OXFORD

JNCI J Natl Cancer Inst (2021) 113(7): djaa210

doi: 10.1093/jnci/djaa210
First published online December 30, 2020
Article

Assessing Cancer History Accuracy in Primary Care Electronic Health Records Through Cancer Registry Linkage

Megan Hoopes, MPH,[1,*] Robert Voss, MS,[1] Heather Angier, PhD, MPH,[2] Miguel Marino, PhD,[2] Teresa Schmidt, PhD,[1] Jennifer E. DeVoe, MD, DPhil,[2] Jeffrey Soule, JD,[3] Nathalie Huguet, PhD[2]

[1]OCHIN, Inc, Portland, OR, USA; [2]Department of Family Medicine, Oregon Health and Science University, Portland, OR, USA; and and [3]Oregon State Cancer Registry, Oregon Health Authority, Portland, OR, USA

*Correspondence to: Megan Hoopes, MPH, OCHIN, Inc, Research Department, 1881 SW Naito Pkwy, Portland, OR 97201, USA (e-mail: hoopesm@ochin.org).

Abstract

Background: Many cancer survivors receive primary care in community health centers (CHCs). Cancer history is an important factor to consider in the provision of primary care, yet little is known about the completeness or accuracy of cancer history data contained in CHC electronic health records (EHRs). **Methods:** We probabilistically linked EHR data from more than1.5 million adult CHC patients to state cancer registries in California, Oregon, and Washington and estimated measures of agreement (eg, kappa, sensitivity, specificity). We compared demographic and clinical characteristics of cancer patients as estimated by each data source, evaluating distributional differences with absolute standardized mean differences. **Results:** A total 74 707 cancer patients were identified between the 2 sources (EHR only, n = 22 730; registry only, n = 23 616; both, n = 28 361). Nearly one-half of cancer patients identified in registries were missing cancer documentation in the EHR. Overall agreement of cancer ascertainment in the EHR vs cancer registries (gold standard) was moderate (kappa = 0.535). Cancer site–specific agreement ranged from substantial (eg, prostate and female breast; kappa > 0.60) to fair (melanoma and cervix; kappa < 0.40). Comparing population characteristics of cancer patients as ascertained from each data source, groups were similar for sex, age, and federal poverty level, but EHR-recorded cases showed greater medical complexity than those ascertained from cancer registries. **Conclusions:** Agreement between EHR and cancer registry data was moderate and varied by cancer site. These findings suggest the need for strategies to improve capture of cancer history information in CHC EHRs to ensure adequate delivery of care and optimal health outcomes for cancer survivors.

The majority of cancer survivors living today were diagnosed more than 5 years ago and have transitioned from oncological to primary care settings (1-3). Cancer survivors are at greater risk for cardiovascular disease, depression, and secondary cancers (1,4)—conditions routinely managed in primary care—than the general population. However, a fundamental challenge to providing optimal survivor care is the ability to identify and track cancer survivors within primary care electronic health records (EHRs) (5,6).

Outpatient EHRs may not capture complete or accurate cancer history information, especially at community health centers (CHCs), which are typically not connected to cancer centers. These "safety net" clinics provide health care to patients regardless of insurance status; their populations are largely low income, publicly insured or uninsured, and racial and ethnic minorities. CHCs represent an ideal setting to assess the accuracy of cancer information in EHRs because CHC patients also have disproportionate cancer risk profiles. For example, socioeconomically disadvantaged populations are less likely to receive timely cancer screenings, have higher rates of delayed diagnosis, and have higher rates of smoking (a modifiable risk factor for multiple cancer sites) (7-9) than more advantaged populations.

Approximately 7.1% of the US adult population are cancer survivors (10); however, 1 recent study that identified cancer survivors among patients in a large network of CHCs found a 3% prevalence of cancer history among adults (11), suggesting unreported cases. Yet, the accuracy and completeness of information on cancer history in outpatient EHRs has not been well described (6,12-14).

Population-based cancer registries exist in every US state and are repositories for complete, high-quality, standardized data on all incident cancers and cancer deaths (15,16). Thus, they can serve as the gold standard for validation studies. In 1

Received: June 23, 2020; **Revised:** September 23, 2020; **Accepted:** December 14, 2020

study that linked a general practice database to a population-based cancer registry in England, high sensitivity (>90%) was found for colorectal, lung, and gastro-esophageal cancers and moderate sensitivity (85%) for urological cancers (12). Another study successfully linked approximately 60% of primary care recorded cases of breast, prostate, lung, and colon cancers to a national cancer registry (6). Much greater levels of completeness and agreement were demonstrated within the EHRs of integrated health-care systems. For example, a study using data from a large health-care system in Northern California linked to the statewide cancer registry found only 2% of patients with a history of lung or bronchus, colorectal, female breast, or prostate cancer were missing from the EHR (14). This finding was likely because this multi-specialty system included hospitals and specialized cancer care and shared an EHR. The validity of clinical data contained in outpatient EHRs has been examined for cancer screening (17,18) and chronic disease preventive care (18), but less is known about information on cancer history. To address this gap, we linked EHR data from a national network of CHCs to 3 state cancer registries to assess the completeness and accuracy of EHR-recorded cancer history. We assessed (1) how well EHR-recorded cancer history data agree with the gold standard of cancer registries; (2) how accurate EHR-recorded cancer history data are with regard to primary site, date of diagnosis, and age at diagnosis; and (3) how demographic and clinical characteristics of patients with cancer compare when obtained from a single data source. Our study expands on previous literature by assessing agreement of all leading cancer sites in a large multistate CHC population.

