

# Workflow Differences Affect Data Accuracy in Oncologic EHRs: A First Step Toward Detangling the Diagnosis Data Babel

Franck Diaz-Garelli, PhD<sup>1,2</sup>; Roy Strowd, MD<sup>2</sup>; Virginia L. Lawson, MS<sup>1,2</sup>; Maria E. Mayorga, PhD<sup>3</sup>; Brian J. Wells, MD, PhD<sup>2</sup>; Thomas W. Lycan Jr, DO<sup>2</sup>; and Umit Topaloglu, PhD<sup>2</sup>

**PURPOSE** Diagnosis (DX) information is key to clinical data reuse, yet accessible structured DX data often lack accuracy. Previous research hints at workflow differences in cancer DX entry, but their link to clinical data quality is unclear. We hypothesized that there is a statistically significant relationship between workflow-describing variables and DX data quality.

**METHODS** We extracted DX data from encounter and order tables within our electronic health records (EHRs) for a cohort of patients with confirmed brain neoplasms. We built and optimized logistic regressions to predict the odds of fully accurate (ie, correct neoplasm type and anatomic site), inaccurate, and suboptimal (ie, vague) DX entry across clinical workflows. We selected our variables based on correlation strength of each outcome variable.

**RESULTS** Both workflow and personnel variables were predictive of DX data quality. For example, a DX entered in departments other than oncology had up to 2.89 times higher odds of being accurate ( $P < .0001$ ) compared with an oncology department; an outpatient care location had up to 98% fewer odds of being inaccurate ( $P < .0001$ ), but had 458 times higher odds of being suboptimal ( $P < .0001$ ) compared with main campus, including the cancer center; and a DX recoded by a physician assistant had 85% fewer odds of being suboptimal ( $P = .005$ ) compared with those entered by physicians.

**CONCLUSION** These results suggest that differences across clinical workflows and the clinical personnel producing EHR data affect clinical data quality. They also suggest that the need for specific structured DX data recording varies across clinical workflows and may be dependent on clinical information needs. Clinicians and researchers reusing oncologic data should consider such heterogeneity when conducting secondary analyses of EHR data.

JCO Clin Cancer Inform 4:529-538. © 2020 by American Society of Clinical Oncology

## INTRODUCTION

Secondary analysis of electronic health record (EHR) data is essential to the development of learning health care systems,<sup>1-3</sup> oncologic comparative effectiveness research,<sup>4,7</sup> and precision oncology decision support.<sup>8-11</sup> This often relies on patient diagnosis (DX) data for cohort selection<sup>12,13</sup> in spite of its quality limitations.<sup>14</sup> Decades of research have shown alarmingly high DX inaccuracy rates,<sup>15-17</sup> which can greatly affect secondary analysis results.<sup>15</sup> Despite error rate improvements from 20%-70% in the 1970s to 20% in the 1980s, their reliability remains questioned.<sup>5,18</sup> The complex nature of clinical knowledge, variable clinical workflows, and billing-oriented data recording are partially to blame<sup>5,19-22</sup> but the challenges are exacerbated by EHR systems that provide multiple descriptions for individual DX codes.<sup>23</sup>

Inaccurate DX data entry is particularly complex in oncology because of both cancer DX coding structures

and oncology workflow constraints. Standard DX code descriptions are not designed to support oncology data reuse,<sup>5,24,25</sup> leaving information locked in progress notes.<sup>5,26</sup> For example, International Classification of Diseases (10th revision) codes C71.XX correspond to “malignant neoplasm of the brain” DX codes and allow encoding anatomic site (eg, C71.1 represents a malignant neoplasm of the frontal lobe). However, unlike International Classification of Diseases for Oncology (3rd edition) codes,<sup>27</sup> which are not as broadly adopted, they do not encode neoplasm type (eg, *IDH* wild-type glioma, glioblastoma, and so on), which is crucial to treatment selection and patient classification. Customized DX descriptions provided by EHR vendors include some neoplasm type information but present wide-varying levels of specificity, complicating structured DX entry.<sup>28</sup> Because cancer care is team based and requires patients to interact with multiple specialties and units (eg, scheduling, imaging, surgery, and so on); EHR systems rarely support

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on March 18, 2020 and published at [ascopubs.org/journal/cci](https://ascopubs.org/journal/cci) on June 16, 2020; DOI <https://doi.org/10.1200/CCI.19.00114>

## CONTEXT

### Key Objective

To uncover the relationship between oncologic electronic health record (EHR) diagnosis (DX) data quality and both workflow and user profiles to inform reliable reuse of oncologic EHR data in clinical care for precision oncology and clinical decision support.

### Knowledge Generated

Both clinical department and care location were predictive of accurate or suboptimal DX entry (eg, DX entered in nononcology departments had higher odds of being accurate; outpatient care locations had lower odds of recording inaccurate DX). A user's clinical role was predictive of accurate, inaccurate, or suboptimal DX data entry (eg, DX recoded by a physician assistant had fewer odds of being suboptimal compared with those entered by physicians).

### Relevance

Clinical oncologists using EHR data to make patient care decisions or reusing clinical data to develop clinical decision support tools should account for these significant accuracy differences across clinical workflows and clinical personnel. There may be different levels of ideal DX specificity across clinical workflows.

consistent data recording across clinical workflows, making DX logging burdensome to oncologists.<sup>29-31</sup> Although both aspects cause structured DX data unreliability, much more has been published on resulting data issues<sup>14,21,22</sup> and limitations of coding terminology<sup>25,21</sup> than on the impact of clinical workflows on EHR data.<sup>32-35</sup>

Oncologic records suffer from high DX data variability,<sup>28</sup> EHR section-dependent accuracy levels,<sup>36</sup> and statistically significant differences in DX entry accuracy across subspecialties.<sup>37</sup> This hints at a link between clinical workflows and resulting data quality, in concordance with prior work.<sup>5,38</sup> However, the literature is still unclear on which factors affect data quality. To address this gap, we conducted a statistical analysis. We assessed which clinical workflow factors correlate with accurate DX entry. We hypothesized that there is a statistically significant relationship between workflow-describing variables (eg, care location, department, and users) and accurate, inaccurate, and suboptimal (ie, correct but imprecise) data entry. We tested this hypothesis on EHR data from patients diagnosed with brain neoplasms. We selected this disease for its large number of textual diagnosis descriptions of varying levels of precision for a limited list of specific diagnosis codes and the availability of a definitive histopathology report stating the most precise and accurate DX description possible. This analysis improves our understanding of oncologic data entry within clinical workflows and its impact on clinical data quality. Our findings identify a new avenue for clinical data quality improvement, thereby facilitating reliable secondary uses of oncology data within learning health care systems and future clinical oncology applications, such as clinical data-driven clinical decision support for cancer treatment selection.

