3-29-2022

Developing an algorithm across integrated healthcare systems to identify a history of cancer using electronic medical records

Jennifer C. Gander

Mahesh Maiyani

Larissa L. White

Andrew T. Sterrett

Brianna Güney

See next page for additional authors

Authors

Jennifer C. Gander, Mahesh Maiyani, Larissa L. White, Andrew T. Sterrett, Brianna Güney, Pamala A. Pawloski, Teri DeFor, YuanYuan Olsen, Benjamin A. Rybicki, Christine Neslund-Dudas, Darsheen Sheth, Richard Krajenta, Devaki Purushothaman, Stacey Honda, Cyndee Yonehara, Katrina A. B. Goddard, Yolanda K. Prado, Habibul Ahsan, Muhammad G. Kibriya, Briseis Aschebrook-Kilfoy, Chun-Hung Chan, Sarah Hague, Christina L. Clarke, Brooke Thompson, Jennifer Sawyer, Mia M. Gaudet, and Heather Spencer Feigelson

Journal of the American Medical Informatics Association, 00(0), 2022, 1–8
https://doi.org/10.1093/jamia/ocac044
Research and Applications

Research and Applications

Developing an algorithm across integrated healthcare systems to identify a history of cancer using electronic medical records

Jennifer C. Gander[1], Mahesh Maiyani[2], Larissa L. White[2], Andrew T. Sterrett[2], Brianna Güney[1], Pamala A. Pawloski[3], Teri DeFor[3], YuanYuan Olsen[3], Benjamin A. Rybicki[4], Christine Neslund-Dudas[4], Darsheen Sheth[4], Richard Krajenta[4], Devaki Purushothaman[4], Stacey Honda[5,6], Cyndee Yonehara[5], Katrina A.B. Goddard[7], Yolanda K. Prado[7], Habibul Ahsan[8], Muhammad G. Kibriya[8], Briseis Aschebrook-Kilfoy[8], Chun-Hung Chan[9], Sarah Hague[9], Christina L. Clarke[2], Brooke Thompson[2], Jennifer Sawyer[2], Mia M. Gaudet[10], and Heather Spencer Feigelson[2]

[1]Center for Research and Evaluation, Kaiser Permanente Georgia, Atlanta, Georgia, USA, [2]Institute for Health Research, Kaiser Permanente Colorado, Aurora, Colorado, USA, [3]HealthPartners Institute, Bloomington, Minnesota, USA, [4]Department of Public Health Sciences, Henry Ford Health System, Detroit, Michigan, USA, [5]Center for Integrated Healthcare, Kaiser Permanente Hawaii, Honolulu, Hawaii, USA, [6]Hawaii Permanente Medical Group, Kaiser Permanente Hawaii, Honolulu, Hawaii, USA, [7]Department of Translational and Applied Genomics, Center for Health Research, Kaiser Permanente Northwest, Portland, Oregon, USA, [8]Institute for Population and Precision Health, University of Chicago, Chicago, Illinois, USA, [9]Sanford Research, Sanford Health, Sioux Falls, South Dakota, USA, and [10]Trans Divisional Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA

Corresponding Author: Jennifer C. Gander, PhD, Center for Research and Evaluation, Kaiser Permanente Georgia, 1375 Peachtree Rd NE, Atlanta, GA 30309, USA; jennifer.c.gander@kp.org

Received 27 October 2021; Revised 5 February 2022; Editorial Decision 9 March 2022; Accepted 16 March 2022

ABSTRACT

Objective: Tumor registries in integrated healthcare systems (IHCS) have high precision for identifying incident cancer but often miss recently diagnosed cancers or those diagnosed outside of the IHCS. We developed an algorithm using the electronic medical record (EMR) to identify people with a history of cancer not captured in the tumor registry to identify adults, aged 40–65 years, with no history of cancer.

Materials and Methods: The algorithm was developed at Kaiser Permanente Colorado, and then applied to 7 other IHCS. We included tumor registry data, diagnosis and procedure codes, chemotherapy files, oncology encounters, and revenue data to develop the algorithm. Each IHCS adapted the algorithm to their EMR data and calculated sensitivity and specificity to evaluate the algorithm's performance after iterative chart review.

Results: We included data from over 1.26 million eligible people across 8 IHCS; 55 601 (4.4%) were in a tumor registry, and 44848 (3.5%) had a reported cancer not captured in a registry. The common attributes of the final algorithm at each site were diagnosis and procedure codes. The sensitivity of the algorithm at each IHCS was 90.65%–100%, and the specificity was 87.91%–100%.

Discussion: Relying only on tumor registry data would miss nearly half of the identified cancers. Our algorithm was robust and required only minor modifications to adapt to other EMR systems.

Conclusion: This algorithm can identify cancer cases regardless of when the diagnosis occurred and may be useful for a variety of research applications or quality improvement projects around cancer care.

Key words: electronic health records, cancer, algorithm

INTRODUCTION

Longitudinal cohort studies have been used for decades to understand the etiology of chronic disease and have made important contributions to public health.[1–4] Many cohort studies have drawn from specific populations, such as teachers,[5,6] agricultural workers,[7] or healthcare providers[8–10] and are designed to study a wide range of outcomes. Currently, the All of Us research program is assembling a cohort across the United States designed to study a variety of disease outcomes.[11] Similar to other cohorts built with a specific disease endpoint in mind, such as the Framingham Heart Study and the Cancer Prevention Studies,[3,4,12,13] we are developing a new prospective cohort, Connect for Cancer Prevention Study (https://dceg.cancer.gov/research/who-we-study/cohorts/connect).

