

Why cancer phenotypes in RWE studies are often wrong?

Asieh Golozar^{1,2}, Henry Morgan Stewart¹, Jasmin Carus³, Sam T Patnoe⁴, Tommi Kauko⁵, Otto Ettala⁵, Zarah Van Schoor⁶, Annelies Verbiest⁶, Kimmo Porkka⁶, Patricia L Mabry⁴, Christian Reich^{1,2}

¹Nemesis Health, ²OHDSI Oncology WG, ³University Medical Center Hamburg-Eppendorf (UKE), Germany,

⁴HealthPartners Institute, USA, ⁵The wellbeing services county of Southwest Finland, Auria Clinical Informatics (VARHA), ⁶Antwerp University Hospital, University of Antwerp, Belgium, ⁶Helsinki University Hospital Comprehensive Cancer Center, Finland

Background

A cancer diagnosis is defined by a combination of attributes, including the anatomic site of origin, histology, stage, grade, and, increasingly, molecular markers. Together, these characteristics determine prognosis, guide treatment selection, and define the population relevant to a given research question. During the initial diagnostic workup, these attributes are established at the highest possible level of granularity to inform prognostic assessment and treatment decisions. Afterward, diagnostic evaluation typically becomes less intensive, apart from occasional re-staging as the disease progresses.

These features of cancer diagnosis and treatment have important implications for phenotyping in observational research. Phenotype algorithms define study populations by building a deterministic logic of inclusion and exclusion and the timing between them. This logic often reflects a pre-conceived clinical sequence: diagnosis is recorded first, followed by detection of metastases, completion of staging, and initiation of treatment. They also assume that malignancies are always coded at with sufficient specificity to distinguish multiple cancers over time. This conceptual sequence is commonly embedded in clinical trial protocols and, by extension, shapes expectations for real-world data.

In reality, the picture is much more complicated. Records serve administrative and billing functions in addition to clinical care. During diagnostic workup, coding typically evolves from broad to specific as diagnostic certainty increases, then often returns to broader codes once a treatment plan is established. This is aggravated by the fact that the main coding system for billing, ICD-10, is notoriously vague in histology and lacks all other cancer attributes except some metastases. Timing also rarely follows a strict primary cancer diagnosis-attribute detection-treatment sequence. In some cases, cancer is first recognized through a metastasis or related complication. Confirmatory pathology and radiology findings may be communicated to support rapid treatment initiation but entered into the record only after treatment has begun.

In the literature and in OHDSI studies, cancer phenotypes are often developed from what is called "clinical description" and therefore tend to encode the above idealized diagnostic sequence. To improve specificity, algorithms also borrow heuristics from pharmacoepidemiology, such as requiring repeated records within a prespecified time window. In oncology, these assumptions can produce substantial misclassification:

- As in clinical trials, some phenotypes exclude patients with multiple malignancies to attribute outcomes to the index cancer. However, longer survival and broader treatment use across tumor types have increased the prevalence of multiple malignancies in real-world populations.
- These exclusions are commonly implemented using diagnostic codes for any malignancy other than the cancer of interest, although nonspecific codes may represent the same underlying cancer rather than a distinct primary tumor.

- Because detection of prior or concurrent malignancies depends on available observation time, the impact of these exclusions varies across sites and can introduce selection bias.
- Phenotypes often assume a 60–90-day diagnostic workup before treatment initiation to capture biomarkers, staging, and eligibility assessments. In practice, this interval may be substantially shorter or longer depending on site-specific care processes and documentation patterns.
- Requiring repeated records within a fixed time window to increase specificity may be inappropriate in cancer, where coding granularity changes over time and event timing is often uncertain.

Here, we aim to determine empirically timing of events, using metastatic urothelial carcinoma (UC) in the FALCON-Bladder study on guideline adherence and relevance that is underway currently as the case study. We developed a diagnostic framework that looks at the cancer patient journey and describes important temporal details that is needed for phenotyping and does it at scale. Preliminary results from 6 partners are presented here and later extended to 40 partners with the full results presented at the Symposium.

Methods

We conducted this study within the federated FALCON-Bladder research network, in which queries are developed centrally, executed locally at participating sites, and again aggregated centrally. The query characterized the cancer patient journey and generated aggregate measures that inform phenotyping.

Patients with at least one diagnosis code of UC at any time in their data were included in the analysis. The index date was set to the first metastasis code. We then measured the temporal relationship to events in any direction before or after:

- The nearest UC diagnosis
- The nearest unspecific cancer diagnosis
- The nearest other cancer diagnosis
- The treatment initiation, defined as drug exposures at the ingredient level (ATC L01).

All metrics are reported at the aggregate level for each data partner. No patient-level data were shared across sites. Analyses were implemented using the open-source [onco-pre-study package](#).

Results

Results are reported from six databases across four countries: Antwerp University Hospital (UZA, Belgium, N=1,016), University Medical Center Hamburg-Eppendorf (UKE, Germany, N=2,245), VARHA Hospital District of Southwest Finland (VARHA, Finland, N=3,870), Helsinki University Hospital (HUS, Finland, N=8,833), HealthPartners Institute (HPI, USA, N=5,915), and a US commercial claims database (PMTX, N=58,540). From a total of 80,419 patients identified with a UC diagnosis, 12,655 (15.7%) had a recorded metastasis code during their follow up.

