ICD-coded Information on Sites of Metastasis in Oncology Real-World Data is Specific but not Sensitive

Christina M. Parrinello, PhD, MPH; Katharina N. Seidl-Rathkopf, PhD; Caroline Bennette, PhD, MPH; Ariel B. Bourla, MD, PhD; Nathan C. Nussbaum, MD; Kenneth R. Carson, MD, PhD; Amy P. Abernethy, MD, PhD

Flatiron Health, New York, NY

Background

- Real-world data (RWD) sources include claims data, electronic health records (EHRs), and registries. The reliability of these data sources for important research questions is uncertain.
- In cancer, extent of disease, including site(s) of metastasis is critical for characterizing the population.
- This study used RWD from the Flatiron Health EHR-derived Database, a specialty Oncology EHR, to evaluate the accuracy of ICD-coded sites of metastasis in metastatic breast cancer (mBC) and advanced melanoma (aMel) as compared to abstracted data (considered the gold standard for this study).

Methods

Cohort Selection

- This study included patients from the Flatiron Health EHR-derived Database, with data recency through 11/30/2017.
- The Flatiron Health database is a longitudinal, demographically and geographically diverse database derived from EHR data. It includes data from > 265 cancer clinics and > 2 million active cancer patients available for analysis.
- Patient-level data include structured and unstructured data. Structured data are pulled directly from the EHR. To extract data from unstructured data sources (such as clinic notes), Flatiron has developed a "technology-enabled" chart abstraction methodology.

Statistical Analysis

- Calculated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of site of metastasis as documented by ICD codes compared to site as abstracted from clinical notes and pathology/radiology reports ("gold standard")
- Analyses conducted separately for mBC and aMel patients, and for each site of metastasis

Results

Metastatic Breast Cancer (mBC)

- 14,699 mBC patients were included in this analysis
- Sensitivity was highest for bone (67%, 95% CI: 66-68%) and lowest for distant lymph node (9%, 95% CI: 8-10%)
- Specificity was high across all sites, ranging from 94-100%

Table 1: Accuracy of ICD-coded sites of metastasis as compared to abstraction, N = 14,699 mBC patients

					Site of metastasis identified via:	
	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	ICD code (N)	Abstraction (N)
Any Site	0.66 (0.65, 0.66)	0.75 (0.66, 0.83)	1.00 (0.99, 1.00)	0.02 (0.01, 0.02)	9,593	14,597
Bone	0.67 (0.66, 0.68)	0.95 (0.94, 0.96)	0.97 (0.96, 0.97)	0.55 (0.54, 0.56)	7,138	10,307
Lung	0.25 (0.24, 0.26)	0.98 (0.98, 0.99)	0.91 (0.89, 0.92)	0.67 (0.66, 0.68)	1,588	5,770
Liver	0.31 (0.30, 0.33)	0.99 (0.99, 0.99)	0.94 (0.93, 0.95)	0.74 (0.73, 0.75)	1,634	4,905
Brain	0.40 (0.38, 0.42)	0.99 (0.99, 0.99)	0.86 (0.84, 0.88)	0.90 (0.89, 0.90)	1,056	2,283
Other CNS Site	0.13 (0.11, 0.16)	1.0 (1.0, 1.0)	0.64 (0.55, 0.72)	0.96 (0.96, 0.96)	133	640
Distant Lymph Node	0.09 (0.08, 0.10)	0.98 (0.98, 0.99)	0.70 (0.67, 0.74)	0.70 (0.70, 0.71)	594	4,597
Other Site	0.20 (0.18, 0.21)	0.94 (0.93, 0.94)	0.48 (0.45, 0.50)	0.81 (0.80, 0.81)	1,346	3,233

Advanced Melanoma (aMel)

- 6,429 aMel patients were included in this analysis
- Sensitivity was highest for brain (50%, 95% CI: 48-53%) and lowest for soft tissue and distant skin/subcutaneous (0%, 95% CI: 0-1% and 9%, 95% CI: 7-12%, respectively)
- Specificity was high across all sites, ranging from 92-100%

Table 2: Accuracy of ICD-coded sites of metastasis as compared to abstraction, N = 6,429 aMel patients

