BMC Cancer

RESEARCH Open Access

The Kaiser Permanente Research Bank Cancer Cohort: a collaborative resource to improve cancer care and survivorship

Heather Spencer Feigelson[1*], Christina L. Clarke[1], Stephen K. Van Den Eeden[2], Sheila Weinmann[3], Andrea N. Burnett-Hartman[1], Sarah Rowell[4], Shauna Goldberg Scott[1], Larissa L. White[1], Monica Ter-Minassian[5], Stacey A. A. Honda[6], Deborah R. Young[7], Aruna Kamineni[8], Terrence Chinn[2], Alexander Lituev[9], Alan Bauck[3] and Elizabeth A. McGlynn[10]

Abstract

Background: The Kaiser Permanente Research Bank (KPRB) is collecting biospecimens and surveys linked to electronic health records (EHR) from approximately 400,000 adult KP members. Within the KPRB, we developed a Cancer Cohort to address issues related to cancer survival, and to understand how genetic, lifestyle and environmental factors impact cancer treatment, treatment sequelae, and prognosis. We describe the Cancer Cohort design and implementation, describe cohort characteristics after 5 years of enrollment, and discuss future directions.

Methods: Cancer cases are identified using rapid case ascertainment algorithms, linkage to regional or central tumor registries, and direct outreach to KP members with a history of cancer. Enrollment is primarily through email invitation. Participants complete a consent form, survey, and donate a blood or saliva sample. All cancer types are included.

Results: As of December 31, 2020, the cohort included 65,225 cases (56% female, 44% male) verified in tumor registries. The largest group was diagnosed between 60 and 69 years of age (31%) and are non-Hispanic White (83%); however, 10,076 (16%) were diagnosed at ages 18–49 years, 4208 (7%) are Hispanic, 3393 (5%) are Asian, and 2389 (4%) are Black. The median survival time is 14 years. Biospecimens are available on 98% of the cohort.

Conclusions: The KPRB Cancer Cohort is designed to improve our understanding of treatment efficacy and factors that contribute to long-term cancer survival. The cohort's diversity - with respect to age, race/ethnicity and geographic location - will facilitate research on factors that contribute to cancer survival disparities.

Keywords: Cancer, Survivorship, Cohort study, Disparities, Genetic

Background

Between 2016 and 2040, the projected prevalence of cancer survivors in the United States (U.S.) will increase from an estimated 15.5 million (approximately 4.8% of the population) to 26.1 million survivors (nearly 7% of the total population) across all age groups [1]. With advances in early detection and the increasing effectiveness of cancer treatments, long-term survivorship is becoming a reality for more than half of those diagnosed with cancer. Currently, an estimated 64% of all survivors have lived at least 5 years beyond diagnosis and 40% have lived at least 10 years beyond diagnosis [2, 3].

In 2015, then President Obama announced the Precision Medicine Initiative [4] with the intent to accelerate research that integrates molecular and genomic

*Correspondence: heather.s.feigelson@kp.org
[1] Institute for Health Research, Kaiser Permanente, 2550 S. Parker Rd, Suite 200, Aurora, CO 80014, USA
Full list of author information is available at the end of the article

information into medical care. This was followed in 2016 by the Enactment of the Cancer Moonshot within the twenty-first Century Cures Act, to accelerate progress to prevent, diagnose and treat cancer [5]. While these initiatives have provided enthusiasm and funding for the identification of highly targeted agents that improve cancer treatment efficacy, to date, the focus on precision medicine has largely ignored other factors that may influence treatment and survival, such as co-morbid conditions, medication use, and lifestyle habits. The promise of precision medicine should include medical care that goes beyond the acute treatment period and maximizes long-term quality of life for cancer survivors.

Kaiser Permanente (KP) is an ideal environment to study the spectrum of cancer care and long-term survival because it provides comprehensive medical care, including cancer care, to a large and diverse membership. For decades, KP has used electronic health records (EHR) to track patient care, including preventive services, diagnoses, and treatments. The vast majority of KP members diagnosed with cancer receive all their follow-up care within KP facilities and the details of that care are captured in the EHR. The KP Research Bank (KPRB), a research resource within KP, is seeking to collect biospecimens and surveys from at least 400,000 adult members across the U.S. and linking that information to the EHR. Within the KPRB, we have developed a Cancer Cohort to specifically address issues related to cancer survival, and to understand how genetic, lifestyle, and environmental factors impact cancer treatment, treatment sequalae, and prognosis. In this paper, we describe the design and implementation of the KPRB Cancer Cohort, present characteristics of the cohort after 5 years of enrollment and discuss future directions.

Methods

The Kaiser Permanente Research Bank

KP provides health care for approximately 12.5 million individuals across eight regions: Colorado, Georgia, Hawai'i, Mid-Atlantic, Northern California, Northwest, Southern California, and Washington. All KP members aged 18 and older whose preferred language is English or Spanish are eligible to join the KPRB. Launched in May 2016, the KPRB is built on an existing repository (Research Program on Genes, Environment and Health, or RPGEH) of approximately 220,000 KP members with data and biospecimens collected primarily between 2007 and 2011 from members of KP Northern California, and to a smaller extent, Southern California [6]. The KPRB consists of three cohorts: a general cohort, a pregnancy cohort [7], and a cancer cohort. The primary recruitment mode is by email invitation; up to 5 invitations to enroll are sent. KP Washington members are invited to participate by postal mail letters only, and a small percentage of members across the remaining regions who are not registered on the member portal also receive mailed invitations. Several regions also use in-person recruitment to increase enrollment among racially and ethnically diverse members. KP members can also initiate enrollment without a specific initiation by visiting the KPRB website (https://researchbank.kaiserpermanente.org/). Approximately 396,000 KP members have consented to join the KPRB as of July 2021, and biospecimens have been collected on approximately 371,000 adults (94% of those consented).