## METHODS

We extracted structured DX data and relevant covariates across multiple clinical workflows from the Wake Forest

Baptist Medical Center's EHR database. Our study was approved by Wake Forest University School of Medicine's Institutional Review Board (IRB; No. 00044728). Our initial extract contained oncologic DX entries for a set of 36 patients treated for brain neoplasms. The data set contained DX descriptions entered during care and covariates describing clinical workflows and involved personnel. The covariates selection was driven by relevance to clinical workflow and personnel-EHR interaction descriptors, but also data availability within the EHR's data and metadata tables. Our initial list of covariates included days between biopsy and DX entry, DX chronologic rank in entry sequence, clinical department and location where the DX was entered, visit providers and their specialties, users entering the data and their clinical role (eg, physician, nurse, assistant, and so on), the authorizing provider, and the order type for order DX.

To identify accurate, inaccurate, and suboptimal DX description, we used a clinician-generated gold standard containing each patient's accurate diagnosis. Patient charts were preselected from an existing chart review-based glioma registry (IRB No.: 00038719). Patient inclusion was based on completeness of information within the registry. Patients having received cancer care outside our institution were excluded to ensure that all relevant care information was available within our EHR. Comprehensive medical record review was performed by 2 independent reviewers. The primary postoperative diagnosis was determined based on a review of pathology reports and clinician notes. All treating clinicians were available for consultation when needed. Discrepancies between the 2 reviewers were resolved by an independent neuro-oncologist. Two features were used to determine DX accuracy: neoplasm type (eg, astrocytoma, glioblastoma, and so on) and anatomic site (eg, frontal lobe, temporal lobe, and so on). We compared the standard's DX description

with each DX entry. Accurate DX entries matched the standard's neoplasm type and anatomic site; inaccurate DX entries contradicted neoplasm type or anatomic site. For example, a frontal lesion DX for a patient with a temporal lesion would be inaccurate; an astrocytoma DX for a patient treated for a glioblastoma would also be inaccurate. Partially accurate (ie, suboptimal) DX descriptions were categorized separately because they did not contradict the standard DX because they failed to provide a specific DX description.

Our final data set contained 10,052 DX attached to procedure orders and 3,718 encounter DX observations of 31 patients, recorded from January 1, 2016, to June 1, 2018. This time frame was defined to ensure ICD coding version consistency (ie, to include DX after October 2015; ICD10 implementation date). We only analyzed data after a biopsy (BX) to allow for accurate recording on the clinical side. However, 4 patients did not have BX data within the selected time window and were excluded. One patient was excluded because of a confirmed neurofibromatosis DX, which made the patient not clinically comparable to other patients.

We selected our predictors based on the strength of correlation (Nagelkerke  $R^2$ )<sup>39</sup> with each outcome variable, because it solves Cox-Snell correlation's<sup>40</sup> upper bound issues and provides a generalized correlation metric similar to our modeling approach.<sup>41</sup> We used a stepwise selection approach for covariate inclusion. To test our hypothesis, we built logistic regressions<sup>41</sup> to predict the odds of accurate, inaccurate, and suboptimal DX across patient charts using R's generalized linear model package.<sup>42</sup> Some of our variables had too many categories to be useful in our model. Thus, we reclassified clinical departments into oncology and nononcology departments and represented user factors via provider type rather than using user identification numbers. We maximized goodness of fit using Akaike's information criterion.<sup>43</sup> We tested for variable interactions and collinearity effects in all models with more than one predictor. Adjustments for multiple comparison were made using R's `p.adjust` function<sup>44</sup> selecting Holm's correction method.<sup>45</sup> We reran each regression using a time window of 90 days before and after the BX to confirm the effect's robustness.<sup>46</sup> Data extraction was performed with DataGrip software (version 2017.2.2; JetBrains, Prague, Czech Republic), exploratory analysis relied on Tableau (version 10.2.4; Tableau Software, Seattle, WA), and graphics were generated using Prism (version 8; GraphPad Software, San Diego, CA). Data cleaning and statistical analyses relied on R version 3.6.1<sup>30</sup> and RStudio (version 1.2.1335; RStudio, Boston, MA).

## RESULTS

The final analytical data set contained 10,052 order DX entries and 3,718 primary encounter DX entries for 31 patients (Table 1). Each patient had at least one encounter

per visit and at least one order per encounter that would each have DX entries attached for clinical care and billing purposes. Minimum and maximum follow-up times were 119 and 1,185 days, respectively, with an average follow-up time of  $654 \pm 308$  (mean  $\pm$  standard deviation). Order DX entries contained 1,899 (18.9%) accurate records and 1,536 (15.3%) inaccurate records; the remaining were suboptimal. There were 180 visit providers, 108 order-authorizing providers for 27 different kinds of orders, and 431 distinct users over 63 departments, across 8 care locations, covering 23 clinical specialties recorded in this data set. Encounter DX entries contained 712 (19.1%) accurate DX and 522 (14.0%) inaccurate DX, recorded in 66 departments across 8 care locations, covering 28 clinical specialties by 162 visit providers.

