Comparison of Structured Versus Abstracted Comorbidities Using Oncology EHR Data from Cancer Patients in the Flatiron Health Network

Christina M. Parrinello, PhD, MPH; Katharina N. Seidl-Rathkopf, PhD; 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

- The Charlson Comorbidity Index (CCI) is used to characterize risk in populations, and is typically calculated using ICD code algorithms from real-world data (RWD) sources, mainly claims data.
- Electronic health record (EHR) data are becoming a more common source of RWD for research, and are themselves a source of ICD codes.
- Using data captured in a specialty Oncology EHR, this study compared the accuracy of the CCI calculated using ICD codes versus data abstracted from the medical record. The abstracted data were considered to be the gold standard.

Methods

Study Population

- This study included advanced non-small cell lung cancer (NSCLC) patients from the Flatiron Health Database (**Figure 1**).
- 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.

Figure 1: Cohort selection

ICD code for lung cancer; 2 visits on or after 1/1/2011; advanced dx date on or after 1/1/2011 as confirmed via abstraction
n = 24,191

→ Random sample of patients with *structured* evidence of first line platinum doublet chemo and pemetrexed maintenance
n = 1,162

→ Diagnosis with advanced nonsquamous NSCLC between 1/1/2011 and 10/31/2015
n = 1,087

→ Confirmation of treatment regimen via abstraction and no concurrent cancer within 3 years of advanced dx
n = 715

Table 2: Calculation of diagnostic accuracy of ICD codes for CCI as compared to the gold standard of abstraction

		Abstraction		
		CCI ≥ 1	CCI = 0	
ICD Codes	CCI ≥ 1	True Positive (A)	False Positive (B)	**PPV:** A/(A+B)
	CCI = 0	False Negative (C)	True Negative (D)	**NPV:** D/(C+D)
		Sensitivity: A/(A+C)	**Specificity:** D/(B+D)	

[1]Quan H, et al. Updating and Validating the Charlson Comorbidity Index and Score for Risk Adjustment in Hospital Discharge Abstracts Using Data From 6 Countries. *Am J Epidemiol* 2011; 173:676–682.
*Cancer and metastasis were *not* included in the CCI calculation for this study.

Statistical Analysis

- We identified comorbidities used in the calculation of the modified CCI, not including cancer or metastasis, using structured data (ICD 9/10 codes) and unstructured data (from an Oncology EHR) (**Table 1**).
- We assessed sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of presence of any comorbidity (CCI ≥ 1) as well as each individual comorbidity, *considering the abstracted data to be the gold standard* for purposes of calculation (**Table 2**).
- We calculated the percent of patients whose CCI changed when determined using ICD codes versus abstraction.
- Comorbidities were only considered if they occurred on or before the advanced diagnosis date.

Table 1: Modified CCI scoring[1]

Comorbidity	Score
Congestive heart failure	2
Dementia	2
Chronic pulmonary disease	1
Rheumatologic disease	1
Mild liver disease	2
Moderate or severe liver disease	4
Diabetes with chronic complications	1
Hemiplegia or paraplegia	2
Renal disease	1
AIDS/HIV	4
Any malignancy*	2
Metastatic solid tumor*	6

Results

- In 715 patients, capture of CCI by ICD codes was poor. The sensitivity of CCI ≥ 1 as measured by ICD codes as compared to abstracted data (gold standard) was 9% (95% CI: 5-14%) (**Table 3**).
- Using ICD codes to identify patients *without* any comorbidities works well most of the time. This is reflected in the high observed specificity and NPV (99%, 98-100% and 72%, 69-75%, respectively) (**Table 3**).
- A similar pattern was observed across individual comorbidities: sensitivity < 30%, specificity 99-100%, and NPV 75-100%. Estimates of PPV had poor precision because of the low number of comorbidities identified via ICD codes (**Table 3**).
- Abstraction captured more comorbidities than ICD codes. Among patients whose CCI score differed depending on whether it was calculated using ICD codes or abstracted data, most had a *higher* CCI score when using abstracted data. Reclassification to a higher CCI was almost entirely driven by patients being classified as having *no* comorbidities from ICD codes (CCI = 0) but at least 1 comorbidity based on abstracted data (CCI ≥ 1) (**Table 4**).