Methods

EHR Patient Population

We used data from the OCHIN community health information network, a multistate collaboration of CHCs sharing a common instance of the EpicCare EHR. At the time of the study, OCHIN's primary care EHR data covered 68 health centers and 328 clinic sites serving more than 1.5 million adults (≥18 years of age) in California, Oregon, and Washington.

We collected identifying information on patients for linkage with each state's cancer registry. Patients with at least 1 CHC visit in a state were sent to that state's registry for linkage (California: N = 769 962; Oregon: N = 557 594; Washington: N = 199 619).

Cancer Registry Linkage

We conducted probabilistic linkage of EHR patient data to cancer registry records using first and last name, sex, date of birth, social security number, and zip code; street address, and race-ethnicitywere used to support the comparison of records for manual review but did not contribute to link scores. For Oregon and Washington, study staff conducted each linkage using Registry Plus Link Plus software (developed by the Centers for Disease Control and Prevention) (19) and a prespecified set of matching criteria. California Cancer Registry staff conducted that state's linkage using Match*Pro software (developed by the National Cancer Institute) (20). Each linkage used last name, date of birth, and social security number as blocking variables; California additionally used first name and zip code for the blocking step. Clerical review of potential matches was completed by individuals experienced with probabilistic linkage. See the Supplementary Methods (available online) for additional linkage details.

Each registry identified all cancer site records for linked patients, returning International Classification of Disease (ICD) for Oncology (ICD-O-3) codes (21), date of diagnosis, and age at diagnosis for individual cancer sites. Each registry released cancer cases diagnosed in that state, excluding those diagnosed at Veteran's Administration hospitals (per state agreements with the Veteran's Administration). Due to the timing of our requests for linkage and years of available data, the diagnosis years reported by the 3 registries differed slightly (California: 1998-2017; Oregon: 1996-2016; Washington: 1992-2016).

Identifying Cancer History in EHR Data

We identified EHR cancer history from ICD-9-CM and ICD-10-CM codes in problem list, medical history, and encounter diagnosis records. We limited our search to diagnosis codes indicating malignant cancers and benign brain or central nervous system tumors, excluding nonmelanoma skin cancers to align with cancer registry inclusion criteria. To standardize cancer site groupings, we mapped ICD-9-CM and ICD-10-CM codes to ICD-O-3 using National Cancer Institute's Surveillance, Epidemiology, and End Results Program documentation (22).

Determining Cancer History Overlap Between EHR and Registry Data

Cancer diagnoses were grouped into primary sites following classifications from the Surveillance, Epidemiology, and End Results Program (21). A small percentage of dates in each data source was missing month or day information; for these, we imputed mid-month or mid-year to construct a complete date, as is standard in cancer reporting.

Hereafter, we reference 3 units of analysis. Cancer patients are patients with any documented cancer history. Sites refer to individual cancer sites (eg, breast, colorectal); patients may have documentation of more than 1 site, and/or different sites recorded in the EHR and registry. Patient*site is the unique combination of a given cancer site for a given patient. Overall cancer status was ascertained from both EHR and registry data, with diagnosis date assigned as the earliest date of any cancer documented across the 2 data sources. For patient-level comparisons, patients were not required to have the same cancer site as cancer survivors to be considered in agreement. For patient*site comparisons, we required the same patient to have the same site documented in both the EHR and registry to be a match.

Statistical Analysis

We computed measures of agreement (sensitivity, specificity, kappa statistic, predictive value positive, and predictive value negative) (23) comparing overall patient-level cancer history and patient*site–level agreement using the registry as the gold standard. Kappa statistics were interpreted to have slight (0.00-0.20), fair (0.21-0.40), moderate (0.41-0.60), or substantial (>0.60) agreement (24). For each leading cancer site, we computed the percent of total cases identified by each data source. We described demographic and clinical characteristics of cancer patients observed by each data source independently. Distributional differences were assessed using absolute standardized mean differences (ASMD), an effect size measure