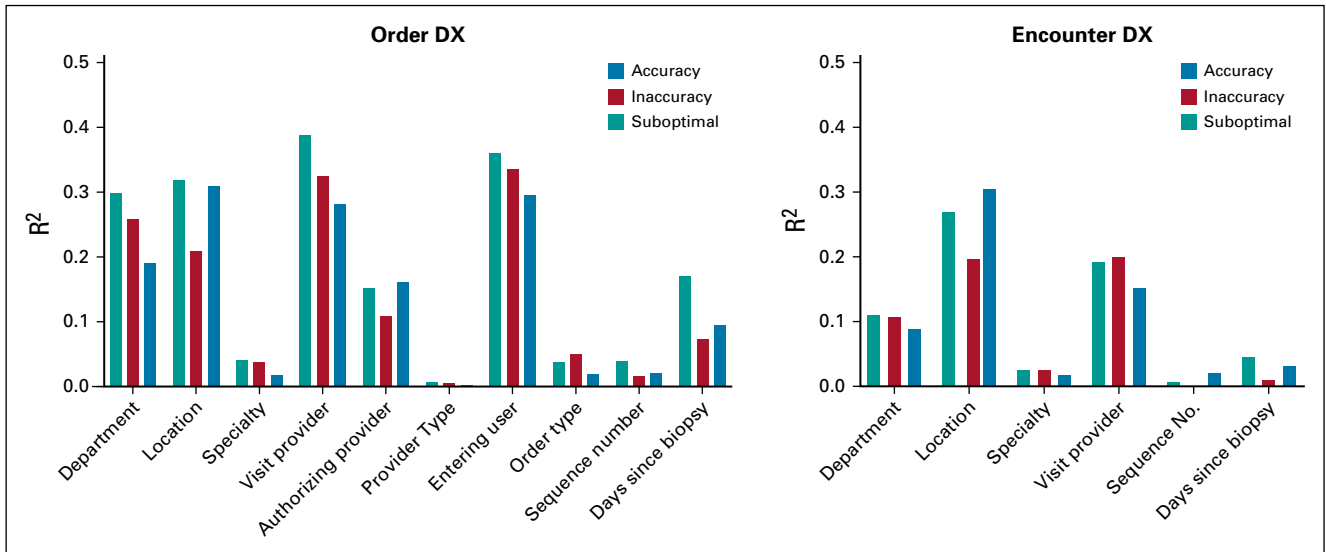
Correlation analysis revealed relatively strong relationships between workflow variables and accuracy (Fig 1). Department and care location had high correlation for most outcomes on both data sets (ie, order and encounter DX). Department presented 0.31, 0.26, and 0.30 correlation values for order accurate, inaccurate, and suboptimal DX, respectively. Care location presented 0.31, 0.21, 0.32 correlation values, respectively. Clinical personnel variables also returned higher correlation values. User entering the data ( $R^2 = 0.3, 0.34, 0.36$ ), visit provider ( $R^2 = 0.29, 0.33, 0.39$ ), and authorizing provider ( $R^2 = 0.16, 0.11, 0.16$ ), respectively, had high correlation values for all outcomes for order DX data. For encounter DX data, department

**TABLE 1.** Data Set Size and Features

Feature	DX Data Provenance	
	Order DX	Encounter DX
Patients	31	31
DX entries	10,052	3,718
Accurate DX entries	1,899	712
Inaccurate DX entries	1,536	552
Hospital departments	63	66
Care locations	8	8
Specialties	23	28
Visit providers	180	162
Authorizing providers	108	—
Entering users	431	—
Order types	27	—
Days since biopsy (mean $\pm$ SD)	$259 \pm 348$	$253 \pm 313$

NOTE. Data are No. unless otherwise indicated. We built our regressions using 2 data sets containing DX data from 2 segments of our oncologic EHR. One data set contained all DX entries attached to clinical orders across 31 patient charts. The second data set contained all DX entries attached to encounter data structures for 31 patients in our oncology EHR. No entering user data were available within encounter DX EHR data tables.

Abbreviations: DX, diagnosis; EHR, electronic health record; SD, standard deviation.



**FIG 1.** Correlation analysis results. DX, diagnosis.

( $R^2 = 0.31, 0.21, 0.32$ ), care location ( $R^2 = 0.30, 0.20, 0.27$ ), and visit provider ( $R^2 = 0.15, 0.20, 0.19$ ), respectively, had relatively high correlation values. All high correlation values were significant ( $P < .0001$ ).

We found significant statistical relationships between clinical workflow and user variables in the order DX data set (Table 2). For clinical workflow variables, DX entered in a nononcology department (eg, surgery, magnetic resonance imaging, outpatient laboratories, and so on) had 2.89 times higher odds of being fully accurate (ie, correct and specific histology and anatomic site; adjusted [adj]- $P < .0001$ ) and 49% lower odds of being suboptimal (adj- $P < .0001$ ) compared with oncology departments (eg, oncologic hematology, radiation oncology, and so on). We also found that DX data recorded at care locations where patients were seen regularly presented increased accuracy and reduced inaccuracy odds. For example, a DX entered at a care location including a cancer survivorship center had 68 times higher odds of being fully accurate (adj- $P < .0001$ ), was 95% less likely to be inaccurate (adj- $P < .0001$ ), and had 62% lower odds of being suboptimal (adj- $P < .0001$ ) compared with the main campus care location, which included the comprehensive cancer center. Similar results were found for a care location including a geriatric outpatient clinic and a chronic disease management facility. DX entered at these facilities had 19 and 8 times higher odds of being accurate (adj- $P < .0001$ ), respectively, 88% and 62% lower odds of being suboptimal (adj- $P < .0001$ ), respectively, and 98% lower odds of being inaccurate in the chronic disease management center (adj- $P < .0001$ ). Imaging facilities were the only care locations with significantly less accurate logging. A DX entered at an imaging facility location had 99% lower odds of being accurate (adj- $P < .0001$ ) and over 400 times higher odds

of being suboptimal (adj- $P < .0001$ ) compared with DX data recorded at the main campus. For our user variable, we found that a DX entered by a physician assistant had 16% lower odds of being fully accurate (adj- $P < .0001$ ) but only 9% higher odds of being inaccurate ( $P = .0015$ ) compared with DX entered by physicians. We also found differences in the logging habits of pharmacists. DX entered by users with pharmacist roles in the system tended to be much more accurate (odds ratio [OR], 2.99;  $P = .012$ ; adj- $P = .059$ ) and had 85% lower odds of being suboptimal (adj- $P = .005$ ).

