Defining international approaches for the detection of emergent metastasis and the classification of site of metastasis from hospital EHR

Stelios Theophanous[1], Sue Cheeseman[1], Elin Hallan Naderi[2], Elisabeth Ross[2], Anne-Lore Bynens[3], Prabash Galgane Banduge[4,5], Petros Kalendralis[4], Aiara Lobo Gomes[4], Piers Mahon[6]

[1]Leeds Teaching Hospitals NHS Trust, Leeds, UK
[2]Oslo University Hospital, Oslo, Norway
[3]Maastricht University Medical Centre+, Maastricht, The Netherlands
[4]Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre, Maastricht, The Netherlands
[5]DataHub, Maastricht University, Maastricht, The Netherlands
[6]DIGICORE, Brussels

Background

Metastasis, the spread of cancer from its primary site to distant healthy organs or tissues, represents a critical clinical event impacting treatment strategies and patient outcomes[1]. Metastases can be detected either at the time of initial diagnosis or emerge subsequently following relapse from localised disease[2]. The identification of sites of metastasis is important for guiding treatment strategies and formulating prognosis, and therefore crucial in real-world evidence research. For example, survival rates differ significantly for patients with progressed breast cancer, depending on whether metastasis is in bone or brain[3,4]. Despite its clinical importance, the recording of metastatic sites is often inconsistent and incomplete within EHR systems due to incomplete biopsy at recurrence, incomplete reclassification of TNM codes and poorly coded information in clinical notes. This hampers analysis of disease progression patterns and the calculation of clinically meaningful end points.

A multicentre, European DigiONE study[5], investigating treatment and outcomes in metastatic non-small cell lung cancer (mNSCLC), required the identification of sites of metastases at diagnosis and at relapse to define clinical groups of importance for statistical analysis. Here, we describe how we leveraged the OMOP databases across the study sites, to identify sites of metastases and devise a common classification system for these for mNSCLC within the OMOP framework.

Methods and Results

Detection of emergent metastases

Various solutions have been implemented to detect the presence and sites of metastases in the three participating hospitals, each with a different electronic health record (EHR) structure (Table 1).

Table 1. Summary of approaches used to detect emergent metastasis in the three participating hospitals.

	Oslo University Hospital	**Leeds Teaching Hospitals NHS Trust**	**Maastricht University Medical Centre+**
Changes to underlying ICD-10 codes	Yes	No	No
Formal M=1 on staging at re-biopsy	No	No	No
Pathology reports after 1st presentation with location of metastasis coded	Presence but no location	Yes	Partially
Imaging notes or reports with mention of likely (new) metastasis	No	Yes	Yes
Clinical letters	No	Yes	Yes

At Oslo University Hospital (OUH), only non-curated source data have been included in the local OMOP database. A metastasis event was captured in three ways: (1) a hospital cancer episode in the EHR system where a metastasis diagnosis was recorded in the form of an ICD-10 code, (2) a recording of TNM staging with M1, or (3) a pathology result that specified that the malignant histology originated from a metastatic site. The presence of a metastasis event following a TNM staging with M0 signified a *definite* emergent metastasis event. A metastasis event following a primary cancer diagnosis without a concurrent metastasis diagnosis was *assumed* to be a case of an emergent metastasis. However, although the ICD-10 coding system of primary cancer is very comprehensive, the recording of metastasis in ICD-10 has a higher degree of missingness.

The Leeds Teaching Hospitals NHS Trust (LTHT) uses an in-house EHR system that allows the recording of staging data and associated sites of metastases for each site-specific cancer diagnosis and recurrence event. However, data completion varies across indication-specific cancer teams. For the NSCLC study cohort, the staging data was manually curated from source data by clinical review of all relevant imaging reports, pathology reports and clinical letters. Metastases and dates of identified relapse were updated in available fields in the structured EHR before translation to OMOP. In some cases, surrogates were also used to identify relapse events, such as new anti-cancer treatments or procedures starting after a specified time window from first ever diagnosis, and disease codes for specific cancer-related problems, such as spinal cord compression.

At Maastricht University Medical Centre+ (MUMC+), the identification of metastasis events, including site(s) of metastasis, involved a combination of natural language processing (NLP) and subsequent manual validation by an oncology nurse. Clinical notes that describe disease management and evolution (*decursus*) are automatically extracted from the EHR system called SAP (System Analysis Program development) into health analytics software (CTCue[6]), which provides in-built NLP. The occurrence of the word “metastasis” (in English or Dutch), M1 stage, or overall stage IV was assessed by the NLP software, including pre-specified synonyms for these three concepts. The output included the string where “metastasis”, M1, stage IV, or any of their synonyms were mentioned, as well as the site(s) and date(s) of presentation of the metastases, allowing review before update of structured

fields. It was possible to distinguish between metastasis at diagnosis and emergent metastasis after disease relapse, using report and diagnosis dates.