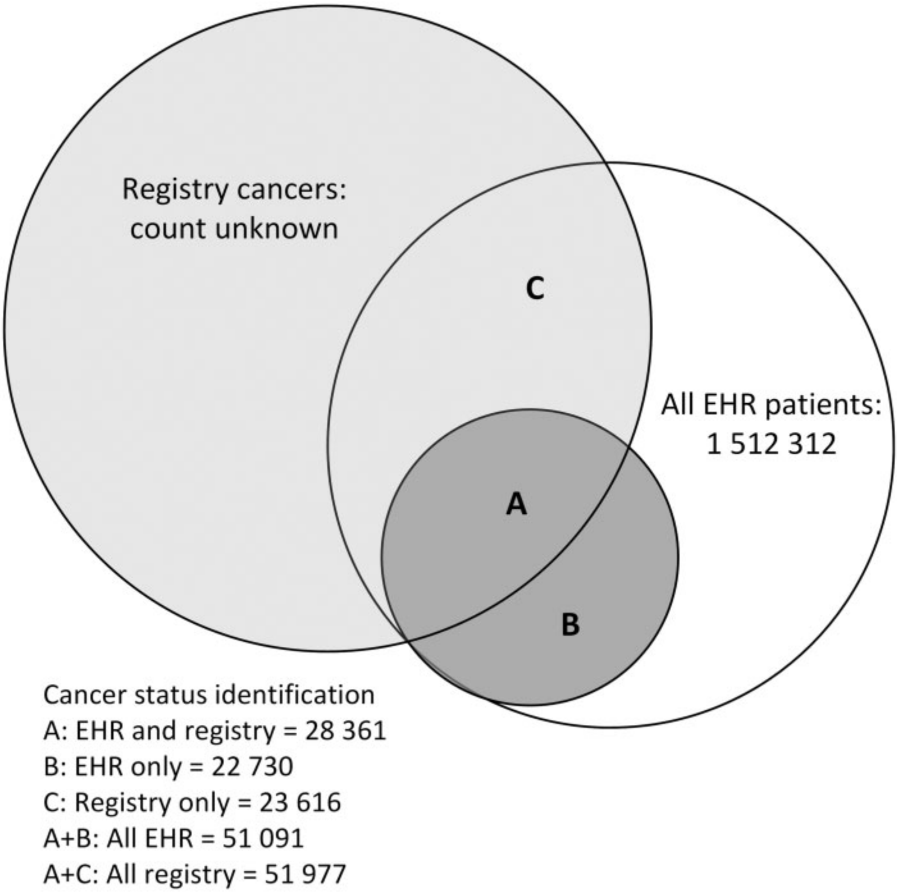
Registry cancers:
count unknown
C
All EHR patients:
1 512 312
A
B
Cancer status identification
A: EHR and registry = 28 361
B: EHR only = 22 730
C: Registry only = 23 616
A+B: All EHR = 51 091
A+C: All registry = 51 977

Figure 1. Diagram of electronic health record (EHR) linkage to California, Oregon, and Washington state cancer registries and resulting subgroups for validation analyses.

unaffected by sample size and appropriate for overlapping groups (25). We considered ASMDs greater than 0.1 to denote meaningful differences between the comparison groups (26).

All demographic and clinical variables came from the EHR. These included household income as percent of federal poverty level, insurance type, race or ethnicity, preferred language, and Charlson Comorbidity Index (27); all time-varying characteristics were calculated using each patient's latest visit. Because cancer is one of the components of the Charlson Comorbidity Index, we calculated a modified score by removing the cancer-related components from the calculation to provide an estimate of noncancer comorbidity (11). We limited analysis to adults (age ≥18 years as of January 1, 2019). We retained pediatric cancer histories if diagnosed before adulthood (<2% of cases). A subanalysis was conducted to examine agreement measures among the subset of patients with more recent cancer diagnoses (2012-2018). Data management and analysis were conducted using SAS software version 9.4 (SAS Institute, Inc, Cary, NC).

Results

Approximately 1.5 million patients from the EHR were eligible for linkage. Patients with EHR records in more than 1 state (n = 12 414) were sent to multiple registries. A total 51 977 patient records (3.4% of EHR patients) matched 1 or more of the registries. Approximately one-half of the patients identified in each source were also found in the other (Figure 1).

Across all 3 states, 74 707 cancer patients were identified between the 2 sources (EHR only, n = 22 730; registry only, n = 23 616; both, n = 28 361). Only 54.6% of those recorded in registries were found to have cancer history in both sources. Approximately equal numbers of patients were indicated to have cancer in EHR (n = 22 730) or registry (n = 23 616) data only. Results by state were similar (Table 1). For all states, the overall agreement of any cancer history in the EHR compared with registries was moderate (kappa = 0.535) and sensitivity was 0.546 (Table 2). Within EHR data, patients identified by a cancer record in the problem list and/or medical history section had moderate agreement with registry data (kappa = 0.524), and those solely identified from encounter diagnoses showed slight agreement (kappa = 0.063). Cancer history was more accurately recorded in the EHR for patients with an assigned primary care provider, those with higher visit rates, and those established longer with their CHC (Table 2).

In total, we identified 90 121 patient*site combinations using either source; 60 415 eligible cancer patient*sites were identified in the EHR data, 56 732 patient*sites from the registries, and 27 026 patient*sites (30.0%) were coascertained in both sources. Multiple cancer sites were documented for some patients using either source: 13.7% of patients had 2, and 2.8% had 3 and more sites. Occurrence of multiple cancer sites was similar within EHR and registry data (not shown).