We further tested this relationship between clinical workflow variables and data quality by rebuilding this regression using the encounter DX data (Table 3) and by rerunning our initial regression on a data set containing only data within a 90-day range of the BX. We found that a DX entered in a nononcology department had 66% higher odds of being fully accurate (adj- $P = 0.008$ ) and 28% lower odds of being suboptimal (adj- $P = .025$ ) compared with oncology departments. A DX entered at a care location including a cancer survivorship center had over 36 times higher odds of being fully accurate (adj- $P < .0001$ ), 89% lower odds of being inaccurate (adj- $P = .017$ ), and 50% lower odds of being suboptimal (adj- $P = .003$ ) compared with the main campus care location. DX entered at a geriatric outpatient clinic and a chronic disease management facility had over 7 and 3 times the odds of being accurate (adj- $P < .0001$ ), respectively, 96% lower odds of being inaccurate (adj- $P < .0001$ ) for the chronic disease management center, and 50% lower odds of being suboptimal in the outpatient geriatric care facility (adj- $P = .003$ ). A DX entered at an imaging facility had 95% lower odds of being accurate (adj- $P < .0001$ ) and 66 times higher odds of being suboptimal (adj- $P = .025$ )

compared with DX data recorded at the main campus. We also rebuilt our models for data sets containing data for the first 90 days after each patient's BX to further confirm our findings. The resulting model for order DX data confirmed our finding on user differences revealing differences between physicians and physician assistants (OR, 6.3,  $adj-P < .0001$  for accurate DX; OR, 0.38,  $adj-P < .0001$  for inaccurate DX; OR, 0.33,  $adj-P < .0001$  for suboptimal DX). This model also revealed the same differences for clinical departments (OR, 3.5,  $adj-P < .0001$  for accurate DX; OR, 0.44,  $adj-P < .0001$  for suboptimal DX) and for the outpatient geriatric care facility (OR, 8.0,  $adj-P < .0001$  for accurate DX; OR, 0.44,  $adj-P < .0001$  for suboptimal DX). Interestingly, we found the opposite effect for the

chronic disease management outpatient facility (OR, 0.47,  $adj-P < .0001$  for accurate DX; OR, 6.8,  $adj-P < .0001$  for suboptimal DX) hinting at a transient effect where such clinics might be more likely to record less accurate and more suboptimal DX data early in cancer treatments. Rebuilding the model for encounter DX confirmed that geriatric care facilities were more likely to record accurate DX (OR, 10.8;  $adj-P < .0001$ ) and less likely to record suboptimal DX (OR, 0.29;  $adj-P = .035$ ).

## DISCUSSION

We used statistical regressions to uncover potential relationships between accurate, inaccurate, and suboptimal DX data entry in oncologic EHRs and clinical

**TABLE 2.** Order DX Regression Results

Model	Term	Odds Ratio Exp ( $\beta$ )	Estimate ( $\beta$ )	SE	95% CI		P	Adjusted P
					Lower	Upper		
Accurate DX	Department							
	Oncology (ref)	1	0	—	—	—	—	—
	Nononcology	2.89	1.06	0.10	0.88	1.25	< .0001	< .0001
	Care location							
	Main campus and cancer center (ref)	1	0	—	—	—	—	—
	Imaging center	0.01	-4.89	0.24	-5.40	-4.46	< .0001	< .0001
	Outpatient geriatric care	19.14	2.95	0.08	2.80	3.11	< .0001	< .0001
	Cancer survivorship center	68.77	4.23	0.26	3.75	4.79	< .0001	< .0001
	Chronic disease management facility	8.74	2.17	0.04	2.08	2.25	< .0001	< .0001
	Other outpatient facilities	0	-16.06	68.89	-93	-109	0.81	1
	Clinical role							
	Physician (ref)	1	0	—	—	—	—	—
	Physician assistant	0.84	-0.18	0.03	-0.24	-0.11	< .0001	< .0001
	Pharmacist	2.99	1.10	0.44	0.21	1.94	.012	.059
Nurse practitioner	3.36	1.21	1.16	-0.81	3.62	.29	1	
Inaccurate DX	Department							
	Oncology (ref)	1	0	—	—	—	—	—
	Nononcology department	1.15	0.14	0.07	-0.01	0.29	.062	.49
	Care location							
	Main campus and cancer center (ref)	1	0	—	—	—	—	—
	Imaging center	0.001	-18.49	106.51	-135	-167	.86	1
	Outpatient geriatric care	0.00	-18.47	325.70	-313	-278	.95	1
	Cancer survivorship center	0.06	-2.75	0.46	-3.80	-1.97	< .0001	< .0001
	Chronic disease management facility	0.02	-3.78	0.22	-4.25	-3.37	< .0001	< .0001
	Other outpatient facilities	9.5e-9	-18.47	188.27	-223	-283	.92	1
	Clinical role							
	Physician (ref)	1	0	—	—	—	—	—
	Physician assistant	1.09	0.09	0.03	0.03	0.14	.0016	0.015
	Pharmacist	2.24	0.81	0.44	-0.09	1.67	.068	0.49
Nurse practitioner	1.85	0.61	1.17	-2.42	2.64	.6	1	

(Continued on following page)