Participation in the KPRB involves completing the consent form, survey, and providing a saliva or blood sample. Except for those who request paper materials, members enroll by visiting the KPRB website and completing an online consent and survey. This triggers the placement of a blood draw order for specimen collection. The consent form gives permission to access the EHR, ascertain and store clinically collected specimens (including tissue biopsies and surgical resections), allow future contact, as well as use of survey data, DNA, and other blood components for research. The baseline survey collects information on demographics; neighborhood walkability/safety; work exposures and shift work; general health (PROMIS 10); social media use; chronic pain; diet, physical activity, sleep, and sedentary behavior; multi-vitamin, vitamin D, calcium, and non-steroidal anti-inflammatory drugs; personal and family history, including cancer; alcohol, tobacco, marijuana, and opiate drugs; stress, isolation, support, and discrimination; reproductive history in women; urinary function and erectile dysfunction in men; genetic testing, sigmoidoscopy and colonoscopy; and health literacy. The KPRB recruitment protocol was approved by the Institutional Review Board (IRB) in KP Mid-Atlantic States which serves as the single IRB for KPRB recruitment.

Biospecimen collection

Blood samples are collected at the member's convenience at any KP clinical laboratory and include up to 4 ml of whole blood (collected in one EDTA tube) for DNA and up to 8.5 ml of serum (in two SST tubes). Specimens are collected and shipped Monday through Friday (and Saturday in some regions) by overnight courier to the biorepository in Berkeley, CA. Specimens are tracked, and "needle to freezer" time is recorded for each sample; 56 and 83% are processed and frozen within 48 and 72 h of collection, respectively. DNA is extracted from whole blood using an automated Thermo Kingfisher system. DNA samples are dried down in Micronic tubes with Biomatrica DNAstable Plus storage medium for ambient storage using Thermo Savant Explorer SpeedVac system.

DNA samples are re-hydrated, quantified, and normalized prior to distribution for research projects. Serum and whole blood are transferred into Micronic tubes for freezing and long-term storage at -80° C.

Linkage to electronic health data

Each KP region captures EHR data for research using a standardized, structured common data model, known as the Virtual Data Warehouse (VDW). The VDW includes comprehensive data on patient characteristics, diagnoses, medical procedures, and medication use [8, 9]. EHR data from KPRB members are updated in the KPRB data repository every 6 months. Most data are current at the time of the update, but data that come from outside sources, such as vital status and cause of death, may lag up to 12 months. Sources of vital status data include: the EHR, state vital statistics, and local (e.g. tumor registry) or national (e.g. Social Security Administration, National Death Index) registries.

The Cancer Cohort

The KPRB Cancer Cohort uses the same website, consent form, and blood draw protocol as the general cohort. The Cancer Cohort survey includes all questions from the general cohort with additional questions specifically relating to cancer about genetic testing, cancer screening, and family history of specific cancers. Study materials are available in English and Spanish (beginning in 2018) languages.

We developed recruitment strategies and materials specifically designed to increase the enrollment of people with cancer. First, we developed a rapid case ascertainment (RCA) algorithm (described below) to identify cancer cases shortly after diagnosis. Second, from January–June of 2021, we used tumor registry and EHR data to identify members with a history of cancer (regardless of when the diagnosis occurred) and invited them to enroll using recruitment materials with content specifically about cancer survivorship. Some regions used direct outreach to oncology departments and care navigators, and in-clinic recruiters to increase enrollment; patients received flyers about enrolling into the KPRB in materials provided during their oncology visits. In addition to these methods, cancer cases are identified from members of the general cohort who report a history of cancer, and incident cases that arise in the general cohort after enrollment. Cancer stage is defined using Surveillance, Epidemiology and End Results (SEER) general summary stage [10] as follows: in-situ, localized, regional, distant, benign/borderline, unknown, or missing. Study enrollment was paused in March 2020 due to the SARS-CoV-2 pandemic. Cancer cases identified using the RCA were invited to enroll again starting in October 2020.

Development of the rapid case ascertainment (RCA) algorithm

To maximize the value of this resource, we aimed to identify and enroll cancer cases shortly after diagnosis, allowing us to minimize survival bias and, in some cases, collect specimens prior to the start of treatment.

KP maintains regional tumor registries in 6 of 8 KP regions to identify and track cancers that were diagnosed and treated within the health system (2 regions have relied on linkage to SEER registries) [10]. While tumor registries are considered the gold standard for the identification of incident cancers, data in the tumor registries can lag 6–18 months behind the date of diagnosis because tumor registrars identify potentially eligible cases and then manually abstract to confirm the diagnosis and ascertain the state or federally required data elements about the cancer [11]. Thus, we developed RCA algorithms to identify new cancer cases within days of diagnosis and assign an anatomical site with a high degree of accuracy using data pulled directly from electronic pathology files (Fig. 1).

Because each KP region uses different software and systems for pathology data, region-specific approaches to RCA were required. However, each region with a tumor registry followed a similar process in algorithm development. The tumor registry was used as the gold standard, and the best data source from the pathology department was determined. In some regions, a file is sent directly from pathology to the tumor registrars, and when available this was the file of choice. Tumor registry data from a prior year was used to develop the algorithm to ensure that the data were complete. Once the pathology data were acquired, the development of the algorithm was iterative; computer code (using SAS, version 9.4 (SAS Institute, Cary NC)) was developed to identify cancer cases from non-cancer pathology specimens. Sensitivity and specificity were calculated for each iteration, and chart review was used as necessary to understand at each iteration why false positive and false negative cases were mis-specified. Because the algorithms were intended to identify people with a high probability of cancer, we focused on achieving high specificity, but also evaluated the loss of sensitivity for each incremental gain of specificity to make the algorithm as robust as possible. As a final step, the algorithm was tested on "live" data prior to launch, and chart review was conducted on a small number of records (< 50 in each region) to verify that the algorithm was performing as expected.