In collaboration with the National Cancer Institute, Connect aims to inform new approaches in precision prevention and early detection of cancer.[14] This prospective cohort will incorporate recent developments in digital technologies, biomarkers, and exposure assessments to advance the study of suspected and emerging factors that influence cancer development, treatment, and survival. The cohort aims to enroll 200 000 adults aged 40–65 years who are members or patients of a participating integrated healthcare systems (IHCS) across the United States with no personal history of invasive cancer.

Tumor registries are considered the "gold standard" for identifying incident cancer within a population.[15–18] However, they usually do not include recently diagnosed cases, as fully annotated data can lag up to 2 years. Further, IHCS tumor registries usually do not track cases diagnosed and/or treated outside of the health system. Utilizing both tumor registries and large electronic medical record (EMR) databases can provide a more complete capture of all cancer cases, regardless of when or where they were diagnosed. Previously, we presented an algorithm that was developed to identify a history of cancer within a single IHCS.[19] This original algorithm was developed and tested within Kaiser Permanente Colorado (KPCO) and contained 2013 administrative data utilizing diagnosis codes, chemotherapy treatment codes, oncology encounters, and the KPCO tumor registry to identify a member's history of cancer. Our current works build off the Clarke and Feigelson publication through external validation of the algorithm at multiple IHCS and use more updated diagnosis, treatment, and procedures codes. We aimed to develop an algorithm with a high specificity that leverages both our EMR systems and our local tumor registries to identify cancer-free adults who are eligible to participate in Connect. We then adapted and tested the algorithm for use at 7 other IHCS.

MATERIALS AND METHODS

Setting

The algorithm was developed at KPCO; 7 other IHCS adapted and validated the algorithm. Each organization maintains an Epic-based EMR for each health plan member and/or patients aligned with primary care and specialty providers. KPCO serves approximately 550 000 members in the greater Denver metropolitan area. Kaiser Permanente Georgia (KPGA) has clinics located around the Atlanta, Georgia metropolitan and Northern Georgia area with over 300 000 members. Kaiser Permanente Hawaii (KPHI) serves the entire state of Hawaii including more than 250 000 members with facilities on 4 islands. Kaiser Permanente Northwest (KPNW) provides medical coverage for over 600 000 members in northwest Oregon and southwest Washington. HealthPartners (HP) Institute is the largest consumer-governed nonprofit healthcare organization in the United States, serving more than 1.8 million medical and dental health plan members nationwide; more than 1.2 million of these patients are in Minnesota and western Wisconsin. Henry Ford Health System (HFHS) serves southeast Michigan including Detroit with over 450 000 aligned patients and/or health plan members. University of Chicago is an academic medical center located in Hyde Park on the South Side of Chicago that serves approximately 375 000 patients. Sanford Health is a nonprofit health system located in the upper Midwest serving over 2 million patients across a 250 000 square mile rural catchment area spanning South Dakota, North Dakota, and western Minnesota.

The EMR systems maintained at each site contain information on diagnoses, medical procedures including biopsies, laboratory and imaging tests, pharmaceutical orders, and clinical encounters and encounter types. The robust information available in the EMR enables each site to identify cancer diagnostic codes, blood or radiology orders, chemotherapy treatment sessions, and oncology department visits.[20–22] KPCO's EMR was established in 1998 and includes tumor registry data beginning in 2000. The EMRs from the other participating 7 IHCS were established between 1996 and 2006 with tumor registries created between 1974 and 2010. This project was reviewed and approved by the Institutional Review Board at each participating IHCS, and the requirement for informed consent was waived.

Algorithm initial development

The algorithm was adapted at KPCO from an earlier version based on International Classification of Diseases Ninth Edition (ICD-9) codes,[19] updated to include contemporary codes and to match the criteria for the new cohort. Specifically, we included ICD-9 and International Classification of Diseases Tenth Edition (ICD-10), as well as International Classification of Diseases for Oncology Third Edition (ICD-O-3) coding, and attempted to identify both *in situ* (stage 0) and invasive disease (stages I–IV). Most invasive cancer diagnoses are included in tumor registries which follow the standards set forth by the North American Association of Central Cancer Registries (NAACCR);[23] however, most *in situ* cases, with the exception of *in situ* breast cancers, are not routinely recorded. Nonmelanoma skin cancers are not considered reportable cancers and thus, not included in tumor registries.