**TABLE 2.** Order DX Regression Results (Continued)

Model	Term	Odds Ratio			95% CI		P	Adjusted P
		Exp ( $\beta$ )	Estimate ( $\beta$ )	SE	Lower	Upper		
Suboptimal DX	Department							
	Oncology (ref)	1	0	—	—	—	—	—
	Nononcology department	0.51	-0.68	0.06	-0.80	-0.55	< .0001	< .0001
	Care location							
	Main campus and cancer center (ref)	1	0	—	—	—	—	—
	Imaging center	458	6.13	0.24	5.70	6.63	< .0001	< .0001
	Outpatient geriatric care	0.18	-1.72	0.08	-1.88	-1.57	< .0001	< .0001
	Cancer survivorship center	0.04	-3.32	0.31	-3.99	-2.76	< .0001	< .0001
	Chronic disease management facility	0.38	-0.97	0.04	-1.05	-0.88	< .0001	< .0001
	Other outpatient facilities	3.2e7	17.30	69.01	76.38	86.57	.8	1
	Clinical role							
	Physician (ref)	1	0	—	—	—	—	—
	Pharmacist	0.15	-1.87	0.56	-3.07	-0.86	.0008	.005
	Physician assistant	1.04	0.04	0.02	-0.01	0.09	.087	.43
Nurse practitioner	0.27	-1.31	1.22	-4.39	0.80	.28	1	

NOTE. Each DX regression model predicts the number of DX descriptions based on department, care location, and the data-entering user's clinical role for 3 DX types: Accurate DX descriptions (ie, descriptions containing the accurate tumor type and anatomic site), Inaccurate DX descriptions (ie, descriptions containing inaccurate tumor type or anatomic site), and Suboptimal DX descriptions (ie, DX descriptions that were neither accurate nor inaccurate).

Abbreviations: DX, diagnosis; ref, reference.

workflow-describing variables, such as care location, clinical department, and EHR user types. We found that both clinical department and care location were predictive of accurate or suboptimal DX entry. Care location was also predictive of inaccurate recording. We also found that a user's clinical role predicted the odds of accurate, inaccurate, or suboptimal entry. Our findings support the hypothesis that clinical workflow factors affect the accuracy of clinical data recording; they also suggest that there may be significant differences across clinical workflows and clinical personnel's logging habits. Clinical oncologists using EHR data to make patient care decisions or reusing clinical data to develop clinical decision support tools should be aware of this variability.

Our study expands beyond existing literature by exploring DX code assignment beyond accuracy<sup>1,22,47</sup> and shifts the view of static clinical data as raw analysis material to data as the product of clinical workflows. Our findings are congruent with the existing literature<sup>15-17,36</sup> but also unlock new dimensions of oncologic data quality assurance. We provide quantitative evidence of the heterogeneous nature of clinical workflows and, most importantly, their impact on clinical data quality. This work also provides preliminary evidence to suggest that different clinical departments, care locations, and clinical roles may have different data logging, which we were not able to find reported quantitatively in prior publications. This is the core contribution of our analysis.

Our findings hint at differences in the ideal level of DX specificity across clinical workflows, care locations, and clinical departments. One interesting finding is the higher degree of accuracy and lower suboptimal logging odds in nononcology departments, potentially explained by oncologic progress note accessibility at oncologic departments. There may be a reliance on unstructured clinical data for information foraging<sup>48</sup> during clinical practice at these sites. This is congruent with the idea that the most accurate and precise DX is contained in the clinical progress note.<sup>49-51</sup> However, this raises questions about the accessibility of oncology notes to users outside oncology departments and in the postclinical data lifecycle.<sup>52</sup> Current clinical data reuse research seeks to develop data quality assessment methods separately from clinical practice.<sup>6,53-64</sup> However, our work shows that clinical data and clinical workflows are closely related and should be viewed as a product and production process. There may also be a link between interface design and entry accuracy, given that order and encounter entry interfaces are different in our EHR system. This may counter current data aggregation and warehousing trends<sup>65,66</sup> but would allow for a data quality assessment and reuse approach more in tune with the paradigm of learning health care.<sup>67</sup>

Our analysis has 5 core limitations that will be addressed in future work. First, we analyzed data for 36 patients with cancer. Because we relied on a clinician-defined gold standard, developing a larger cohort would have been labor

**TABLE 3.** Encounter DX Regression Results

Model	Term	Odds Ratio			95% CI		P	Adjusted P
		Exp ( $\beta$ )	Estimate ( $\beta$ )	SE	Lower	Upper		
Accurate DX	Department							
	Oncology (ref)	1	0	—	—	—	—	—
	Nononcology department	1.66	0.51	0.16	0.19	0.84	.002	.008
	Care location							
	Main campus and cancer center (ref)	1	0	—	—	—	—	—
	Imaging center	0.05	-2.94	0.58	-4.34	-1.97	< .0001	< .0001
	Outpatient geriatric care	7.10	1.90	0.20	1.55	2.37	< .0001	< .0001
	Cancer survivorship center	38.60	3.60	0.40	2.92	4.54	< .0001	< .0001
	Chronic disease management facility	3.58	1.27	0.15	0.98	1.56	< .0001	< .0001
Other outpatient facilities	5.0e-8	-16.80	401	-188	-277	.96	1	
Inaccurate DX	Department							
	Oncology (ref)	1	0	—	—	—	—	—
	Nononcology department	1.07	0.06	0.14	-0.21	0.35	.65	1
	Care location							
	Main campus and cancer center (ref)	1	0	—	—	—	—	—
	Imaging center	2.7e-8	-17.41	378	-180	-288	.96	1
	Outpatient geriatric care	2.7e-8	-17.41	642	-264	3.53	.97	1
	Cancer survivorship center	0.11	-2.18	0.72	-3.99	-1.01	.002	.017
	Chronic disease management facility	0.06	-2.83	0.51	-4.01	-1.97	< .0001	< .0001
Other outpatient facilities	2.7e-8	-17.41	402	-189	-272	.96	1	
Suboptimal DX	Department							
	Oncology (ref)	1	0	—	—	—	—	—
	Nononcology department	0.72	-0.33	-0.57	-0.10	-2.80	.005	.025
	Care location							
	Main campus and cancer center (ref)	1	0	—	—	—	—	—
	Imaging center	66.00	4.19	3.22	5.59	7.20	< .0001	< .0001
	Outpatient geriatric care	0.50	-0.70	-1.11	-0.31	-3.44	.0005	.003
	Cancer survivorship center	0.06	-2.77	-3.83	-1.95	-5.89	< .0001	< .0001
	Chronic disease management facility	0.90	-0.11	-0.39	0.18	-0.74	.46	1
Other outpatient facilities	2.8e7	17.15	170	121	0.07	.94	1	

NOTE. Each DX regression model predicts the number of DX descriptions based on department and care location for 3 DX types: Accurate DX descriptions (ie, descriptions containing the accurate tumor type and anatomic site), Inaccurate DX descriptions (ie, descriptions containing inaccurate tumor type or anatomic site), and Suboptimal DX descriptions (ie, DX descriptions that were neither accurate or inaccurate).