The percentage of cases in both data sources varied substantially by cancer site (Figure 2). Prostate cancer was the site most

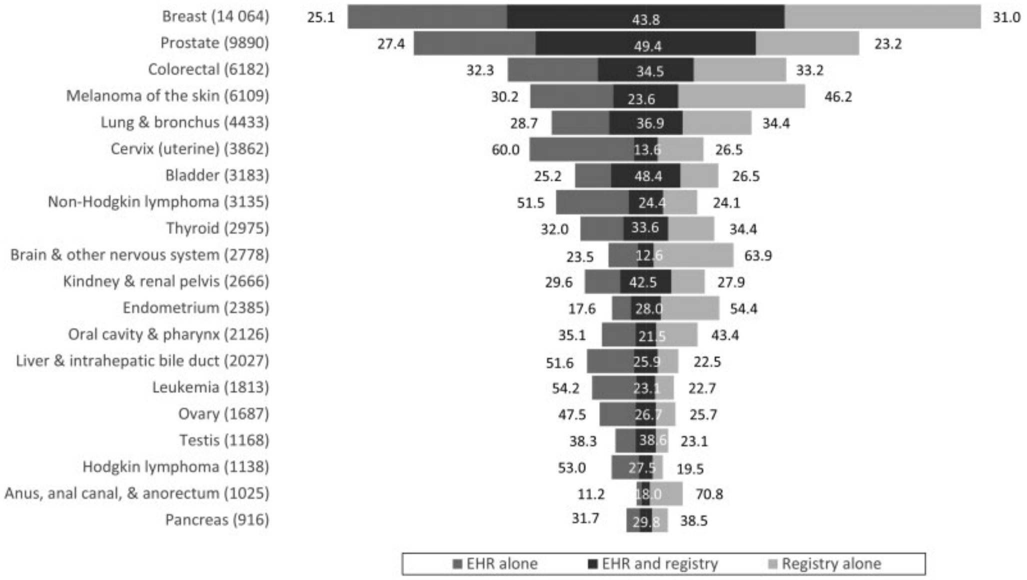

Figure 2. Percentage of leading cancers identified by source of ascertainment. Total number of cases ascertained from either source presented in parentheses. Width of bars is proportional to this combined case count. EHR = electronic health record.

Table 1. Cross-tabulation of patient-level cancer ascertainment by EHR and cancer registries, overall and by state[a]

	Cancer in registry		
Cancer history in EHR data	Yes	No	Total
3 states combined, No. (%)			
Yes	28 361 (1.9)	22 730 (1.5)	51 091 (3.4)
No	23 616 (1.6)	1 437 605 (95.1)	1 461 221 (96.6)
Total	51 977 (3.4)	1 460 335 (96.6)	1 512 312 (100)
California, No. (%)			
Yes	11 012 (1.4)	9972 (1.3)	20 984 (2.7)
No	11 749 (1.5)	737 154 (95.7)	748 903 (97.3)
Total	22 761 (3.0)	747 126 (97.0)	769 887 (100)
Oregon, No. (%)			
Yes	14 499 (2.6)	11 107 (2.0)	25 606 (4.6)
No	8894 (1.6)	520 162 (93.8)	529 056 (95.4)
Total	23 393 (4.2)	531 269 (95.8)	554 662 (100)
Washington, No. (%)			
Yes	2850 (1.4)	2114 (1.1)	4964 (2.5)
No	2973 (1.5)	190 993 (96.0)	193 966 (97.5)
Total	5823 (2.9)	193 107 (97.1)	198 930 (100)

[a] Counts are distinct patients with any cancer. Some cancers were identified by multiple registries, so the single-state EHR cancer and overall patient count values do not sum to the EHR total. EHR = electronic health record.

often identified by both the registry and EHR: 49.4% of patient*-sites had matching records. Overlap was also relatively high for female breast (43.8%), bladder (48.4%), and kidney or renal pelvis (42.5%) cancers. Overlap was low for cervix (13.6%), brain and nervous system (12.6%), and oral cavity (21.5%) cancers. Additionally, nonmatching cervical cancers were more often identified in EHR than cancer registry data (60.0% vs 26.5%).

Prostate, bladder, and female breast cancers had substantial agreement (kappa > 0.60), and cervix and brain or central nervous system cancers had fair agreement (kappa < 0.30; Table 3). Some of the most common cancers had the highest agreement and sensitivity, but cancer prevalence was not consistently related to agreement or sensitivity.

We compared distributions of diagnosis year and age at diagnosis for matched patient*sites appearing in both EHR and registry data (Figure 3). Overall distributions were similar, but EHR diagnosis dates skewed later in time, shifting the observed age at diagnosis distribution slightly older. According to cancer registries, cancers were diagnosed on average 3.2 years earlier (SD = 5.2 years) and at a younger age than was documented in the EHR (mean age at diagnosis: registry = 57.6 years [SD = 15.0 years]; EHR = 60.9 years [SD = 14.9 years]).