To develop and test the algorithm, KPCO used 2016 data to ensure the tumor registry, 'gold standard' data were complete at each IHCS participating in this study. The algorithm matched the Connect for Cancer Prevention Study eligibility criteria and identified all KPCO members aged 40–65 years, and flagged invasive cancers using EMR and tumor registry records with an invasive cancer behavior code. Records with a behavior code identifying *in situ* disease, and nonmelanoma skin cancers, were not flagged as having a history

of cancer, per the eligibility criteria. The developed algorithm initially flagged any record with an inpatient or outpatient claim with an ICD-9, ICD-10, or ICD-O-3 code indicating cancer. The algorithm included the following codes for incident cancer cases, ICD-9: 140-209; ICD-10: C00-C97; and the following codes for the history of cancer cases, ICD-9: V10 and ICD-10: Z85. The algorithm excluded ICD-9: 173, 209.4, 209.5, 209.6, V10.83, or V13.89; ICD-10: C44, Z85.828 or Z85.821. ICD-O-3 procedural codes were also included: C000–C999 with a behavior code of 3 or greater except C440–C449 with histology codes of 8051–8098. The algorithm also flagged any record with at least 3 visits in the oncology department on separate days, or at least 2 records of receiving a chemotherapeutic drug on separate days (see a complete list of codes in Supplementary Table S1). The developed algorithm was then tested using manual chart review and revised based on the chart review findings until the algorithm met the stated goals. SAS version 9.4 (SAS Institute, Cary, NC, USA) was used to develop and disseminate the algorithm.

Chart review

The developed algorithm is intended to go beyond the tumor registry's data and account for limitations in the tumor registry which may include cancer cases that were (1) not reportable cases because they were stage 0; (2) recently diagnosed and not yet captured in the tumor registry; or (3) diagnosed outside the health plan or prior to health plan membership. Chart review was specific to records not reported in the tumor registry. To ensure the algorithm was accurately identifying members without cancer, we selected a random sample of 100 members not in the tumor registry according to the following stratification parameters: (a) cancer status as identified by the algorithm (1:4 cancer vs not cancer; 20 flagged with cancer vs 80 flagged as cancer-free) and (b) age group (1:2 age 40–55 years vs >55 years) because cancer is more common in older adults. Age group stratification was applied in each cancer status subgroup, such that 14 of 20 members the algorithm flagged as having cancer were over age 55. Chart review was performed on each individual to indicate an ever diagnosis of cancer before or during 2016.

For internal validation, we tested the algorithm on 100 chart review cases at KPCO. Each chart was fully reviewed by trained research staff to find any documentation of cancer, or to confirm there was no history of cancer using all notes in the chart dating back to the patient's first enrollment, or to the beginning of the EMR (1998). For the 20 patients, the algorithm identified as having a history of cancer, the abstractor reviewed the EMR at the exact date of diagnosis; if a diagnosis for cancer was not found, the reviewer then examined the chart from the administrative diagnosis date in the EMR going forward in time for a mention of cancer. For the 80 patients, the algorithm identified as not having a history of cancer, the abstractor examined the patient's full EMR dating back to the first enrollment through 2016. The algorithm's performance was based on a goal to achieve ≥80% sensitivity and ≥80% specificity to detect any cancer.[24] Using chart review batches of 25, we iteratively reviewed the errors revealed by the chart review to modify the algorithm until the desired sensitivity and specificity were achieved. When the algorithm identified a patient as having cancer and chart abstraction found no cancer (false positive) or the algorithm flagged a patient as cancer-free and chart abstraction reported a history of cancer (false negative), the IHCS study team met to determine any pattern to the false positives or negatives that could be improved upon. After each iteration, a new chart review sample was selected, as described above, and sensitivity and specificity were recalculated.

External validation

After the desired sensitivity and specificity were achieved at KPCO, the algorithm was disseminated to the remaining 7 participating IHCS. To aid in external validation, KPCO developed and refined a chart abstraction guide that was disseminated to each IHCS. Each IHCS modified the algorithm to fit their IHCS' EMR and began chart review as described above. While it was recommended that each IHCS complete a minimum of 100 chart reviews to test their algorithm, each site determined how many chart reviews to complete. KPGA, KPHI, and KPNW maintained the KPCO-developed chart abstraction and completed 100 chart reviews in 25-batch iterations. KPNW completed 99 chart reviews, instead of 100, due to discovering 1 member randomly selected for chart review was deceased. Three sites slightly modified the procedures to better match their data collection and EMR system. HP used 2020 data and included all patients age 40–65 years that completed at least 1 primary care office visit or at least 1 telehealth visit in internal medicine, family practice, or obstetrics in the past 5 years at a HP-owned facility and excluded those already in the tumor registry. HFHS utilized the Health Care System Research Network Virtual Data Warehouse[25] 2013–2016 to identify 100 cases for chart review. Sanford Health was unable to pull their total eligible population and did not rely on a tumor registry; however, Sanford completed 188 chart reviews from a 2020 sample of 66 390 based on a sample size calculation with a 2% margin of error. University of Chicago's Center for Informatics Clinical Research Data Warehouse used 2020 data to modify the algorithm and identify 100 cases to review. The participating IHCS completed their chart review and modified the algorithm, as needed, to achieve the sensitivity and specificity goals.

Statistical analysis

Patient-level descriptive statistics were stratified based on inclusion in the tumor registry, identification of a history of cancer by the algorithm, or cancer-free. For HFHS, KPCO, KPGA, KPHI, and KPNW, we calculated the sensitivity and specificity of the final algorithm using the tumor registry, for each region. For these 5 IHCS, the overall sensitivity to detect cancer is a weighted average of the sensitivities, where the weight of each subgroup (patients in the tumor registry vs patients not in the tumor registry) is the proportion of all cancer cases in that subgroup. Figure 1 describes the components and calculation of the weighted sensitivity (H).

The remaining IHCS calculated the sensitivity of the algorithm based solely on their chart review results, Table 1 box (G).