Two regions (Northwest and Northern California) used the Systematized Nomenclature of Medicine (SNOMED) coding system [12]. From SNOMED, we used a selection of morphology, "M" codes that indicated cancer and the corresponding topography/

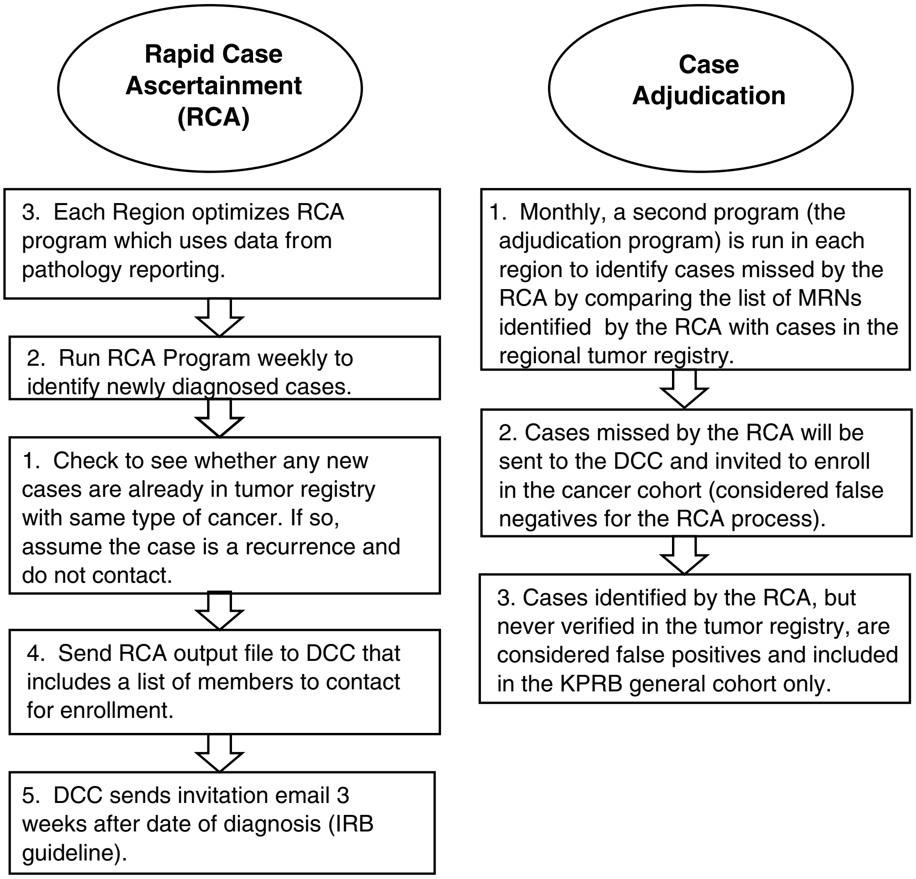
Rapid Case Ascertainment (RCA)
Case Adjudication
3. Each Region optimizes RCA program which uses data from pathology reporting.
2. Run RCA Program weekly to identify newly diagnosed cases.
1. Check to see whether any new cases are already in tumor registry with same type of cancer. If so, assume the case is a recurrence and do not contact.
4. Send RCA output file to DCC that includes a list of members to contact for enrollment.
5. DCC sends invitation email 3 weeks after date of diagnosis (IRB guideline).
1. Monthly, a second program (the adjudication program) is run in each region to identify cases missed by the RCA by comparing the list of MRNs identified by the RCA with cases in the regional tumor registry.
2. Cases missed by the RCA will be sent to the DCC and invited to enroll in the cancer cohort (considered false negatives for the RCA process).
3. Cases identified by the RCA, but never verified in the tumor registry, are considered false positives and included in the KPRB general cohort only.

Fig. 1 Diagram of the Rapid Case Ascertainment Process. Rapid Case Ascertainment (RCA) process uses pathology data to identity new cancer cases within days of diagnosis. Probable cancer cases are adjudicated using linkage to the tumor registry data in each Kaiser Permanente region. Abbreviations: DCC: Data Coordinating Center; IRB: Institutional Review Board; MRN: Medical Record Number

anatomical terms "T" from the same pathology report. The "T" codes were used to identify the anatomical site from the pathology report and created a "probable cancer site" variable within the algorithm. The KP Colorado tumor registry uses the E-Path software system (https://www.inspirata.com/solutions/cancer-registry-automation/). Working with the local tumor registry at KP Colorado, we developed a direct feed from E-Path for the RCA system. A similar process was adopted at KP Mid-Atlantic States, which uses a direct feed from the EHR and CoPath Plus to the Cancer Alert system (CAS) which scans both feeds for defined International Classification of Diseases (ICD) codes and keywords. The E-Path and CAS algorithms are the primary source of records by tumor registrars. KP Southern California used the software system, Linguamatics product I2E (https://www.linguamatics.com/products/i2e), to identify cases from pathology reports. Linguamatics (Marlborough, MA) is a commercial software coding product to mine text using natural language processing (NLP). Linguamatics developed the NLP coding, and it was distributed to KP Southern California to test and refine on their pathology reports. KP Hawai'i developed a program to identify cancer cases from the EHR Problem List and claims data. KP Washington also used the Problem List to identify recent cancer diagnoses among enrolled members attributed to the internal delivery system and residing within the 13-county Seattle-Puget Sound SEER registry catchment area. Among these members with Problem List-noted cancer diagnoses, pathology records with cancer-related text strings (using a list provided by the Seattle-Puget Sound SEER registry) and specimen collection dates within 60 days prior to the cancer diagnosis dates noted on the Problem List were identified. Pathology records with non-melanoma skin cancers, cancers from primary site metastasis, or those with clear cancer-negating text were excluded.