The proportion of patients with a cancer diagnosis code within ± 30 days of the first metastasis code ranged from 45.3% at UKE to 69.2% at PMTX. Most of this capture came from diagnosis codes recorded after the metastasis date, ranging from 35.0% at UKE to 65.6% at UZA, while codes recorded before the metastasis date contributed much less, ranging from 1.0% at UZA to 16.5% at VARHA.

Extending the window to ± 90 days increased capture modestly at most sites, from 51.4% at UKE to 77.1% at PMTX. This suggests that most proximate cancer diagnosis codes are recorded within 30 days of the metastasis date, and that additional lookback or follow-up time adds relatively little for most sites.

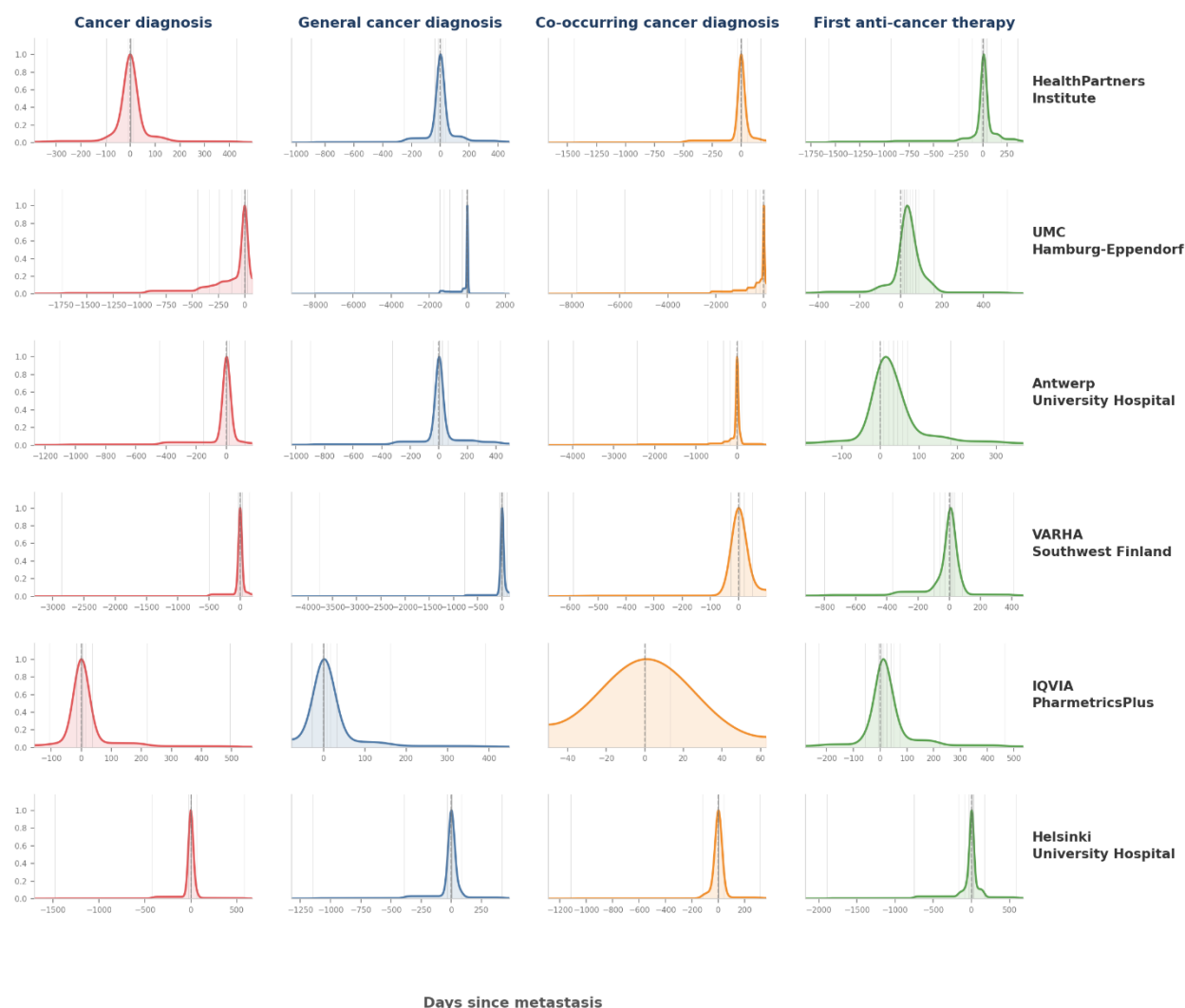


Figure 1. Cancer diagnosis codes and treatment initiation relative to the time of metastasis detection.

Recommendations

Our results show that assumption about events in complex disease like cancer does not follow the intuitive "ideal" sequence, and coding specificity varies. It also shows that timing is highly dependent on the site. In order to develop reliable and robust phenotypes, we recommend:

1. Do not exclude other malignancies. But if that is required, ensure that generic cancer codes are excluded from the exclusion.
2. Measure the distribution of events over time and then adjust the phenotype algorithm accordingly.