
- C7: LC + treatment recommended for SCLC only
- C8: LC + treatment indicated for both conditions

NSCLC and SCLC specific treatments were identified through HemOnc (https://hemonc.org/wiki/Main_Page), the largest freely available medical wiki of interventions, regimens, and general information relevant to the fields of hematology and oncology. HemOnc compiles updated and detailed information on treatment recommendation for different cancer from various guidelines including ASCO, ESMO, KSMO and NCCN.

NSLC and SCLC diagnosis in OncoEHR were used as gold standard. In the next stage of this project, the

same approach will be used in a series of academic EHRs with linkage to tumor registry where tumor registry information will be used as the gold standard.

Results and Discussion

Any misclassification between NSCLC and SCLC has the tendency to significantly alter the composition of the patient population: Compared to NSCLC, patients with SCLC tend to have higher rates of metastasis at the time of diagnosis, specifically metastasis to liver and brain. Brian and lung metastasis have both been associate with poor prognosis, suboptimal response to treatment and overall survival. Additionally, there seems to be differences in the distribution of treatments administered after diagnosis (index date) (Figure 1).

The simplest way to create a cohort of NSCLC is to use LC (C1), assuming that the proportion of SCLCS is low and a positive predictive value of 80% is sufficient. In OncoEMR, the distribution of the two histological types follows the expected rates from the literature: from a total of 39,471 with 2 diagnosis of lung cancer, 15,428 (39%) patients had a confirmed NSCLC or SCLC; 13,003 (85%) were NSCLC (C2) and 2,347 (15%) were SCLC (C3). However, that may or may not hold for another database.

Only 20% of the lung cancer patients have a record of treatment with an antineoplastic agent; 19% of patients with NSCLC (C4) and 27% of patients with SCLC (C5). Most of the treated population in both cohorts (66% of NSCLC treated cohort and 86% of SCLC treated cohort), however, received a treatment that is not unique to either condition (C8), limiting the performance of these definitions to correctly identify NSCLC and SCLC patients. Using unique treatments (C6 and C7) would correctly identify 333 NSCLC (2%) and 5 SCLC (<0.01%) patients in this data.

Variable and significant misclassification was observed across phenotypes. The impact of such misclassification depends on how phenotypes are used. In the absence of data granularity, detailed evaluation of the performance of (rule-based and probabilistic) phenotypes and their potential impact on different oncology research questions should be comprehensively investigated.

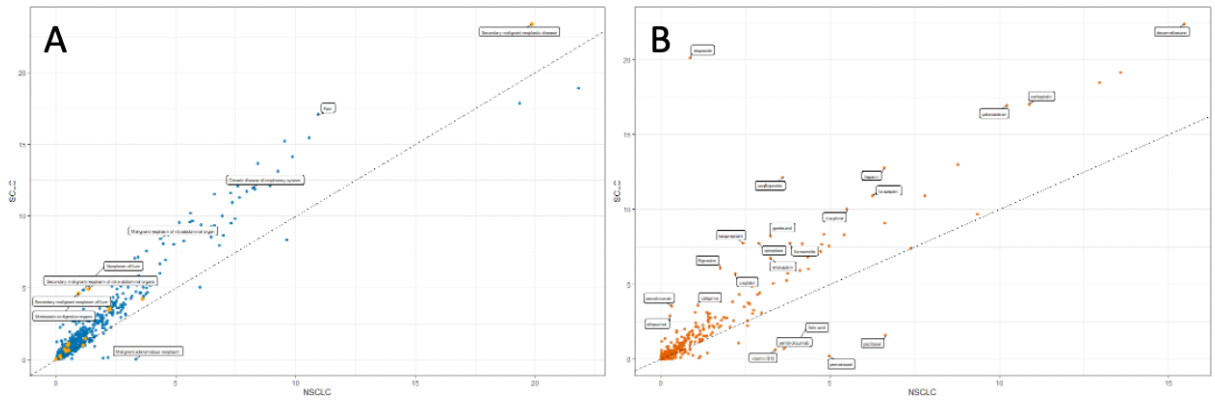

Figure 1. Frequency of pre-index comorbidities (A) and treatment patients received after index (B)