Specificity was not affected by the distribution of cases and was calculated for all IHCS as:

$$\frac{\text{True Negatives : No. of cancer} - \text{free cases identified by algorithm and chart review}}{\text{Total Negatives from chart review}}.$$

RESULTS

Table 1 provides a description of the study population for each participating IHCS. Except for Sanford Health for reasons mentioned in the Materials and methods section, there was a total of 1 265 076 members (HP = 472 714; HFHS = 170 283; KPCO = 173 197; KPGA = 104 633; KPHI = 76 910; KPNW = 170 181; University of Chicago = 97 158), with 54.4% female and 45.6% male included across the IHCS's algorithm development (Table 1). Cancer-free individuals were 46.3% male, 63.5% White, and 5.7% Hispanic. University of Chicago had 9.2% of eligible members with a record

Among Eligible Population:	Percent of Cancer Cases	Sensitivity
A = Number of Cancer Cases identified in Tumor Registry	**D** = A / C	**F**= 100% Tumor Registry Sensitivity
B = Number of Cancer Cases identified by Algorithm	**E** = B / C	**G** = (Number of true positives from chart review) / (Number of all cancer cases identified by chart review)
C = A + B Total Number of Cancer Cases identified in Tumor Registry or by Algorithm		**H** = (D * F) + (E * G) **Weighted Sensitivity**

Figure 1. Formula description of the weighted sensitivities calculated for HFHS, KPCO, KPGA, KPHI, and KPNW to account for cancer cases identified in either the tumor registry or by the algorithm.

Table 1. Aggregated characteristics among the integrated healthcare systems participating in the project to develop and modify an algorithm to identify a history of cancer

	Total eligible population *N* (%)	No indication of cancer via tumor registry or cancer algorithm *N* (%)	Cancer identified by tumor registry *N* (%)	Cancer identified by algorithm, not in tumor registry *N* (%)
Integrated Healthcare System[a]	1 265 076	1 164 627 (92.14)	55 601 (4.40)	44 848 (3.55)
HealthPartners Institute[b]	472 714	442 618 (93.6)	17 072 (3.6)	13 024 (2.8)
Henry Ford Health Systems	170 283	157 818 (92.7)	7223 (4.2)	5242 (3.1)
Kaiser Permanente Colorado	173 197	161 185 (93.1)	5650 (3.3)	6362 (3.67)
Kaiser Permanente Georgia	104 633	96 960 (92.7)	3331 (3.2)	4342 (4.15)
Kaiser Permanente Hawaii	76 910	70 188 (91.3)	4502 (5.9)	2220 (2.89)
Kaiser Permanente Northwest	170 181	150 923 (88.7)	8846 (5.2)	10 412 (6.12)
University of Chicago	97 158	84 935 (87.4)	8977 (9.2)	3246 (3.38)
Sanford Health[c]	–	–	–	–
Patient demographics				
Age group				
40–44	251 083 (19.8)	243 037 (20.9)	3812 (6.9)	4149 (9.3)
45–49	234 764 (18.6)	223 179 (19.2)	5909 (10.6)	5639 (12.6)
50–54	245 322 (19.4)	227 900 (19.6)	9359 (16.8)	8053 (18.0)
55–59	258 404 (20.4)	233 357 (20.0)	14 105 (25.4)	10 955 (24.4)
60–65	275 503 (21.8)	237 154 (20.4)	22 407 (40.3)	16 052 (35.8)
Sex				
Male	576 824 (45.6)	539 568 (46.3)	21 413 (38.5)	15 691 (35.0)
Female	687 955 (54.4)	624 768 (53.6)	34 183 (61.5)	29 155 (65.0)
Other	–	–	–	–
Unknown	297 (0.02)	288 (0.02)	5 (0.01)	2 (0.00)
Race				
Asian	81 826 (6.5)	77 286 (6.6)	2946 (5.3)	1600 (3.6)
Black	183 693 (14.5)	170 847 (14.7)	8279 (14.9)	4581 (10.2)
Hawaiian/Pacific Islander	10 405 (0.8)	9761 (0.8)	416 (0.7)	229 (0.5)
American Indian/Alaska Native	4451 (0.4)	4239 (0.4)	149 (0.3)	157 (0.4)
Multiple Race	20 411 (1.6)	27 763 (2.4)	1844 (3.3)	801 (1.8)
Other	38 425 (3.0)	36 667 (3.1)	888 (1.6)	747 (1.7)
Unknown	112 380 (8.9)	108 276 (9.3)	1640 (2.9)	2382 (5.3)
White	813 394 (64.3)	739 788 (63.5)	39 439 (70.9)	34 351 (76.6)
Ethnicity				
Hispanic	74 812 (5.9)	66 266 (5.7)	2558 (4.6)	6181 (13.8)
Non-Hispanic	846 108 (66.9)	768 228 (66.0)	47 725 (85.8)	30 358 (67.7)
Unknown	344 156 (27.2)	326 166 (28.0)	5340 (9.6)	12 456 (27.8)

Note: The table reports the 2016 aggregated data from the participating integrated healthcare systems, with 2 exceptions noted below.

[a]Percentiles reported across integrated healthcare systems (IHCS) are based on the IHCS' total population to report the percent of cancer and cancer-free adults used in the algorithm development. Percentiles in the remaining table describe the characteristics within each classification.