Abbreviations: DX, diagnosis; ref, reference.

intensive and cost prohibitive at this stage. Still, our final data set provided adequate statistical power to test our hypothesis and previous work.<sup>37</sup> Second, we relied on data from a single institution, which may limit the external validity of our findings. We will address this limitation in future work by replicating our analysis at multiple sites. Third, we had a simple definition of DX accuracy. The definition was based on information available in our EHR and information needed to identify patient charts for secondary analysis. This is in concordance with the clinical data quality literature,<sup>68,69</sup> which recommends accuracy definitions to fit the intended data use. Fourth,

limited clinical workflow and user factors were explored because of availability and analytic method limitations; other variables will be explored in future work. Finally, we only studied one type of cancer. Future work will include reproducing this analysis for other patient cohorts to assess generalizability. We will use a larger cohort of patients and cohorts diagnosed with other oncologic conditions.<sup>37</sup> We will also explore the availability of user interaction data within our EHR database to better understand whether data entry modes (eg, drop-down menu selection and search interfaces) have an impact on the accuracy of DX data entry in oncologic EHRs. Additional secondary

analyses of clinical data using analytic methods, such as machine learning<sup>70</sup> and simulation modeling techniques,<sup>71</sup> will also be conducted.

In conclusion, clinical departments and care locations were predictive of DX data quality. A user's clinical role (eg, physician, assistant) was also predictive of accurate,

inaccurate, and suboptimal DX data recording. Clinical oncologists using EHR data to make care decisions or reusing data to develop clinical decision support tools should take such differences into account. Additional analytic work is needed to tease out this heterogeneity in clinical data recording.<sup>1,2,5</sup>

## AFFILIATIONS

<sup>1</sup>University of North Carolina at Charlotte, Charlotte, NC

<sup>2</sup>Wake Forest School of Medicine, Winston Salem, NC

<sup>3</sup>North Carolina State University, Raleigh, NC

## CORRESPONDING AUTHOR

Franck Diaz-Garelli, PhD, University of North Carolina at Charlotte, Department of Public Health Sciences, College of Health and Human Services, 9201 University City Blvd, Charlotte, NC 28223-0001; e-mail: franck.diaz@uncc.edu.

## SUPPORT

Partially supported by the Cancer Center Support Grant from the National Cancer Institute to the Comprehensive Cancer Center of Wake Forest Baptist Medical Center (P30 CA012197), by the National Institute of General Medical Sciences' Institutional Research and Academic Career Development Award program (K12-GM102773), and by WakeHealth's Center for Biomedical Informatics' first Pilot Award.

## AUTHOR CONTRIBUTIONS

**Conception and design:** Franck Diaz-Garelli, Roy Strowd, Brian J. Wells, Umit Topaloglu

**Financial support:** Franck Diaz-Garelli, Umit Topaloglu

**Administrative support:** Franck Diaz-Garelli

**Provision of study materials or patients:** Franck Diaz-Garelli, Roy Strowd, Umit Topaloglu

**Collection and assembly of data:** Franck Diaz-Garelli, Roy Strowd, Virginia L. Lawson, Umit Topaloglu

**Data analysis and interpretation:** Franck Diaz-Garelli, Roy Strowd, Maria E. Mayorga, Brian J. Wells, Thomas W. Lycan, Umit Topaloglu

**Manuscript writing:** All authors

**Final approval of manuscript:** All authors

**Accountable for all aspects of the work:** All authors

## AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to [www.asco.org/rwc](http://www.asco.org/rwc) or [ascopubs.org/cci/author-center](http://ascopubs.org/cci/author-center).

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians ([Open Payments](http://Open Payments)).

### Roy Strowd

**Consulting or Advisory Role:** Monteris Medical, Novocure

**Research Funding:** Southeastern Brain Tumor Foundation, Jazz Pharmaceuticals, National Institutes of Health

**Other Relationship:** American Academy of Neurology

### Virginia L. Lawson

**Employment:** Walgreens

**Stock and Other Ownership Interests:** Walgreens

### Brian J. Wells

**Research Funding:** Boehringer Ingelheim (Inst)

### Thomas W. Lycan

**Travel, Accommodations, Expenses:** Incyte Corporation

**Open Payments Link:** <https://openpaymentsdata.cms.gov/physician/4354581>

No other potential conflicts of interest were reported.

## ACKNOWLEDGMENT

The authors acknowledge use of the services and facilities, funded by the National Center for Advancing Translational Sciences, National Institutes of Health (UL1TR001420), and support provided by the National Science Foundation Research Experience for Undergraduates IMPACT award (1559700).