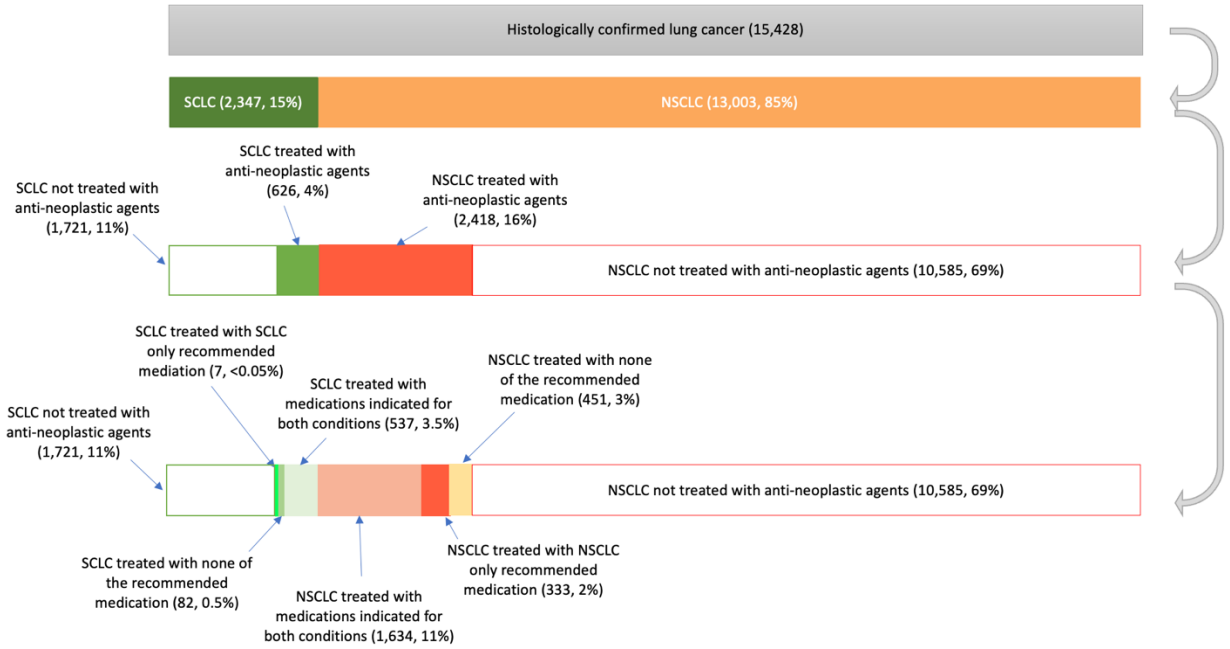

Figure 2. Lung cancer patient breakdown according to different definitions of NSCLC and SCLC

References

1. Gridelli C, Rossi A, Carbone DP, Guarize J, Karachaliou N, Mok T, et al. Non-small-cell lung cancer. Nat Rev Dis Primer. 2015 May 21;1(1):15009.

2. Duh MS, Reynolds Weiner J, Lefebvre P, Neary M, Skarin AT. Costs associated with intravenous chemotherapy administration in patients with small cell lung cancer: a retrospective claims database analysis. Curr Med Res Opin. 2008 Apr;24(4):967–74.

3. Turner RM, Chen YW, Fernandes AW. Validation of a Case-Finding Algorithm for Identifying Patients with Non-small Cell Lung Cancer (NSCLC) in Administrative Claims Databases. Front Pharmacol. 2017 Nov 30;8:883–883.

4. Balzi W, Roncadori A, Danesi V, Massa I, Manunta S, Gentili N, et al. How to discriminate non-small cell lung cancer (NSCLC) cases from an Italian administrative database? A retrospective, secondary data use study for evaluating a novel algorithm performance. BMJ Open. 2021 Sep 24;11(9):e048188–e048188.