[b]HealthPartners Institute used 2020 data instead of 2016 to modify the KPCO algorithm, as described in the Materials and methods section.

[c]Sanford Health was unable to report on the total eligible population and cases within their tumor registry based on the methodology used to modify the KPCO algorithm.

[d]The study included aggregated electronic medical record (EMR) data from 7 integrated healthcare institutions. HealthPartners Institute's EMR was established in 2000 and includes a tumor registry created in 1984. Henry Ford Health System's EMR was established in 2013 and includes a tumor registry created in 1985. Kaiser Permanente Colorado's EMR was established in 1998 and includes tumor registry data beginning in 2000. Kaiser Permanente Georgia's EMR was established in 1996 and includes a tumor registry created in 2010. Kaiser Permanente Hawaii's EMR was established in 2004 and includes a tumor registry established in the 1980s and converted to an electronic database in 2003. Kaiser Permanente Northwest's EMR was established in 1996 and includes a tumor registry created in 1974. University of Chicago EMR was established in 2006 and includes a tumor registry created in 1954. Sanford Health's EMR was established in 2005 and includes a tumor registry created in: Aberdeen, SD (2015), Bemidji, MN (2011), Fargo, ND (2004), Sioux Falls, SD (1983), Worthington, MN (2015), and Bismarck, ND (2002).

Table 2. Final algorithm attributes across each of the participating integrated healthcare systems

	Integrated healthcare systems specific algorithm final version attributes							
	HealthPartners Institute	Henry Ford Health System	Kaiser Permanente Colorado	Kaiser Permanente Georgia	Kaiser Permanente Hawaii	Kaiser Permanente Northwest	University of Chicago	Sanford Health
Tumor registry record		X	X	X	X	X	X	
Diagnosis codes								
ICD-9	X	X	X	X	X	X	X	
ICD-10	X	X	X	X	X	X	X	X
Encounters								
Oncology department visit		X	X	X	X	X		
Chemotherapy treatment		X	X	X	X	X		
Chemotherapy ICD-9	X	X	X	X	X	X		
Chemotherapy ICD-10	X	X	X	X	X	X		
Revenue Codes		X	X	X	X	X		
Current Procedural Terminology-4		X	X	X	X	X		
Healthcare Common Procedure Coding System		X	X	X	X	X		

Table 3. Example of calculating KPCO's weighted sensitivity to account for cancer cases identified in either the tumor registry or by the algorithm

Among eligible population	Percent of cancer cases	Sensitivity
A = 5650	**47.04% = 5650/12 012**	F = 100%
Number of cancer cases identified in tumor registry	D = A/C	Tumor registry sensitivity
B = 6632	**52.96% = 6632/12 012**	**82.35% = (14 true positives)/ (17 total cancer cases identified by chart review)**
Number of cancer cases identified by algorithm	E = B/C	G = (Number of true positives from chart review)/ (Number of all cancer cases identified by chart review)
12 012 = 5650 + 6632		90.65% = (47.04% × 100%) + (52.96% × 82.35%)
C = A + B		H = (D × F) + (E × G)
Total number of cancer cases identified in tumor registry or by algorithm		**Weighted sensitivity**

The bold text represents the chart review results from KP CO to provide an example of weighted sensitivity calculation.

in their tumor registry, followed by KPHI (5.9%), KPNW (5.2%), HFHS (4.2%), HP (3.6%), KPCO (3.3%), and KPGA (3.2%). Among individuals with a reported cancer not captured in the tumor registry, 35.8% were aged 60–65, 65.0% were female, 3.6% Asian, 10.2% Black, 0.5% Hawaiian/Pacific Islander, 76.6% White, and 13.8% Hispanic. Compared to cases identified in the tumor registry, cancer cases identified by the algorithm not in the tumor registry had a higher percentage of people age 40–44 years (6.9% vs 9.3%) and 45–49 years (10.6% vs 12.6), more females (61.5% vs 65.0%), and a higher percentage of Hispanic members (4.6% vs 13.8%).

Each site revised the algorithm distributed by KPCO through systematic chart review standardized across IHCS. Table 2 displays the attributes in the final algorithm. All 8 IHCS included ICD-10 codes to identify a history of cancer. HFHS, KPCO, KPGA, KPHI, KPNW, and University of Chicago included tumor registry data. All 4 Kaiser Permanente systems and HFHS included revenue codes, procedure codes, oncology department visits, and chemotherapy in their final algorithm. Supplementary Table S1 describes the changes KPCO made to their initial algorithm, including the addition of ICD-10 codes Z86.0, Z86.00, and Z86.000.

To calculate weighted sensitivity for HFHS, KPCO, KPGA, KPHI, and KPNW (Figure 1), we assumed the tumor registry sensitivity was 100% and the overall sensitivity was a weighted average of the tumor registry sensitivity and the chart review sensitivity. Using Figure 1 described in the Materials and methods section and the KPCO results, Table 3 provides an example of how the weighted sensitivity was calculated. For example, KPCO's tumor registry's sensitivity is 100% and includes 47.04% of the total number of cancer cases identified. The KPCO calculated algorithm sensitivity for members not in the tumor registry and based on chart review was 82.35% (sensitivity =14/17), and the percentage of cancers not in the tumor registry was 52.96%. The weighted sensitivity combines each of these percentages and sensitivities to report a final weighted sensitivity of 90.65% (weighted sensitivity = (47.04% × 100%) + (52.96% × 82.35%)). Supplementary Table S2 provides a more detail on the weighted sensitivities from HFHS, KPCO, KPGA, KPHI, and KPNW.