Classification of metastasis location data

Metastasis location and classification for metastatic mNSCLC were harmonised between the three participating hospitals (Table 2). Using the expertise of medical oncologists, metastatic locations were categorised into six groups: brain, liver, adrenal glands, bone, lung, and "other" anatomical structures. Variations identified in coding systems across centres were (1) grouped brain and leptomeningeal metastases at OUH, but separately reported at LTHT, (2) grouped bone and bone marrow metastases at OUH, but separately reported at LTHT, and (3) adrenal metastases were specified at OUH but classed as "other" at LTHT. Once these differences were characterised, a common coding system was agreed to harmonise data across LTHT and OUH. MUMC+ subsequently altered the rules of the NLP software to capture the metastasis location data with the highest granularity (e.g. differentiating between brain and leptomeningeal metastases).

Table 2. Final OMOP metastatic site concept coding for the mNSCLC cohort in the three participating hospitals.

		Oslo University Hospital	**Leeds Teaching Hospitals NHS Trust**	**Maastricht University Medical Centre+**
Local coding system for metastasis location		ICD-10	EHR-specific drop-down	Structured output from NLP software, validated by a human
OMOP concept ID used to denote site of metastasis	Brain	35225775	36768862	36768862
	Leptomeninges		4235348	35226096
	Lung	254591, 36770283	36770283	36770283
	Pleura	35226258, 72266	35226258	35226258
	Bone	36769301, 35226074	36769301	36769301
	Bone Marrow		35226074	35226074
	Liver	36770544	36770544	36770544
	Adrenal glands	193144 and 35225568	36769180	35225568
	Other anatomical structures	Lymph nodes: 434298, 35225542, 434875, 318096, 442182, 192568, 200959, 439751, 320342, 318096; Metastasis	Lymph nodes: 4110086, 4215878, 35225550, 36768587, 4057702, 35226326, 4081801, 4075974; Omentum: 35226218;	36769180

		Gastrointestinal and retroperitoneal: 35225543, 35226222, 704985, 198371, 35225719; Genitourinary: 35225580, 78987, 35225580, 199752, 35226230; Other and unspecified parts of nervous system: 35225775, 373425; Other and unspecified respiratory organs: 253717, 35226280; Other specified sites: 36769180, 432851 Skin: 136354, 35225673; Unspecified site: 36769180, 4158910	Peritoneum: 35226253; Skin: 35225673; Soft tissue: 35226117	

Conclusion

The identification of cancer relapse and recurrence events, as well as sites of metastases, is important for the definition of clinically important end points and to improve clinical care. Much of this information is not routinely recorded in EHR systems, and the quality of completeness and detail varies across platforms. The three hospitals in this study employed various methods to identify this important information, all of them human resource and time intensive and with some limitations. LTHT approaches will likely miss some patients who relapse but do not receive further treatment and OUH may miss-classify patients with emergent metastasis following a primary diagnosis outside the hospital.

The use of more novel techniques, such as NLP and machine learning applied to a wider range of hospital data systems (e.g. pathology and imaging), would increase efficiency. It would be of interest to compare the accuracy of these newer techniques to the manual curation processes described.

References

1 Fares J, Fares MY, Khachfe HH, Salhab HA, Fares Y. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct Target Ther* 2020; **5**: 28.

2 Riggio AI, Varley KE, Welm AL. The lingering mysteries of metastatic recurrence in breast cancer. *Br J Cancer*2021; **124**: 13–26.

3 Li B, Wong M, Pavlakis N. Treatment and Prevention of Bone Metastases from Breast Cancer: A Comprehensive Review of Evidence for Clinical Practice. *J Clin Med* 2014; **3**: 1–24.

4 Bailleux C, Eberst L, Bachelot T. Treatment strategies for breast cancer brain metastases. *Br J Cancer* 2021; **124**: 142–55.

5 Mahon P, Chatzitheofilou I, Dekker A, *et al.* A federated learning system for precision oncology in Europe: DigiONE. *Nat Med* 2024; **30**: 334–7.

6 CTcue. An IQVIA Business: Empowering healthcare with real-world evidence 2024. https://ctcue.com (accessed March 14, 2024).