The development of the RCA process in each KP region was phased in over several months. The first regions (Northern California, Colorado, Northwest) launched in May 2016, Hawai'i launched in June 2016, Southern California launched in November 2016, Mid-Atlantic States launched in 2017, and Washington launched in 2019. The RCA was launched in each region after the algorithm was optimized and tested. After launch, the process is largely automated. Every week the RCA program runs in each region to identify new cancer cases and generates a data file that is sent to the KPRB Data Coordinating Center (DCC). The DCC checks the file to determine the member's eligibility for recruitment outreach (enrolled in KP, alive, at least 18 years of age, and not on any "do not contact" list) and then invites the member to participate in the study 3 weeks after the date of diagnosis. The IRB required a three-week window after diagnosis prior to study contact. The member is contacted via email or postal mail with an invitation that indicates they are being contacted to join the KPRB because they 'recently received a procedure or test'. The recruitment materials do not indicate that the member has, or may have, cancer. This was done to protect patient privacy and to avoid alarming patients who turn out to be false positives.

Case adjudication

Case adjudication relies on the regional or SEER tumor registries to accomplish three critical functions: (1) validate the results of the RCA process, (2) identify incident cases missed by the RCA, and (3) identify prevalent and incident cases in the KPRB general and pregnancy cohorts. Beginning at least 6 months after the Cancer Cohort is launched in a region, members who have enrolled in the Cancer Cohort are compared against the tumor registry. In most regions, case adjudication is done monthly. Data from the tumor registry are used to confirm the type of cancer, and whether the case identified by the RCA is incident or a cancer recurrence. Members missed by the RCA algorithm, and who are not already enrolled in the KPRB, are sent an invitation to enroll. Cancer cases identified from the general cohort during adjudication are flagged for inclusion in the Cancer Cohort. Because these members already have donated biospecimens and completed a survey, they are not contacted further. Finally, members identified by the RCA algorithm who are not in the tumor registry within 12 months are considered false positives; these members are no longer considered part of the Cancer Cohort and become part of the general cohort.

Results

The positive predictive value (PPV) of the RCA algorithm at each site ranges between 80 and 99%, indicating that we have successfully minimized inviting people who have not been recently diagnosed with cancer (i.e., the false positive rate is low). Response rates for both the RCA process over the entire recruitment period and the cancer survivor outreach which occurred from January–June 2021 to KP members with a history of cancer are shown in Table 1. The response rate from RCA was 9.8% overall and for survivors was 8.8%; however, we observed important differences by race/ethnicity. Response rates in non-Hispanic Whites were 14.6 and 15.3%, respectively, but response rates from RCA were less than 5% for Blacks, Hispanics, Native Hawaiian/Pacific Islanders, and those with missing race/ethnicity information. Among survivors, response rates were also less than 5% for all the above groups as well as American Indian/Alaska Natives and Asians.

Table 2 shows the characteristics of the 65,225 KPRB cancer cases that have been verified in the tumor registries through 2020 (exact dates vary by KP region). This does not include cases who were recently identified by the RCA process that have yet to be verified since tumor registry data can lag up to 18 months behind identification by the RCA. Characteristics are shown for all cases, and for those diagnosed within 6 months (prior to) enrollment, incident cases (after KPRB enrollment), and prevalent cases diagnosed more than 6 months prior to enrollment. Most members diagnosed within 6 months prior to enrollment were identified through the RCA process. Because the KPRB was built upon the preexisting RPGEH cohort of nearly 200,000 people, most members were diagnosed more than 6 months prior to enrollment (N =37,309). However, there are 21,157 incident cases, and an additional 6759 diagnosed within 6 months of KPRB enrollment. Fifty-six percent of the cases are female (N =36,492) and 44% are male (N =28,733). While most members were diagnosed over the age of 60 years, 10,076 (16%) were diagnosed between 18 and 49 years of age. Most participants are non-Hispanic White (83%), but the cohort also includes 4208 (7%) Hispanic, 3393 (5%) Asian, and 2389 (4%) Black participants. The median survival time is 14 years (currently 674,136 total person-years of follow-up) and 17,554 participants have died (data not shown). Blood or saliva specimens are available on 98% of the Cancer Cohort.

Table 3 shows the most common types of cancer included in the KPRB by sex, and the distribution of cases by stage at diagnosis. The most common types of cancer are breast, prostate, melanoma, and colorectal. Most cancers are diagnosed at the localized stage (46% of women and 54% of men), whereas 11% of cancers in women and 14% in men are diagnosed at the distant stage. This distribution reflects, in part, the high screening rates of KP members for breast, colorectal, and cervical cancers. We observed little variation in stage at diagnosis by race/

Table 1 Kaiser Permanente Research Bank Cancer Cohort response rates by method of case identification

	Cases identified by Rapid Case Ascertainment			KP members with a history of cancer in the EHR		
	Number Contacted	Number Consented	% Consented	Number Contacted	Number Consented	% Consented
Sex[a]						
Female	71,550	7439	10.4%	126,372	11,250	8.9%
Male	56,411	5115	9.1%	92,616	7956	8.6%
Age at Diagnosis (years)[b]						
< 50	28,466	1956	6.9%			
50–59	32,372	2460	7.6%			
60–69	48,694	4304	8.8%			
70–79	36,287	3462	9.5%			
≥ 80	17,844	1118	6.3%			
Race/Ethnicity[c]						
White, non-Hispanic	64,415	9394	14.6%	95,018	14,551	15.3%
Hispanic	24,558	1163	4.7%	47,722	1767	3.7%
Asian	16,607	952	5.7%	28,263	1116	3.9%
Black	14,799	564	3.8%	31,603	1061	3.4%
Multi-Racial	2431	226	9.3%	6293	329	5.2%
Native Hawaiian/Pacific Islander	2073	84	4.1%	3573	105	2.9%
American Indian/Alaska Native	687	55	8.0%	1522	69	4.5%
Other/Unknown	2394	117	4.9%	5005	210	4.2%

[a] Excludes 14 people with missing information on sex; [b]Age at diagnosis was not known for all members with a history of cancer at the time of invitation to participate; [c]Race/ethnicity information was obtained from the EHR and may not reflect self-reported data

ethnicity across cancer types (Fig. 2). Between 50 and 57% of cases were diagnosed with localized disease, and 10–17% of cases were diagnosed with distant metastases.