## REFERENCES

1. Safran C: Reuse of clinical data. *Yearb Med Inform* 9:52-54, 2014
2. Safran C: Using routinely collected data for clinical research. *Stat Med* 10:559-564, 1991
3. Thompson ME, Dulin MF: Leveraging data analytics to advance personal, population, and system health: Moving beyond merely capturing services provided. *N C Med J* 80:214-218, 2019
4. Hershman DL, Wright JD: Comparative effectiveness research in oncology methodology: Observational data. *J Clin Oncol* 30:4215-4222, 2012
5. Hersh WR, Weiner MG, Embi PJ, et al: Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 51S30-S37, 2013 (8; suppl 3)
6. Brown JS, Kahn M, Toh S: Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care* 51:S22-S29, 2013 (8; suppl 3)
7. Mailankody S, Prasad V: Comparative effectiveness questions in oncology. *N Engl J Med* 370:1478-1481, 2014
8. Garraway LA, Verweij J, Ballman KV: Precision oncology: An overview. *J Clin Oncol* 31:1803-1805, 2013
9. Johnson A, Zeng J, Bailey AM, et al: The right drugs at the right time for the right patient: The MD Anderson precision oncology decision support platform. *Drug Discov Today* 20:1433-1438, 2015



10. Kurnit KC, Dumbraava EE, Litzenburger B, et al: Precision oncology decision support: Current approaches and strategies for the future. *Clin Cancer Res* 24:2719-2731, 2018
11. Beckmann JS, Lew D: Reconciling evidence-based medicine and precision medicine in the era of big data: Challenges and opportunities. *Genome Med* 8:134, 2016
12. Köpcke F, Prokosch H-U: Employing computers for the recruitment into clinical trials: A comprehensive systematic review. *J Med Internet Res* 16:e161, 2014
13. Moskowitz A, Chen K: Defining the patient cohort, in MIT Critical Data (ed): Secondary Analysis of Electronic Health Records. New York, NY, Springer International Publishing, 2016, pp 93-100
14. Hsia DC, Krushat WM, Fagan AB, et al: Accuracy of diagnostic coding for Medicare patients under the Prospective-Payment System. <http://dx.doi.org/10.1056/NEJM198802113180604>
15. Doremus HD, Michenzi EM: Data quality. An illustration of its potential impact upon a diagnosis-related group's case mix index and reimbursement. *Med Care* 21:1001-1011, 1983
16. Lloyd SS, Rissing JP: Physician and coding errors in patient records. *JAMA* 254:1330-1336, 1985
17. Johnson AN, Appel GL: DRGs and hospital case records: Implications for Medicare case mix accuracy. *Inquiry* 21:128-134, 1984
18. Botsis T, Hartvigsen G, Chen F, et al: Secondary use of EHR: Data quality issues and informatics opportunities. *Summit on Translat Bioinforma* 2010:1-5, 2010
19. Blois MS: Information and Medicine: The Nature of Medical Descriptions. Berkeley, CA, University of California Press, 1984
20. Sacchi L, Dagliati A, Bellazi R: Analyzing complex patients' temporal histories: New frontiers in temporal data mining, in Fernández-Llatas C, García-Gómez JM (eds): *Data Mining in Clinical Medicine*. New York, NY, Humana Press, 2015, pp 89-105
21. Farzandipour M, Sheikhtaheri A, Sadoughi F: Effective factors on accuracy of principal diagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10). *Int J Inf Manage* 30:78-84, 2010
22. O'Malley KJ, Cook KF, Price MD, et al: Measuring diagnoses: ICD code accuracy. *Health Serv Res* 40:1620-1639, 2005
23. Baskaran LNGM, Greco PJ, Kaelber DC: Case report medical eponyms: An applied clinical informatics opportunity. *Appl Clin Inform* 3:349-355, 2012
24. Fleming M, MacFarlane D, Torres WE, et al: Magnitude of impact, overall and on subspecialties, of transitioning in radiology from ICD-9 to ICD-10 codes. *J Am Coll Radiol* 12:1155-1161, 2015
25. Pippenger M, Holloway RG, Vickrey BG: Neurologists' use of ICD-9CM codes for dementia. *Neurology* 56:1206-1209, 2001
26. Sarmiento RF, Demoncourt F: Improving patient cohort identification using natural language processing, in MIT Critical Data: Secondary analysis of electronic health records. New York, NY, Springer International Publishing, 2016, pp 405-417
27. Jack A, Organization WH, Percy CL, et al: International Classification of Diseases for Oncology: ICD-O. Geneva, Switzerland, World Health Organization, 2000
28. Diaz-Garelli J-F, Wells BJ, Yelton C, et al: Biopsy records do not reduce diagnosis variability in cancer patient EHRs: Are we more uncertain after knowing? *AMIA Jt Summits Transl Sci Proc* 2017:72-80, 2018
29. Walji MF, Kalenderian E, Tran D, et al: Detection and characterization of usability problems in structured data entry interfaces in dentistry. *Int J Med Inform* 82:128-138, 2013
30. Howard J, Clark EC, Friedman A, et al: Electronic health record impact on work burden in small, unaffiliated, community-based primary care practices. *J Gen Intern Med* 28:107-113, 2013
31. Asan O, Nattinger AB, Gurses AP, et al: Oncologists' views regarding the role of electronic health records in care coordination. *JCO Clin Cancer Inform* 2:1-12, 2018
32. Mount-Campbell AF, Hosseinzadeh D, Gurcan M, et al: Applying human factors engineering to improve usability and workflow in pathology informatics. *Proc Int Symp Hum Factors Ergon Healthc* 6:23-27, 2017
33. Hribar MR, Biermann D, Read-Brown S, et al: Clinic workflow simulations using secondary EHR data. *AMIA Annu Symp Proc* 2016:647-656, 2017
34. Zheng K, Haftel HM, Hirsch RB, et al: Quantifying the impact of health IT implementations on clinical workflow: A new methodological perspective. *J Am Med Inform Assoc* 17:454-461, 2010
35. Ozkaynak M, Unerti KM, Johnson A, et al: Clinical workflow analysis, process redesign, and quality improvement, in Finnell JT, Dixon BE (eds): *Clinical Informatics Study Guide: Text and Review*. New York, NY, Springer International Publishing, 2016, pp 135-161
36. Diaz-Garelli J-F, Strowd R, Wells BJ, et al: Lost in translation: Diagnosis records show more inaccuracies after biopsy in oncology care EHRs. *AMIA Jt Summits Transl Sci Proc* 2019:325-334, 2019
37. Diaz-Garelli J-F, Strowd R, Ahmed T, et al: A tale of three subspecialties: Diagnosis recording patterns are internally consistent but specialty-dependent. *Jamia Open* 2:369-377, 2019
38. van der Bij S, Khan N, Ten Veen P, et al: Improving the quality of EHR recording in primary care: A data quality feedback tool. *J Am Med Inform Assoc* 24:81-87, 2017
39. Nagelkerke NJD: A note on a general definition of the coefficient of determination. *Biometrika* 78:691-692, 1991
40. Cox DR, Snell EJ: *Analysis of Binary Data* (ed 2). Abingdon-on-Thames, United Kingdom, Taylor & Francis, 1989
41. Tabachnick BG, Fidell LS: *Using multivariate statistics* (ed 5). Boston, MA, Allyn & Bacon/Pearson Education, 2007
42. R Core Team: R: A language and environment for statistical computing. <http://www.R-project.org/>
43. Pan W: Akaike's information criterion in generalized estimating equations. *Biometrics* 57:120-125, 2001
44. Wright SP: Adjusted P-values for simultaneous inference. *Biometrics* 48:1005, 1992
45. Aickin M, Gensler H: Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *Am J Public Health* 86:726-728, 1996
46. Diaz-Garelli JF, Bernstam EV, Mse, et al: Rediscovering drug side effects: The impact of analytical assumptions on the detection of associations in EHR data. *AMIA Jt Summits Transl Sci Proc* 2015:51-55, 2015
47. Escudíé J-B, Rance B, Malamut G, et al: A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: A case study on autoimmune comorbidities in patients with celiac disease. *BMC Med Inform Decis Mak* 17:140, 2017
48. Gibson B, Butler J, Zirkle M, et al: Foraging for information in the EHR: The search for adherence related information by mental health clinicians. *AMIA Annu Symp Proc* 2016:600-608, 2017
49. Wei W-Q, Teixeira PL, Mo H, et al: Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* 23:e20-e27, 2016
50. Fisher ES, Whaley FS, Krushat WM, et al: The accuracy of Medicare's hospital claims data: Progress has been made, but problems remain. *Am J Public Health* 82:243-248, 1992