Table 4 reports the performance metrics of the final algorithm by IHCS. The algorithm's sensitivity, or the ability to detect true cancer cases, remained above 90% across all IHCS (HP = 100%; HFHS = 97.66%; KPCO = 90.65%; KPGA = 100%; KPHI = 100%;

Table 4. Final algorithm performance based on randomized chart review across each integrated healthcare system

		HealthPartners Institute[a]	Henry Ford Health System	Kaiser Permanente Colorado	Kaiser Permanente Georgia	Kaiser Permanente Hawaii	Kaiser Permanente Northwest	University of Chicago	Sanford Health
	Number of cancer cases in the eligible population identified in tumor registry or by algorithm	NA	12465	12012	6895	5154	14421	NA	NA
Sensitivity	Tumor registry sensitivity[b,c]	NA	100%	100%	100%	100%	100%	NA	NA
	Number of cancer cases in tumor registry	NA	7223	5650	3331	3312	8846	NA	NA
	Percent of cancer cases that were in the tumor registry	NA	57.95%	47.04%	48.31%	64.26%	61.34%	NA	NA
	Calculated sensitivity for members not in the tumor registry:	100%	94.44%	82.35%	100%	100%	100%	92.9%	98.4%
	Number of patients in random sample identified as cancer cases by algorithm and chart review, numerator	5	17	14	11	9	12	13	184
	Number of patients in random sample that were identified to have cancer by chart review, denominator	5	18	17	11	9	12	14	187
	Percent of cancer cases identified by the algorithm[b]	NA	42.05%	52.96%	51.69%	35.74%	38.66%	NA	NA
	Weighted average of the sensitivities[b,c]	NA	97.66%	90.65%	100%	100%	100%	NA	NA
Specificity	Number of patients in random sample identified as cancer-free by algorithm and chart review, numerator	40	79	77	80	80	79	94	182
	Number of patients in random sample that were identified as cancer-free by chart review, denominator	45	82	83	89	91	87	96	188
	Calculated Specificity	88.89%	96.3%	92.77%	89.89%	87.91%	90.80%	97.9%	96.8%

[a]HealthPartners Institute developed an algorithm with a high sensitivity that leverages electronic medical record system to identify cancer adults who are not eligible to participate in Connect for Cancer Prevention.
[b]HealthPartners Institute, University of Chicago, and Sanford Health based their algorithm's sensitivity and specificity on chart abstraction and did not include the sensitivity of their tumor registry.
[c]Henry Ford Health System and the Kaiser Permanente (KP) sites calculated a weighted sensitivity to account for the algorithm and tumor registry's sensitivity within the Henry Ford Health System and KP EMR data.

KPNW = 100%; University of Chicago = 92.86%; and Sanford Health = 98.40%). Specificity was 87.91% or higher across all IHCS.

DISCUSSION

A consortium of 8 IHCS partnered to develop and validate an algorithm to identify a patient's history of cancer with high sensitivity and specificity. Each site tailored the algorithm to their data systems to achieve a sensitivity and specificity above 90% for both. False negatives usually were a result of "history of cancer" being documented in the medical note or patient's problem list, but with no other supporting evidence in the EMR. False positives occurred infrequently and were usually a result of the algorithm identifying (1) infusions or use of chemotherapies for conditions other than cancers or (2) "cancer" diagnoses that were incorrectly reported by the patient, were benign lesions, or skin cancers that were documented in notes or the problem list. For example, a patient reported a benign lesion (such as a polyp) as a cancer, or documented a skin cancer that was not reportable, such as basal cell cancer. The common attributes of the final algorithm at each site were diagnosis codes. Despite differences in EMR database systems, our method was robust and required only minor modifications to adapt to other IHCS.

The algorithm development is part of Connect for Cancer Prevention Study, a longitudinal cohort study planned to include 200 000 cancer-free adults to better understand cancer etiology and cancer outcomes (https://www.cancer.gov/connect-prevention-study/). The developed algorithm will be applied to each enrolling IHCS as part of the eligibility criteria used for participant recruitment. While the algorithm's sensitivity and specificity remained above 90%, we acknowledge the algorithm may need to be revised throughout the duration of the project to maintain an acceptable level of performance.

Local tumor registries can be utilized to identify incident cancer cases but may miss individuals with a history of cancer that may have been diagnosed prior to enrolling with a health plan. Furthermore, available tumor registry data can lag up to 2 years to account for newly diagnosed cases and new information received about previously submitted cases. EMR databases contain information on the most recent patient encounters and details on prior diagnoses and diagnostic testing. Prior work by Clarke and Feigelson[19] showed that applying an algorithm within the EMR databases to identify individuals with a distant or recent history of cancer can account for these limitations and provide a suitable workaround. However, that previous algorithm was only tested in a single IHCS and did not account for nuances between tumor registries and IHCS EMR data.