Table 4 demonstrates observed population characteristics of cancer patients if one were to use each data source individually. Age, sex, and federal poverty level distributions were similar (ASMD < 0.10). Those with cancer history identified in EHR data

Table 2. Agreement[a] of EHR and registry cancer ascertainment: full study, by state, EHR source, and CHC use

Sample	Kappa	Strength of agreement	Sensitivity	Specificity	Positive predictive value
Three states combined	0.535	Moderate	0.546	0.984	0.555
California	0.489	Moderate	0.484	0.987	0.525
Oregon	0.573	Moderate	0.620	0.979	0.566
Washington	0.515	Moderate	0.489	0.989	0.574
EHR[b]: problem list and/or medical history	0.524	Moderate	0.508	0.987	0.576
EHR[b]: encounter only	0.063	Slight	0.038	0.998	0.376
PCP assigned	0.5643	Moderate	0.611	0.981	0.554
PCP not assigned	0.3735	Fair	0.292	0.994	0.566
Years established[c]: highest quartile	0.6077	Substantial	0.666	0.977	0.594
Years established[c]: lowest quartile	0.3513	Fair	0.343	0.991	0.381
Overall visit rate[d]: highest quartile	0.5668	Moderate	0.738	0.972	0.484
Overall visit rate[d]: lowest quartile	0.5181	Moderate	0.449	0.990	0.660

[a] Cancer registry is gold standard. CHC = community health center; EHR = electronic health record; PCP = primary care provider.
[b] Refers to section(s) of EHR in which patient's cancer history was documented.
[c] Time between patient's earliest and most recent encounter.
[d] Based on ambulatory visits in 2016-2018.

tended to have more comorbidities than the registry-identified group (66.3% vs 55.4% had Charlson Comorbidity Index score ≥1, ASMD = 0.23). A slightly higher percentage of patients in EHR data reported Hispanic ethnicity (ASMD = 0.12) or Spanish language preference (ASMD = 0.13) compared with the registries. Patients with cancer history in the registries were more commonly uninsured than those found in EHR data (23.8% vs 19.4%, respectively), and a greater proportion of EHR-recorded patients had Medicare (39.5% vs 37.2%, ASMD = 0.10). Visit rates during 2016-2018 were slightly higher for patients with cancer history identified in EHR compared with the cancer registry data only (3.9 vs 3.5, ASMD = 0.35).

The subanalysis limited to more recently diagnosed patients showed greater overlap (Supplementary Figure 1, available online) and higher agreement (Supplementary Table 1, available online) across all cancer sites. Among this subpopulation, all cancer sites except soft tissue or heart, cervix, and brain or nervous system showed moderate or substantial agreement. Observed characteristics of cancer survivors by EHR and registry ascertainment were similar to those observed in the primary analysis (Supplementary Table 2, available online).

Discussion

We linked EHR data from a large network of CHCs to 3 state cancer registries and found nearly one-half of cancer cases recorded in the registries were "missing" from EHR data on matched patients. Primary care providers have previously highlighted issues of inconsistent and incomplete documentation (28,29). There are multiple explanations for these discrepancies, including insufficient data exchange between health systems and care providers, the lack of a systematic workflow to capture this information in primary care, insufficient time for documentation, and incomplete or inaccurate patient understanding or recall. For example, we found the EHR recorded more cases of cervical cancer than the registries. This may represent in situ cervical cancers that are not reportable to these registries and some cases resulting from patient misunderstanding of abnormal cervical cancer screening follow-up.

Overall, we found moderate levels of statistical agreement between EHR and state cancer registries on whether patients had any cancer history, with some variation by cancer site and length of time since cancer diagnosis. Multiple reasons likely contribute to these differential results. In the EHR, dates of diagnosis may be inaccurately recorded as entry date or encounter date. Some cancer history found only in the EHR may represent cancers diagnosed in another state or cancers not meeting registry inclusion criteria (eg, benign tumors, tumors with uncertain histologic behavior). Further, cancer diagnostic codes may appear in EHR records because they are associated with orders for screening (eg, breast cancer when mammography is ordered) but are not necessarily indicative of a cancer diagnosis. Our data showed low agreement when EHR cancer information originated from encounter diagnoses without being noted on the problem list, supporting this idea. Developing cancer-specific EHR flowsheets to systematically capture details about cancer diagnosis and treatment would improve accuracy and completeness. More research is needed to better understand the reasons for missing data and to assess the impact of this missing cancer information on care receipt and cancer survivor outcomes.

Despite the finding that nearly one-half of total cancers were unobserved in either data source alone, the patient populations with cancer history identified by each source independently had similar demographic and clinical profiles. This suggests that population-level research on patients with a history of cancer as noted in the EHR may be reasonably representative of the "true" survivor population with regard to important covariates. However, research that relies on EHR data to assess outcomes or receipt of guideline-concordant care for cancer survivors may be affected by this missing documentation, making it likely that adequate preventive care is hindered by missing cancer history. Implementation and effectiveness trials designed to improve cancer survivor care in primary care settings may underestimate the expected population reach of their intervention by as much as 50%, but accuracy may be improved by limiting populations to certain cancer sites (eg, prostate, bladder, and breast), patients with more recent diagnoses, and/or those with higher levels of use and longer contact with their primary care clinics.