Discussion

The KPRB Cancer Cohort is a unique resource to study the spectrum of cancer care and survival. A key advantage of this cohort is the linkage of EHR records to biospecimens. Historically, member retention in KP has been high, facilitating passive follow-up for research studies using the EHR. A report from the Cancer Research Network that included data from 4 KP regions found that 70% of cancer cases had been health plan members for at least 5 years prior to their diagnosis and only 20% had disenrolled within 5 years after diagnosis [13]. The longitudinal data, both prior to and after cancer diagnosis provides a wealth of information on screening, premalignant lesions, specific treatment modalities, information on medication use and comorbidities, and cancer risk factors such as BMI and smoking status. The EHR can be combined with the KPRB survey data to provide a comprehensive source of risk factor and lifestyle information. Tissues samples are not formally collected as part of the KPRB but clinically collected surgical tissue samples are stored at each KP region and their use is covered in the KPRB consent form. Finally, the Cancer Cohort can be compared against the KPRB general cohort, that currently includes over 330,000 adults with no history of cancer. Detailed information about available data and the application process to use the KPRB resource (for both KP-affiliated and non-KP affiliated investigators) are available on the website (https://researchbank.kaiserpermanente.org/our-research/for-researchers).

An important strength of the Cancer Cohort is its diversity with respect to age, race/ethnicity, and geographic location. The cohort currently includes over 10,000 people diagnosed under age 50 and significant numbers of Hispanic ($N = 4208$), Asian ($N = 3393$), and Black ($N = 2389$) members. Geographic disparities in life expectancy among U.S. counties are large and increasing [14]. Socioeconomic, behavioral and health care factors all contribute to the variation in life expectancy across the U.S. Our cohort includes KP members from across the U.S. (including Hawai'i), providing an opportunity to investigate differences in cancer care and survival by geographic location among an insured population with access to health care.

Including all cancers, rather than targeting enrollment to a specific set of cancers, offers several advantages, including making in-clinic recruitment and study promotion inclusive of all KP members. Providers and clinic staff can discuss the study without considering whether

Table 2 Demographic Characteristics of Cancer Cohort members verified in the tumor registry through 2020

	Incident within 6 months prior to KPRB enrollment ($N = 6759$)		**Incident after KPRB enrollment ($N = 21{,}157$)**		**Prevalent (Diagnosed more than 6 months prior to KPRB enrollment) ($N = 37{,}309$)**		**Total ($N = 65{,}225$)**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Sex								
Female	4075	60%	11,152	53%	21,265	57%	36,492	56%
Male	2684	40%	10,005	47%	16,044	43%	28,733	44%
Age at Diagnosis (years)								
< 18					94	< 1%	94	< 1%
18–49	1124	17%	950	5%	8002	22%	10,076	16%
50–59	1498	22%	1979	9%	9366	25%	12,843	20%
60–69	2314	34%	5791	27%	12,177	33%	20,282	31%
70–79	1419	21%	7405	35%	6398	17%	15,222	23%
≥ 80	404	6%	5032	24%	1272	3%	6708	10%
Year of Diagnosis								
Before 2015	966	14%	11,017	52%	33,300	89%	45,283	70%
2015	38	< 1%	1740	8%	1664	5%	3442	5%
2016	1398	21%	1842	9%	1043	3%	4283	7%
2017	1725	26%	2359	11%	727	2%	4811	7%
2018	1804	27%	2521	12%	497	1%	4822	7%
2019	796	12%	1563	7%	73	< 1%	2432	4%
2020	32	< 1%	115	1%	5	< 1%	152	< 1%
Race/Ethnicity[a]								
White, non-Hispanic	5213	77%	18,381	87%	30,712	83%	54,306	83%
Hispanic	595	9%	1068	5%	2545	7%	4208	7%
Asian	485	7%	951	5%	1957	5%	3393	5%
Black	289	4%	587	3%	1513	4%	2389	4%
Native Hawaiian/Pacific Islander	18	< 1%	38	< 1%	82	< 1%	138	< 1%
American Indian/Alaska Native	21	< 1%	38	< 1%	74	< 1%	133	< 1%
Multi-Racial/Other	106	2%	77	< 1%	359	1%	542	1%
Missing	32	1%	17	< 1%	67	< 1%	116	< 1%

[a] Self-reported race/ethnicity (may differ from information in the electronic health record)

a particular type of cancer is included. The RCA algorithms identify all cases with a diagnosis from pathology, thus, including all cancers is no more resource intensive than only selecting certain cancer types. The disadvantages of inviting all cancer types include enrolling small numbers of uncommon cancers that may not have sufficient sample size to study with sufficient power in this cohort alone. However, these cases can still be valuable for participation in disease-specific consortia studies [15, 16]. These cases could also be used to study multiple cancers defined by a specific exposure or similar treatment (e.g., studies of human papillomavirus-related cancers [17, 18], or studies of adverse outcomes from a specific chemotherapy regimen prescribed for multiple cancers [19]).

Our approach of using RCA algorithms facilitates the identification cancers that are rapidly fatal, and thus difficult to study. For example, our cohort currently includes 2917 lung cancer cases, 1371 pancreatic cancer cases, and nearly 8000 cases diagnosed at the distant summary stage (across all cancer types). Rapid identification and enrollment also provide an opportunity to collect blood specimens prior to the start of treatment; to date, we have 12,272 serum samples from cancer cases diagnosed after enrollment or within 6 months prior to enrollment. Our RCA approach has the disadvantage of missing up to 20% of cases. Most of these cases are missed because they are diagnosed through imaging or blood tests, instead of through histopathological evaluation, and thus do not have a pathology record at the time of diagnosis.