Our intent was to develop an algorithm to leverage the tumor registry's completeness of incident cancer cases and update diagnosis, treatment, and procedure codes that leverage the robust retrospective data in the EMR across several IHCS. By leveraging these data, we can also account for history of cancer that is reported by new members and patients to the IHCS as they continue to interact with their healthcare providers. Had we relied solely on data from each of the IHCS' tumor registries and not included the EMR data, we would have missed nearly half of the reported cancers. In addition to the algorithm being comprehensive of tumor registry and EMR data, we aimed for the algorithm to be accessible across IHCS and to accurately identify adults who did not have a history of cancer. Therefore, to improve the utility of the algorithm across IHCS, we used only discrete data elements. By prioritizing utilization, coupled with high performance, the algorithm could be used for patient surveillance in addition to identifying a cancer-free cohort for the Connect for Cancer Prevention project.

As noted earlier, most invasive cancer diagnoses are included in tumor registries that follow NAACCR standards, while reporting of *in situ* cases is not as consistent. Therefore, it is important to distinguish *in situ* cases as they are represented in ICD-O-3 behavior codes when applying the algorithm to an EMR database.[26] For all tumors diagnosed on or after January 1, 2001, NAACCR requires tumor registries to include all cancers with an ICD-O-3 behavior code of 2 or 3 (*in situ* or malignant), with the exception of several skin carcinomas, and cervical and prostatic neoplasia.[26]

To our knowledge, our project is the first to partner with multiple IHCS to systematically develop an algorithm that will identify individuals with a history of cancer, regardless of when that diagnosis occurred. However, there are limitations to our approach that should be noted. First, the algorithm was developed and validated across Epic-based EMRs and may not be easily applied to other EMR databases. Second, our eligible cohort used to develop and validate the algorithm was restricted to individuals 40–65 years of age; the algorithm may perform differently in a broader age group. Third, each participating IHCS had a robust EMR system that dates back between 15 and 25 years and the accuracy of the algorithm is limited by the completeness of the EMR. Applying the algorithm to an EMR system that is less complete or does not include historic information will decrease the accuracy as there will be less cases for the algorithm to sample. Fourth, the algorithm did not include data elements such as medical notes or prescription drugs which may have improved the sensitivity but decreased the specificity or the ability to implement across IHCS. Future algorithm revisions should consider adding these elements and assess the algorithm for accuracy. Finally, data may be missing for IHCS members who are relatively new to the health plan, or who rarely interact with the healthcare system.

CONCLUSION

We developed and validated a high-performing algorithm to identify individuals with a history of cancer as part of eligibility criteria for a prospective cohort study of cancer-free adults. Combining the strengths of tumor registries and EMR creates a robust algorithm that can identify those with a history of cancer, regardless of whether that diagnosis was recent or in the distant past. The ability to identify this broad group may be beneficial for quality improvement projects that outreach to individuals with a history of cancer, specifically to newly enrolled members and new patients of a health plan to identify unmet or ongoing healthcare needs. The algorithm could be applied to other healthcare systems that employ an EMR and have tumor registry data available, or to supplement the tumor registry to identify incident cancer cases diagnosed within 0–12 months and may not yet be included in the registry.

FUNDING

This work was funded by the National Cancer Institute, Contracts HHSN2612018000211.

AUTHOR CONTRIBUTIONS

All authors have met the International Committee of Medical Journal Editors criteria for authorship. Drs. Gander and Feigelson led

the investigation and the development of the manuscript. The lead investigators at each IHCS (Drs. Gander, Feigelson, Pawloski, Rybicki, Honda, Goddard, Ahsan, and Chan) take full responsibility for the integrity of their IHCS data and the accuracy of the data analysis. All authors have made substantial contributions to the concept and design, drafted and revised it for content, approved the final version for publication, and agree to be accountable for all aspects of the work. A more detailed list of each author's contribution is provided below.

Concept and design: Gander, Maiyani, White, Clarke, and Feigelson. *Acquisition, analysis, or interpretation of data*: Gander, Maiyani, White, Sterrett, Pawloski, DeFor, Olsen, Rybicki, Neslund-Dudas, Sheth, Krajenta, Purushothaman, Honda, Yonehara, Goddard, Prado, Ahsan, Kibriya, Aschebrook-Kilfoy, Chan, Hague, Thompson, Sawyer, and Feigelson. *Drafting of the manuscripts*: Gander, Maiyani, White, Güney, Pawloski, Rybicki, Honda, Goddard, Ahsan, Aschebrook-Kilfoy Chan, and Feigelson. *Critical revision of the manuscript for important intellectual content:* Gander, Pawloski, Rybicki, Honda, Goddard, Ahsan, Aschebrook-Kilfoy, Chan, Clarke, Gaudet, Feigelson. *Statistical analysis*: Gander, Maiyani, Sterrett, DeFor, Olsen, Neslund-Dudas, Sheth, Krajenta, Purushothaman, Yonehara, Prado, Kibriya, Aschebrook-Kilfoy, Hague, and Sawyer. *Obtained funding*: Gander, Pawloski, Rybicki, Honda, Goddard, Kibriya, Chan, Gaudet, and Feigelson. *Administrative, technical, or material support*: Gander, Maiyani, White, Sterrett, Güney, DeFor, Neslund-Dudas, Yonehara, Prado, Kibriya, Aschebrook-Kilfoy, Hague, Thompson, Sawyer, and Feigelson. *Supervision*: Gander, Pawloski, Rybicki, Honda, Goddard, Ahsan, Chan, and Feigelson.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We would like to thank the research staff at Kaiser Permanente (KP) Georgia, KP Colorado, HealthPartners Institute, Henry Ford Health System, KP Hawaii, KP Northwest, University of Chicago, and Sanford Health. The algorithm development, algorithm validation, and chart review across 8 IHCS was a large undertaking that involved participation of multiple coauthors from each IHCS. This work could not have been done without the hard work and contribution of each team member across the IHCS. The authors would also like to thank the National Cancer Institute's Division of Cancer Epidemiology and Genetics investigators and staff for making this work possible.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article cannot be shared publicly due to *the privacy of individuals and IHCS members that participated in the study.* The derived data will be shared on reasonable request to the corresponding author.