Table 3: Comparison of structured and unstructured data for identification of comorbidities in 715 advanced NSCLC patients

	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Any comorbidity (CCI ≥ 1)	0.09 (0.05, 0.14)	0.99 (0.98, 1.00)	0.86 (0.65, 0.97)	0.72 (0.69, 0.75)
Congestive heart failure	0.27 (0.08, 0.55)	1.00 (0.99, 1.00)	1.00 (0.28, 1.00)	0.98 (0.97, 0.99)
Dementia	0.00 (0.00, 0.81)	1.00 (0.99, 1.00)	NA*	1.00 (0.99, 1.00)
Chronic pulmonary disease	0.06 (0.03, 0.11)	0.99 (0.98, 1.00)	0.79 (0.49, 0.95)	0.75 (0.72, 0.79)
Rheumatologic disease	0.10 (0.01, 0.32)	1.00 (0.99, 1.00)	1.00 (0.09, 1.00)	0.97 (0.96, 0.98)
Mild liver disease	0.00 (0.00, 0.81)	1.00 (0.99, 1.00)	0.00 (0.00, 0.99)	1.00 (0.99, 1.00)
Moderate or severe liver disease	0.00 (0.00, 0.99)	1.00 (0.99, 1.00)	NA*	1.00 (0.99, 1.00)
Diabetes with chronic complications	0.00 (0.00, 0.53)	1.00 (0.99, 1.00)	0.00 (0.00, 0.99)	0.99 (0.98, 1.00)
Hemiplegia or paraplegia	0.00 (0.00, 0.91)	1.00 (0.99, 1.00)	NA*	1.00 (0.99, 1.00)
Renal disease	0.09 (0.00, 0.41)	1.00 (0.99, 1.00)	1.00 (0.01, 1.00)	0.99 (0.97, 0.99)
AIDS/HIV	0.00 (0.00, 0.91)	1.00 (0.99, 1.00)	NA*	1.00 (0.99, 1.00)

*Incalculable because no comorbidities were identified via ICD codes

Table 4: CCI reclassification in 715 advanced NSCLC patients as measured by structured versus unstructured data

		CCI from Unstructured (Abstracted) Data					Reclassification of patients from ICD code-based CCI to abstraction-based CCI		
		0	1	2	3+		Lower CCI	Higher CCI	Total Patients Reclassified
CCI from Structured (ICD code) Data	0	500 (69.9%)	156 (21.8%)	21 (2.9%)	16 (2.2%)		--	193 (28%)	193 (28%)
	1	3 (0.4%)	12 (1.7%)	1 (0.1%)	0 (0%)		3 (19%)	1 (6%)	4 (25%)
	2	0 (0%)	2 (0.3%)	0 (0.0%)	2 (0.3%)		2 (50%)	2 (50%)	4 (100%)
	3+	0 (0%)	0 (0%)	1 (0.1%)	1 (0.1%)		1 (50%)	--	1 (50%)

Limitations

- Relatively few comorbidities were identified overall, whether by ICD code or abstraction, which resulted in low precision of our results.
- The Flatiron Health Database is derived from an Oncology EHR, and data from medical records outside of the oncology care a patient receives (i.e., hospital admissions, primary care physicians, claims data) may be incomplete. Oncologists may not have an incentive to enter ICD codes for comorbidities because it is unlikely to lead to higher payment. Furthermore, they are unlikely to describe comorbidities in the medical record unless it will influence treatment decisions. We are therefore likely underestimating the presence of comorbidities from both structured and unstructured data.

Conclusions/Discussion

- Using data from an Oncology EHR, ICD codes were not sufficient for identification of comorbidities as compared to abstracted information from clinician notes, and may not capture CCI-relevant comorbidities that are documented elsewhere in the EHR.
- Comorbidity data abstracted from an Oncology EHR may itself be incomplete and not an ideal gold standard. Oncologists are unlikely to document a comorbidity unless it will affect cancer treatment decision-making
- Next steps include combining Oncology EHR data with additional data sources to increase the completeness of data for use in CCI calculation.
- Additional analysis may include stratification by stage of diagnosis (diagnosed with early stage vs metastatic disease). It is possible that compared to patients who are diagnosed with metastatic disease, those who are diagnosed with early stage disease may have more complete capture of comorbidities in either structured or unstructured data because there are more clinical encounters and opportunities for documentation.

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

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