Table 3 Anatomic Site and Summary Stage at Diagnosis for Cancer Cohort Members by Sex

	Total (*N* = 65,225)	Females (*N* = 36,492)		Males (*N* = 28,733)	
Anatomic Site	N	N	%	N	%
Breast	15,080	14,962	99%	118	1%
Prostate	10,509			10,509	100%
Melanoma	9048	4280	47%	4768	53%
Colorectal	4717	2386	51%	2331	49%
Head and Neck	3055	1615	53%	1440	47%
Leukemia	3042	1343	44%	1699	56%
Lung	2917	1661	57%	1256	43%
Bladder	2416	538	22%	1878	78%
Cervical	2212	2212	100%		
Endometrial	1994	1994	100%		
Lymphoma	1751	869	50%	882	50%
Kidney and Ureter	1562	569	36%	993	64%
Brain and Other Nervous System	1550	1006	65%	544	35%
Pancreatic and Biliary	1371	651	47%	720	53%
Upper Gastrointestinal Tract	991	345	35%	646	65%
Other[a]	3010	2061	68%	949	32%
	Total (*N* = 65,225)	Females (*N* = 36,492)		Males (*N* = 28,733)	
Summary Stage[b]	N	N	%	N	%
In-Situ	10,261	6897	19%	3364	12%
Local	32,354	16,864	46%	15,490	54%
Regional	10,243	6256	17%	3987	14%
Distant	7951	3893	11%	4058	14%
Benign/Borderline	1263	855	2%	408	1%
Unknown	1983	981	3%	1002	4%
Missing	1170	746	2%	424	1%

[a] Other includes cancers of the following systems and sites: eye, musculoskeletal, genitourinary otherwise not listed, gynecologic otherwise not listed, condition-specific sarcomas, and unspecified/unknown cancers. [b]SEER general summary stage [8]

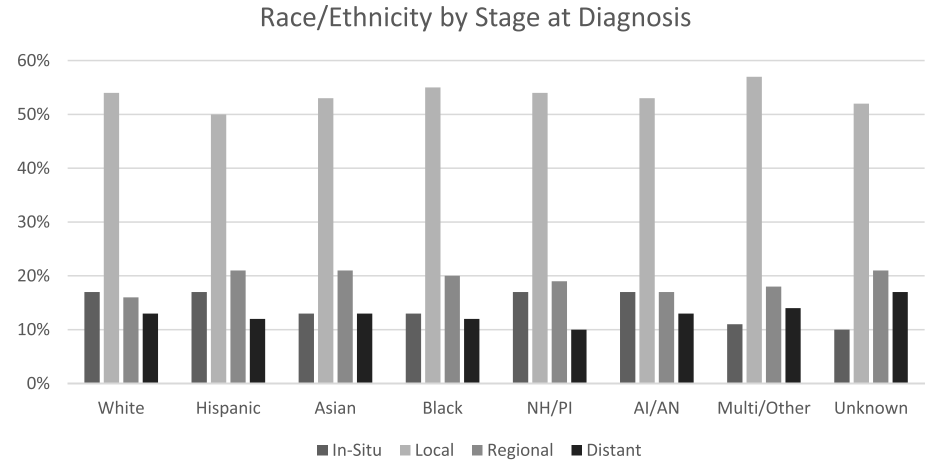

Fig. 2 Stage at Diagnosis by Race/Ethnicity of Cancer Cohort Members. NH/PI: Native Hawaiian/Pacific Islander; AI/AN: American Indian/Alaska Native; Multi/Other: Multi-Racial/or other race/ethnicity

The cases that are missed by the RCA algorithm are ultimately identified via the adjudication process with the tumor registries on average 6 months after diagnosis (but can lag up to 18 months) and invited to enroll in the cohort at that time. This lag in the tumor registry also contributes to the observed variability of the PPV.

While the cohort has many strengths, it is important to acknowledge the limitations, key among them are generalizability and the likelihood of survival bias. Our RCA methodology was designed to invite the participation of all KP members with a cancer newly diagnosed by biopsy. However, our response rate from the RCA process was only 9.8% overall, and notably lower among members of racial/ethnic groups other than non-Hispanic White, where the response rate was 14.6%. The results of our outreach to members with a history of cancer were similar; we had a 15.3% response rate among non-Hispanic Whites, but other racial/ethnic groups were below 5% (except for our "mixed/other" race category (5.2%)). Unfortunately, we did not explore the reasons why people declined to participate. It is likely that people diagnosed at an advanced stage, or otherwise with a poor prognosis, were less likely to enroll. Because members of racial/ethnic minority groups are often diagnosed at a later stage compared to Whites [20, 21], we may disproportionately under-represent advanced cancer cases among these groups. Importantly, we have over 20,000 incident cases who were diagnosed after enrollment in KPRB and this number will increase over time. While it is likely that the Cancer Cohort is not representative of the larger KP population, or cancer survivors across the U.S., we are able to query the EHR data available on all cases to understand the differences between people who enrolled in the cohort and those who did not, allowing for sample weights or other methods to statistically adjust for important biases that are identified.

Future directions

Enrollment into the KPRB Cancer Cohort will continue at least through 2022, and the entire KPRB collection will be genotyped beginning in late 2021. Even while cohort recruitment is active, there are many opportunities to use this rich resource to address key questions in cancer control. A recent portfolio review of grants across the NIH identified important gaps in survivorship research [22]. Most research conducted to date has been among breast cancer survivors, thus, there is a need to understand the unique aspects of survivorship for other cancer types, and especially among men. There is a need to expand research on older and long-term survivors and assess the effectiveness of newer cancer therapies. Further, the Blue-Ribbon Panel established as part of the Cancer Moonshot developed a set of recommendations designed to exploit new advances in cancer prevention, diagnosis, and treatment [7]. Several recommended topics can be addressed within the KPRB Cancer Cohort, including retrospective analysis of biospecimens from patients treated with standard of care and a focus on cancer screening and risk reduction among cancer survivors, about 15% of whom will develop a new incident cancer [23, 24]. The report also noted the importance of including environmental, behavioral, and health-care resource factors in "big data" approaches modeling cancer etiology and outcomes.