51. Spellberg B, Harrington D, Black S, et al: Capturing the diagnosis: An internal medicine education program to improve documentation. *Am J Med* 126:739-743.e1, 2013
52. Lu Z, Su J: Clinical data management: Current status, challenges, and future directions from industry perspectives. *Open Access Journal of Clinical Trials* 2:93-105, 2010
53. Diaz-Garelli J-F, Bernstam EV, Lee M, et al: DataGauge: A practical process for systematically designing and implementing quality assessments of repurposed clinical data. *eGEMS* 7:32, 2019
54. Weiskopf NG, Weng C: Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J Am Med Inform Assoc* 20:144-151, 2013
55. Kahn M, Brown J, Chun A, et al: Transparent reporting of data quality in distributed data networks. *eGEMS* 3, 2015. doi: [10.13063/2327-9214.1052](https://doi.org/10.13063/2327-9214.1052)
56. Kahn MG, Callahan TJ, Barnard J, et al: A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMS* 4, 2016. doi: [10.13063/2327-9214.1244](https://doi.org/10.13063/2327-9214.1244)
57. Callahan TJ, Barnard JG, Helmkamp LJ, et al: Reporting data quality assessment results: Identifying individual and organizational barriers and solutions. *eGEMS* 5, 2017. doi: [10.5334/egems.214](https://doi.org/10.5334/egems.214)
58. Hripcsak G, Duke JD, Shah NH, et al: Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 216:574-578, 2015
59. Kahn MG, Raebel MA, Glanz JM, et al: A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care* 50, 2012
60. Johnson SG, Speedie S, Simon G, et al: A data quality ontology for the secondary use of EHR data. *AMIA Annu Symp Proc* 2015:1937-1946, 2015
61. Johnson SG, Speedie S, Simon G, et al: Application of an ontology for characterizing data quality for a secondary use of EHR data. *Appl Clin Inform* 7:69-88, 2016
62. DQe-v: A database-agnostic framework for exploring variability in electronic health record data across time and site location. <https://egems.academyhealth.org/articles/10.13063/2327-9214.1277/>
63. Jose Franck Diaz: DataGauge: A model-driven framework for systematically assessing the quality of clinical data for secondary use, 2016. [https://digitalcommons.library.tmc.edu/uthshis\\_dissertations/33/](https://digitalcommons.library.tmc.edu/uthshis_dissertations/33/)
64. Holve E, Kahn M, Nahm M, et al: A comprehensive framework for data quality assessment in CER. *AMIA Jt Summits Transl Sci Proc* 2013:86-88, 2013
65. Sahama TR, Croll PR: A data warehouse architecture for clinical data warehousing. <http://dl.acm.org/citation.cfm?id=1274531.1274560>
66. Singh R, Singh K: A descriptive classification of causes of data quality problems in data warehousing. In *J Comput Sci Issues* 7:41-50, 2010
67. Olsen LA, Aisner D, McGinnis JM (eds): *The Learning Healthcare System: Workshop Summary*. Washington, DC, National Academies Press, 2007
68. Weiskopf NG, Rusanov A, Weng C: Sick patients have more data: The non-random completeness of electronic health records. *AMIA Annu Symp Proc* 2013:1472-1477, 2013
69. Weiskopf NG, Hripcsak G, Swaminathan S, et al: Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform* 46:830-836, 2013
70. Madhuri R, Murty MR, Murthy JVR, et al: Cluster analysis on different data sets using K-modes and K-prototype algorithms, in Satapathy SC, Avadhani PS, Udgata SK, et al (eds): *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India*, Visakhapatnam, India, December 13-15, 2013, Volume II. Springer International Publishing, 2014, pp 137-144
71. van Dam KH, Nikolic I, Lukszo Z. *Agent-Based Modelling of Socio-Technical Systems*. New York, NY, Springer Science & Business Media, 2012