Our findings highlight the need for strategies to improve the accuracy of cancer history documentation in primary care settings. Patients with a history of cancer have excess morbidity (1,11) and long-term care management needs following their diagnosis and treatment. Improved data exchange between oncology and primary care (30), and emerging potential for direct

Table 3. Site-specific agreement[a] of electronic health record and registry cancer cases

Cancer site	Cases coascertained, No.	Kappa	Strength of agreement	Sensitivity	Specificity	Positive predictive value
Prostate (male)	4890	0.658	Substantial	0.681	0.9958	0.644
Bladder	1538	0.651	Substantial	0.646	0.9995	0.658
Breast (female)	6139	0.606	Substantial	0.586	0.9959	0.637
Kidney and renal pelvis	1125	0.598	Moderate	0.605	0.9995	0.592
Testis (male)	451	0.560	Moderate	0.626	0.9993	0.508
Lung and bronchus	1635	0.538	Moderate	0.517	0.9992	0.563
Larynx	163	0.521	Moderate	0.463	0.9999	0.595
Esophagus	185	0.518	Moderate	0.573	0.9999	0.473
Colon and rectum	2135	0.512	Moderate	0.511	0.9987	0.517
Thyroid	998	0.503	Moderate	0.494	0.9994	0.514
Pancreas	273	0.460	Moderate	0.437	0.9998	0.487
Endometrium (female)	668	0.437	Moderate	0.340	0.9995	0.615
Hodgkin lymphoma	310	0.432	Moderate	0.585	0.9996	0.343
Ovary (female)	450	0.422	Moderate	0.510	0.9991	0.361
Liver and intrahepatic bile duct	524	0.412	Moderate	0.537	0.9993	0.335
Myeloma	151	0.396	Fair	0.549	0.9998	0.310
Non-Hodgkin lymphoma	763	0.393	Fair	0.503	0.9989	0.323
Melanoma of the skin	1443	0.381	Fair	0.339	0.9988	0.439
Stomach	157	0.375	Fair	0.318	0.9999	0.458
Leukemia	381	0.371	Fair	0.496	0.9994	0.297
Oral cavity and pharynx	457	0.353	Fair	0.331	0.9995	0.380
Soft tissue including heart	145	0.353	Fair	0.395	0.9998	0.320
Vulva (female)	125	0.316	Fair	0.213	0.9999	0.610
Anus, anal canal, and anorectum	184	0.304	Fair	0.202	0.9999	0.617
Cervix uteri (female)	523	0.237	Fair	0.339	0.9973	0.184
Brain and other nervous system	328	0.218	Fair	0.159	0.9996	0.350
Other	746	0.107	Slight	0.162	0.9946	0.084

[a] Cancer registry is gold standard.

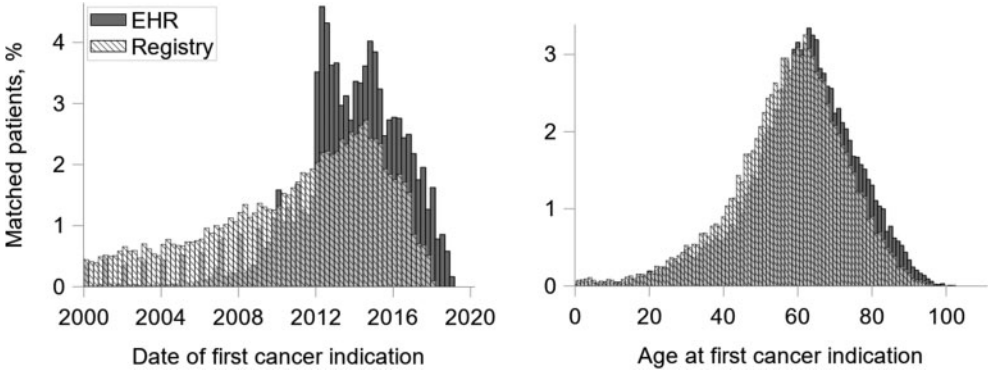

Figure 3. Distribution of year of diagnosis and age at diagnosis among matched cases, electronic health record (EHR) vs cancer registry. Comparisons include matched patient*sites (same cancer site and patient in both EHR and registry data), n = 27 026. We imputed mid-year for dates where only the year was known and mid-month if day was unknown.

linkages and data feeds with centralized cancer registries (5,31) could help track outcomes and lead to better care for cancer survivors and enhanced research potential. Primary care CHCs typically do not provide cancer treatment or immediate follow-up care and are thus unlikely to contain comprehensive cancer-related data. But they do provide longitudinal care to cancer survivors, and EHRs contain rich clinical (eg, comorbidity, screening history) and demographic (eg, insurance status, social determinants of health) data not found in cancer registries. CHCs are also key front-line providers in the prevention and early detection of cancer, providing access to traditionally underserved populations, and should be partners in epidemiologic and outcomes research across the cancer continuum.

We note several limitations. First, probabilistic linkage methods are imperfect, and some patients with records in both the EHR and cancer registry may not have matched using our algorithms. Second, among the cancers identified in the EHR only, we could not determine which were misclassified (not actual cancer diagnoses), true cancers that did not meet registry inclusion criteria, or true malignancies that were not reported to the registry. Future work could consider using EHRs for additional case-finding by centralized cancer registries. Some EHR-only cancers may have stemmed from diagnosis codes that were used as rule-out diagnoses for a screening test, although this scenario is unlikely to represent a large number of cases because the majority of cancer cases

Table 4. Prevalence and characteristics[a] of patients with cancer history according to EHR and cancer registry, stratified by data source