Conclusions

The application of precision medicine to cancer treatment must include not only genetic and molecular characterization, but consideration of lifestyle and health conditions that impact treatment efficacy, quality of life, and long-term survival. The KPRB Cancer Cohort is a rich data resource that will enable research to improve understanding of treatment efficacy and lifestyle factors that contribute to long-term cancer survival. Further, the cohort is diverse with respect to age, race/ethnicity and geographic location which will facilitate research on factors that contribute to the well-documented disparities in cancer survival [25, 26].

Abbreviations

KPRB: The Kaiser Permanente Research Bank; EHR: Electronic health records; KP: Kaiser Permanente; RPGEH: Research Program on Genes, Environment and Health; IRB: Institutional Review Board; RCA: Rapid case ascertainment; SEER: Surveillance, Epidemiology and End Results; SNOMED: Systematized Nomenclature of Medicine; CAS: Cancer Alert system; ICD: International Classification of Diseases; NLP: Natural language processing; DCC: Data Coordinating Center; PPV: Positive predictive value.

Acknowledgments

N/A

Informed consent

Informed consent was obtained from all individuals included in the study.

Authors' contributions

HSF: Conception and design, acquisition and interpretation of data, drafted the manuscript, approved submitted version, agreed to be personally accountable for this work. CLC: Design, analysis and interpretation of data, revised manuscript, approved submitted version, agreed to be personally accountable for this work. SKVDE: Design, acquisition of data, revised manuscript, approved submitted version, agreed to be personally accountable for this work. SW: Design, acquisition of data, revised manuscript, approved submitted version, agreed to be personally accountable for this work. ANBH: Design, interpretation of data, revised manuscript, approved submitted version, agreed to be personally accountable for this work. SR: Conception and design, acquisition of data, revised manuscript, approved submitted version, agreed to be personally accountable for this work. SGS: Acquisition of data, drafted manuscript, approved submitted version, agreed to be personally accountable for this work. LLW: Data analysis and interpretations, drafted manuscript, approved submitted version, agreed to be personally accountable for this work. MTM: Design, acquisition of data, revised manuscript, approved submitted version, agreed to be personally accountable for this work. SAAH: Design, acquisition of data, revised manuscript, approved submitted version, agreed to be personally accountable for this work. DRY: Design, acquisition of data, revised manuscript, approved submitted version, agreed to be

personally accountable for this work. AK: Design, acquisition of data, revised manuscript, approved submitted version, agreed to be personally accountable for this work. TC: Acquisition of data, data analysis, revised manuscript, approved submitted version, agreed to be personally accountable for this work. AL: Acquisition of data, design, revised manuscript, approved submitted version, agreed to be personally accountable for this work. AB: Acquisition of data, design, revised manuscript, approved submitted version, agreed to be personally accountable for this work. EAM: Conception and design, drafted the manuscript, approved submitted version, agreed to be personally accountable for this work.

Funding

Data used in this study were provided by the Kaiser Permanente Research Bank (KPRB) from the KPRB collection, which includes the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH), funded by the National Institutes of Health [RC2 AG036607 (Schaefer and Risch)], the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, The Ellison Medical Foundation, and the Kaiser Permanente Community Benefits Program.

Availability of data and materials

Access to data used in this study may be obtained by application to the KPRB at kp.org/researchbank/researchers.

Declarations

Ethical approval

This article does not contain any studies with animals. This study has been reviewed and approved by the Institutional Review Board at Kaiser Permanente, Mid-Atlantic States and all other Kaiser Permanente regions' IRBs ceded oversight to the Mid-Atlantic States IRB. All methods were carried out in accordance with relevant guidelines and regulations (declaration of Helsinki).

Consent for publication

N/A

Competing interests

The authors declare no competing interests.

Author details

[1]Institute for Health Research, Kaiser Permanente, 2550 S. Parker Rd, Suite 200, Aurora, CO 80014, USA. [2]Division of Research, Kaiser Permanente, 2000 Broadway, Oakland, CA 94612, USA. [3]Center for Health Research, Kaiser Permanente, 3800 N. Interstate Ave, Portland, OR 97227, USA. [4]Kaiser Permanente Program Office, 1800 Harrison, 16th floor, Oakland, CA 94612, USA. [5]Mid-Atlantic Permanente Research Institute, Kaiser Permanente, 2101 East Jefferson St, 3 West, Rockville, MD 20852, USA. [6]Center for Integrated Healthcare Research and Hawai'i Permanente Medical Group, Kaiser Permanente, 501 Alakawa St Suite 201, Honolulu, HI 96817, USA. [7]Department of Research and Evaluation, Kaiser Permanente, 100 S. Los Robles Avenue, Pasadena, CA 91101, USA. [8]Kaiser Permanente Washington Health Research Institute, 1730 Minor Ave Suite 1600, Seattle, WA 98101, USA. [9]Kaiser Permanente Research Bank, Kaiser Permanente, 1795 A Second St, Berkeley, CA 94710, USA. [10]Kaiser Permanente Research & Quality Measurement and Kaiser Permanente Research Bank, 100 S. Los Robles, 3rd floor, Pasadena, CA 91101, USA.