REFERENCES

1. Howard DH, Sentell T, Gazmararian JA. Impact of health literacy on socioeconomic and racial differences in health in an elderly population. *J Gen Intern Med* 2006; 21 (8): 857–61.
2. Bao Y, Bertoia ML, Lenart EB, *et al.* Origin, methods, and evolution of the three Nurses' Health Studies. *Am J Public Health* 2016; 106 (9): 1573–81.
3. Kannel WB, Feinleib M, McNamara PM, *et al.* An investigation of coronary heart disease in families: the Framingham Offspring Study. *Am J Epidemiol* 1979; 110 (3): 281–90.
4. Dawber TR, Meadors GF, Moore FE, Jr. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health* 1951; 41 (3): 279–86.
5. Parikh-Patel A, Allen M, Wright WE; California Teachers Study Steering Committee. Validation of self-reported cancers in the California Teachers Study. *Am J Epidemiol* 2003; 157 (6): 539–45.
6. Bernstein L, Allen M, Anton-Culver H, *et al.* High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). *Cancer Causes Control* 2002; 13 (7): 625–35.
7. Levêque-Morlais N, Tual S, Clin B, *et al.* The AGRIculture and CANcer (AGRICAN) cohort study: enrollment and causes of death for the 2005–2009 period. *Int Arch Occup Environ Health* 2015; 88 (1): 61–73.
8. Colditz GA. Epidemiology of breast cancer: findings from the Nurses' Health Study. *Cancer* 1993; 71 (4 Suppl): 1480–9.
9. Gaziano JM, Sesso HD, Christen WG, *et al.* Multivitamins in the prevention of cancer in men: the Physicians' Health Study II randomized controlled trial. *JAMA* 2012; 308 (18): 1871–80.
10. Gaziano JM, Glynn RJ, Christen WG, *et al.* Vitamins E and C in the prevention of prostate and total cancer in men: the Physicians' Health Study II randomized controlled trial. *JAMA* 2009; 301 (1): 52–62.
11. All of Us Research Program Investigators. The "All of Us" research program. *N Engl J Med* 2019; 381 (7): 668–76.
12. Hammond EC. Smoking in relation to mortality and morbidity. Findings in first thirty-four months of follow-up in a prospective study started in 1959. *J Natl Cancer Inst* 1964; 32 (5): 1161–88.
13. Garfinkel L. Selection, follow-up, and analysis in the American Cancer Society prospective studies. *Natl Cancer Inst Monogr* 1985; 67 : 49–52.
14. Rebbeck TR, Burns-White K, Chan AT, *et al.* Precision prevention and early detection of cancer: fundamental principles. *Cancer Discov* 2018; 8 (7): 803–11.
15. Menck H, Smart CR. *Central Cancer Registries: Design, Management, and Use*. Boca Raton, FL: CRC Press; 1994.
16. Zachary I, Boren SA, Simoes E, *et al.* Information management in cancer registries: Evaluating the needs for cancer data collection and cancer research. *Online J Public Health Inform* 2015; 7 (2)
17. Zachary I. *Improving the Usability and Utilization of Cancer Registry Data: The Need to Identify a Core Data Set*. University of Missouri, Columbia; 2012.
18. US National Institutes of Health. National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch. Surveillance, Epidemiology, and End Results (SEER). Program Research Data; 1973–2008.
19. Clarke CL, Feigelson HS. Developing an algorithm to identify history of cancer using electronic medical records. *EGEMS (Wash DC)* 2016; 4 (1): 1209.
20. Ross TR, Ng D, Brown JS, *et al.* The HMO research network virtual data warehouse: a public data model to support collaboration. *EGEMS* 2014; 2 (1): 2.
21. Ritzwoller DP, Carroll N, Delate T, *et al.* Validation of electronic data on chemotherapy and hormone therapy use in HMOs. *Med Care* 2013; 51 (10): e67–e73.
22. Hornbrook MC. Building a virtual cancer research organization. *JNCI Monographs* 2005; 2005 (35): 12–25.
23. Goldberg DW, Kohler B, Kosary C. *The Texas A&M, NAACCR, NCI Geocoding Service*. 2021. http://geo.naaccr.org. Accessed September 2021.
24. Krickeberg K, Van Trong P, Thi My Hanh P. *Epidemiology: Key to Public Health*. Switzerland: Springer; 2019.
25. Ross TR, Ng D, Brown JS, *et al.* The HMO research network virtual data warehouse: a public data model to support collaboration. *EGEMS (Wash DC)* 2014; 2 (1): 1049.
26. Thornton M. North American Association of Central Cancer Registries; 2011.