	Patients with a history of cancer		
Patient characteristics	Documented in EHR, % (n = 51 091)	Documented in cancer registry, % (n = 51 977)	ASMD
Age on January 1, 2019, y			0.074
18-29	1.8	1.8	
30-39	4.9	4.3	
40-49	8.6	7.2	
50-59	18.2	17.3	
60-69	29.7	29.3	
70-79	21.5	22.6	
≥80	15.4	17.4	
Sex			0.020
Female	56.0	55.5	
Male	44.0	44.5	
Race or ethnicity			0.119[b]
Hispanic	13.4	12.2	
Non-Hispanic White	72.9	71.2	
Non-Hispanic Black	3.4	4.6	
Non-Hispanic Asian	2.8	4.2	
Other and unknown	7.6	7.9	
Preferred language			0.131[b]
English	84.7	84.2	
Spanish	9.1	7.3	
Other	6.2	8.6	
Charlson Comorbidity Index score (excluding cancer component)			0.229[b]
0	33.7	44.6	
1	18.0	15.3	
2-3	22.0	18.6	
4-6	17.8	14.6	
≥7	8.5	6.9	
Federal poverty level			0.085
≤138%	41.5	39.3	
≥139%	16.1	14.0	
Missing or unknown	42.5	46.8	
Primary payer type			0.100[b]
Medicaid	25.1	23.5	
Medicare	39.5	37.2	
Private	14.8	14.8	
Uninsured	19.4	23.8	
Other or missing	1.3	0.7	
Primary care provider assigned			0.271[b]
Yes	89.2	79.4	
No	10.8	20.6	
Average annual encounter rate, 2016-2018			0.352[b]
No visits in 2016-2018	38.9	56.1	
<1/y	12.3	11.9	
1-2/y	15.9	11.5	
2-5/y	18.8	12.0	
5-10/y	10.0	5.9	
>10/y	4.1	2.6	
3-Year encounter rate, 2016-2018, mean (SD)	3.9 (7.2)	3.5 (6.8)	0.065

[a] All characteristics obtained from EHR data. Time-varying characteristics assigned as of latest encounter date, unless otherwise specified. ASMD = absolute standardized mean difference; EHR = electronic health record.
[b] ASMD greater than 0.1 indicates meaningful distributional difference.

ascertained from EHR data were recorded in the problem list or medical history sections. Another limitation is the broad date range and limited inclusion criteria we imposed on the selection of EHR patient records for linkage and cases for analysis; for example, there was no restriction on the timing of EHR visits relative to cancer diagnosis. We took this approach to describe the totality of EHR-recorded cancer data, but subanalyses suggest that limiting to patients with higher CHC use and/or more recently diagnosed results in greater agreement. Such restrictions should be considered to strengthen cancer research conducted using primary care EHR data. Lastly, we did not study clinic or provider characteristics and their impact on agreement; future studies could evaluate this relationship.

Nearly one-half of cancer survivors of CHCs in California, Oregon, and Washington did not have their cancer history documented in the EHR. Agreement between EHR and cancer registry data was moderate and varied by cancer site, length of time since diagnosis, and use, yet demographics were similar. These findings suggest the need for strategies to improve the accuracy and completeness of cancer history data in CHC EHRs to ensure adequate delivery of care and optimal health outcomes for cancer survivors. While caution is warranted when designing cancer-related research using outpatient EHR data, our results highlight several ways in which reliability may be improved.

Funding

This work was supported by the National Cancer Institute (NCI) of the National Institutes of Health (grant numbers R01CA204267 and P50CA244289). This P50 program was launched by NCI as part of the Cancer Moonshot.

Notes

Role of the funder: The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Disclosures: All authors have no conflict of interest or disclosures to report.

Prior presentations: An earlier version of this work was presented at the North American Primary Care Research Group's annual meeting in Chicago, IL; November, 2018.

Disclaimers: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgements: This research article is based in part on cancer data provided by the California Cancer Registry, the Oregon State Cancer Registry (OSCaR), and the Washington State Cancer Registry (WSCR), participating registries funded by the National Program of Cancer Registries (NPCR) under the Centers for Disease Control and Prevention (CDC). We wish to thank Julianne Cooley with California Cancer Reporting and Epidemiologic Surveillance (CalCARES), Todd Beran with OSCaR, and Mahesh Keitheri Cherteri with WSCR for their assistance with data linkage and file preparation. We acknowledge the contributions of Nate Warren with OCHIN, who assisted with references and manuscript preparation and submission.

Author contributions: Megan Hoopes: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, writing—original draft, writing—review and editing; Robert Voss: data curation, formal analysis, writing—original draft, writing—review and editing; Heather Angier: funding acquisition, writing—original draft, writing—review and editing; Miguel Marino: conceptualization, data curation, formal analysis, writing—original draft, writing—review and editing; Teresa Schmidt: data curation, formal analysis, writing—original draft, writing—review and editing; Jennifer E. DeVoe: funding acquisition, writing—original draft, writing—review and editing; Jeffrey Soule: methodology, resources, writing—original draft, writing—review and editing; Nathalie Huguet: conceptualization, funding acquisition, writing—original draft, writing—review and editing.

Data Availability

Raw data underlying this article were generated from multiple agencies and institutions; restrictions apply to the availability and rerelease of data under cross-institution agreements. Data are however available from the authors upon reasonable request and with permission of all relevant parties.