					Site of metastasis identified via:	
	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	ICD code (N)	Abstraction (N)
Any Site	0.50 (0.48, 0.51)	1.0 (0.99, 1.0)	0.99 (0.99, 1.0)	0.55 (0.53, 0.56)	2,009	4,007
Bone	0.43 (0.40, 0.46)	0.98 (0.98, 0.98)	0.83 (0.80, 0.86)	0.89 (0.88, 0.89)	595	1,159
Lung	0.30 (0.28, 0.32)	0.98 (0.98, 0.99)	0.92 (0.90, 0.94)	0.71 (0.70, 0.72)	779	2,369
Liver	0.31 (0.28, 0.33)	0.99 (0.99, 0.99)	0.86 (0.83, 0.89)	0.86 (0.85, 0.87)	428	1,209
Brain	0.50 (0.48, 0.53)	0.99 (0.99, 0.99)	0.94 (0.92, 0.96)	0.86 (0.85, 0.87)	830	1,553
Distant Lymph Node	0.14 (0.12, 0.16)	0.95 (0.95, 0.96)	0.44 (0.39, 0.49)	0.82 (0.81, 0.83)	420	1,293
Distant Skin/ Subcutaneous	0.09 (0.07, 0.12)	0.98 (0.98, 0.98)	0.41 (0.34, 0.49)	0.89 (0.88, 0.89)	177	787
Soft Tissue	0 (0, 0.01)	1.0 (1.0, 1.0)	NA	0.87 (0.86, 0.88)	0	854
Other Site	0.29 (0.27, 0.32)	0.92 (0.91, 0.93)	0.49 (0.46, 0.53)	0.83 (0.82, 0.84)	809	1,364

Limitations/Discussion

- ICD codes were grouped into categories of metastasis site. For instance, "soft tissue" included multiple ICD codes, such as "Kaposi's sarcoma of soft tissue" and "Neoplasm of unspecified behavior of bone, soft tissue, and skin". Method of categorization may have resulted in some measurement error. Next steps may include re-assessing ICD code categorization and additional sensitivity analyses. Furthermore, researchers could consider collapsing certain categories (e.g., combining "Other CNS" and "Other") to minimize potential measurement error.
- Cohort selection included confirmation of advanced/metastatic disease via abstraction, whereas claims-based analyses typically rely on ICD codes for cohort selection. This may have contributed to the high specificity we observed. Additional analysis of a cohort within the Flatiron network selected via ICD codes rather than abstraction could provide insight into the effect of cohort selection method on our results.
- Although patients were initially identified using ICD codes, the majority of the study population was drawn from community oncology clinics, and was therefore a somewhat select group of patients. A sensitivity analysis could restrict to academic sites, which would potentially include a broader patient base.
- Accuracy of metastasis to distant lymph nodes and distant skin/subcutaneous sites was quite poor, and may indicate different coding practices for these types of metastases.
- The current analysis defined "agreement" as presence of metastasis as indicated by both ICD codes and abstraction, regardless of whether the dates of metastasis from both approaches matched. Sensitivity analyses that required ICD codes and abstraction indicating metastasis to be within 90 days of each other resulted in lower sensitivity.
- Although abstraction was used as the "gold standard" here, completeness of sites of metastasis in the EHR could vary by disease, site, or provider.
- Variables that have much higher specificity than sensitivity may be better suited to address certain types of research questions. For instance, ICD codes could be used to identify a group of patients with melanoma who have brain mets to determine their survival after administration of a certain therapy. However, using ICD codes to determine the incidence rate of brain mets in melanoma patients administered a certain therapy would result in underestimation.

Conclusions

- For identification of sites of metastasis, ICD codes from an Oncology EHR were highly specific, yet not very sensitive. In oncology practice, where ICD codes are entered mainly for billing reasons, sites of metastasis are often not captured by ICD codes even though they are commonly captured in routine clinical documentation.
- This study calls into question the reliance on ICD codes alone for research questions where characterization of sites of metastasis is important.
- This study highlights the importance of evaluating accuracy of each variable in the data source, especially as RWD becomes more important for high impact clinical and regulatory decisions.
- It is important to continue to abstract this information from the EHR to maximize capture of sites of metastasis to ensure research-quality data.

Presented at ISPOR: May 19-23, 2018 in Baltimore, MD.
For additional information, contact Christina Parrinello at cparrinello@flatiron.com.

All authors are full-time employees of Flatiron Health, a cancer-focused health technology company in New York, New York, USA and the source of study data.