References

1. Leach CR, Weaver KE, Aziz NM, et al. The complex health profile of long-term cancer survivors: prevalence and predictors of comorbid conditions. *J Cancer Surviv*. 2015;9(2):239–251.
2. Institute of Medicine and National Research Council. *From Cancer Patient to Cancer Survivor: Lost in Transition*. Washington, DC: National Academies Press; 2006.
3. Pollack LA, Adamache W, Ryerson AB, Eheman CR, Richardson LC. Care of long-term cancer survivors: physicians seen by Medicare enrollees surviving longer than 5 years. *Cancer*. 2009;115(22):5284–5295.
4. Roy S, Vallepu S, Barrios C, Hunter K. Comparison of comorbid conditions between cancer survivors and age-matched patients without cancer. *J Clin Med Res*. 2018;10(12):911–919.
5. Krist AH, Beasley JW, Crosson JC, et al. Electronic health record functionality needed to better support primary care. *J Am Med Inform Assoc*. 2014;21(5):764–771.
6. Sollie A, Roskam J, Sijmons RH, Numans ME, Helsper CW. Do GPs know their patients with cancer? Assessing the quality of cancer registration in Dutch primary care: a cross-sectional validation study. *BMJ Open*. 2016;6(9):e012669.
7. Zahnd WE, James AS, Jenkins WD, et al. *Rural-Urban Differences in Cancer Incidence and Trends in the United States*. Cancer Epidmiol Biomarkers Prev. 2018;27(11):1265–1274.
8. Olaku OO, Taylor EA. Cancer in the medically underserved population. *Prim Care*. 2017;44(1):87–97.
9. Fernandez LM, Becker JA. Women's select health issues in underserved populations. *Prim Care*. 2017;44(1):47–55.
10. Centers for Disease Control and Prevention. BRFSS Prevalence and Trends Data. https://www.cdc.gov/brfss/brfssprevalence. 2015. Accessed April 18, 2020.
11. Hoopes M, Schmidt T, Huguet N, et al. Identifying and characterizing cancer survivors in the US primary care safety net. *Cancer*. 2019;125(19):3448–3456.
12. Dregan A, Moller H, Murray-Thomas T, Gulliford M. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. *Cancer Epidemiol*. 2012;36(5):425–429.
13. Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiol Drug Saf*. 2013;22(2):168–175.
14. Thompson CA, Jin A, Luft HS, et al. Population-Based Registry Linkages to Improve Validity of Electronic Health Record-Based Cancer Research. Cancer Epidmiol Biomarkers Prev. 2020;29(4):796–806.
15. Tucker TC, Durbin EB, McDowell JK, Huang B. Unlocking the potential of population-based cancer registries. *Cancer*. 2019;125(21):3729–3737.
16. North American Association of Central Cancer Registries (NAACCR). https://www.naaccr.org/about-naaccr. Accessed April 20, 2020.
17. Petrik AF, Green BB, Vollmer WM, et al. The validation of electronic health records in accurately identifying patients eligible for colorectal cancer screening in safety net clinics. *Fam Pract*. 2016;33(6):639–643.
18. Bailey SR, Heintzman JD, Marino M, et al. Measuring preventive care delivery: comparing rates across three data sources. *Am J Prev Med*. 2016;51(5):752–761.
19. Centers for Disease Control and Prevention. Link Plus I Registry Plus. https://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm. Accessed March 31, 2020.
20. National Cancer Institute. Download MatchPro Software. https://surveillance.cancer.gov/matchpro/download. Accessed March 31, 2020.
21. Surveillance, Epidemiology, and End Results. SEER ICD-O-3 Coding Materials. https://seer.cancer.gov/icd-o-3/. Accessed March 31, 2020.
22. ICD Conversion Programs. https://seer.cancer.gov/tools/conversion. Accessed March 31, 2020.
23. Cunningham M. More than just the kappa coefficient: a program to fully characterize inter-rater reliability between two raters. Paper presented at SAS Global Forum; March 22-25, 2009; Washington, DC.

24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
25. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083–3107.
26. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat-Simul Comput*. 2009;38(6):1228–1234.
27. Charlson ME, Charlson RE, Peterson JC, Marinopoulos SS, Briggs WM, Hollenberg JP. The Charlson comorbidity index is adapted to predict costs of chronic disease in primary care patients. *J Clin Epidemiol*. 2008;61(12):1234–1240.
28. Hudson SV, Miller SM, Hemler J, et al. Cancer survivors and the patient-centered medical home. *Behav Med Pract Policy Res*. 2012;2(3):322–331.
29. Rubinstein EB, Miller WL, Hudson SV, et al. Cancer survivorship care in advanced primary care practices: a qualitative study of challenges and opportunities. *JAMA Intern Med*. 2017;177(12):1726–1732.
30. Tsui J, Howard J, O'Malley D, et al. Understanding primary care-oncology relationships within a changing healthcare environment. *BMC Fam Pract*. 2019;20(1):164.
31. Hiatt RA, Tai CG, Blayney DW, et al. Leveraging state cancer registries to measure and improve the quality of cancer care: a potential strategy for California and beyond. *JNCI*. 2015;107(5):djv047.