Received: 23 September 2021 Accepted: 21 January 2022
Published online: 25 February 2022

References

1. Bluethmann SM, Mariotto AB, Rowland JH. Anticipating the "silver tsunami": prevalence trajectories and comorbidity burden among older Cancer survivors in the United States. Cancer Epidemiol Biomark Prev. 2016;25(7):1029–36. https://doi.org/10.1158/1055-9965.EPI-16-0133.
2. de Moor JS, Mariotto AB, Parry C, et al. Cancer survivors in the United States: prevalence across the survivorship trajectory and implications for care. Cancer Epidemiol Biomark Prev. 2013;22(4):561–70. https://doi.org/10.1158/1055-9965.EPI-12-1356.
3. Rowland JH, Yancik R. Cancer survivorship: the interface of aging, comorbidity, and quality care. J Natl Cancer Inst. 2006;98(8):504–5. https://doi.org/10.1093/jnci/djj154.
4. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372(9):793–5. https://doi.org/10.1056/NEJMp1500523.
5. Jaffee EM, Dang CV, Agus DB, et al. Future cancer research priorities in the USA: a lancet oncology commission. Lancet Oncol. 2017;18(11):e653–706. https://doi.org/10.1016/S1470-2045(17)30698-8.
6. Kvale MN, Hesselson S, Hoffmann TJ, et al. Genotyping informatics and quality control for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics. 2015;200(4):1051–60. https://doi.org/10.1534/genetics.115.178905.
7. Hedderson MM, Ferrara A, Avalos LA, et al. The Kaiser Permanente Northern California research program on genes, environment, and health (RPGEH) pregnancy cohort: study design, methodology and baseline characteristics. BMC Pregnancy Childbirth. 2016;16(1):381. Published 2016 Nov 29. https://doi.org/10.1186/s12884-016-1150-2.
8. Hornbrook MC, Hart G, Ellis JL, et al. Building a virtual cancer research organization. J Natl Cancer Inst Monogr. 2005;35:12–25. https://doi.org/10.1093/jncimonographs/lgi033.
9. 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. Published 2014 Mar 24. https://doi.org/10.13063/2327-9214.1049.
10. Ruhl JL, Callaghan C, Hurlbut A, Ries LAG, Adamo P, Dickie L, et al., editors. Summary stage 2018: codes and coding instructions. Bethesda, MD: National Cancer Institute; 2020.
11. Thoburn KK, German RR, Lewis M, Nichols PJ, Ahmed F, Jackson-Thompson J. Case completeness and data accuracy in the centers for disease control and prevention's national program of cancer registries. Cancer. 2007;109(8):1607–16. https://doi.org/10.1002/cncr.22566.
12. Cote RA. Architecture of SNOMED: its contribution to medical language processing. Proc Annu Symp Comput Appl Med Care. 1993:695–9.
13. Chubak J, Ziebell R, Greenlee RT, et al. The cancer research network: a platform for epidemiologic and health services research on cancer prevention, care, and outcomes in large, stable populations. Cancer Causes Control. 2016;27(11):1315–23. https://doi.org/10.1007/s10552-016-0808-4.
14. Dwyer-Lindgren L, Bertozzi-Villa A, Stubbs RW, et al. Inequalities in life expectancy among US counties, 1980 to 2014: temporal trends and key drivers. JAMA Intern Med. 2017;177(7):1003–11. https://doi.org/10.1001/jamainternmed.2017.0918.
15. Wentzensen N, Poole EM, Trabert B, et al. Ovarian Cancer risk factors by histologic subtype: an analysis from the ovarian Cancer cohort consortium. J Clin Oncol. 2016;34(24):2888–98. https://doi.org/10.1200/JCO.2016.66.8178.
16. Sampson JN, Wheeler WA, Yeager M, Panagiotou O, Wang Z, Berndt SI, et al. Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. J Natl Cancer Inst. 2015;107(12):djv279. https://doi.org/10.1093/jnci/djv279. Erratum in: J Natl Cancer Inst. 2016;108(4). https://doi.org/10.1093/jnci/djw106.
17. Kreimer AR, Johansson M, Waterboer T, et al. Evaluation of human papillomavirus antibodies and risk of subsequent head and neck cancer. J Clin Oncol. 2013;31(21):2708–15. https://doi.org/10.1200/JCO.2012.47.2738.
18. Kreimer AR, Brennan P, Lang Kuhs KA, et al. Human papillomavirus antibodies and future risk of anogenital cancer: a nested case-control study in the European prospective investigation into cancer and nutrition study. J Clin Oncol. 2015;33(8):877–84. https://doi.org/10.1200/JCO.2014.57.8435.
19. Stone JB, DeAngelis LM. Cancer-treatment-induced neurotoxicity--focus on newer treatments. Nat Rev Clin Oncol. 2016;13(2):92–105. https://doi.org/10.1038/nrclinonc.2015.152.
20. American Cancer Society. Cancer Facts & Figures for African Americans 2019–*2021*. https://www.cancer.org/research/cancer-facts-statistics/cancer-facts-figures-for-african-americans.html
21. American Cancer Society. *Cancer Facts & Figures for Hispanics/Latinos 2018–2020*. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-facts-and-figures-for-hispanics-and-latinos/cancer-facts-and-figures-for-hispanics-and-latinos-2018-2020.pdf
22. Rowland JH, Gallicchio L, Mollica M, Saiontz N, Falisi AL, Tesauro G. Survivorship science at the NIH: lessons learned from Grants funded in fiscal

year 2016. J Natl Cancer Inst. 2019;111(2):109–17. https://doi.org/10.1093/jnci/djy208.

23. Miller KD, Siegel RL, Lin CC, et al. Cancer treatment and survivorship statistics, 2016. CA Cancer J Clin. 2016;66(4):271–89. https://doi.org/10.3322/caac.21349.
24. Donin N, Filson C, Drakaki A, et al. Risk of second primary malignancies among cancer survivors in the United States, 1992 through 2008. Cancer. 2016;122(19):3075–86. https://doi.org/10.1002/cncr.30164.
25. Albano JD, Ward E, Jemal A, et al. Cancer mortality in the United States by education level and race. J Natl Cancer Inst. 2007;99(18):1384–94. https://doi.org/10.1093/jnci/djm127.
26. Mariotto AB, Zou Z, Johnson CJ, Scoppa S, Weir HK, Huang B. Geographical, racial and socio-economic variation in life expectancy in the US and their impact on cancer relative survival. *PLoS One*. 2018;13(7):e0201034. Published 2018 Jul 25. https://doi.org/10.1371/journal.pone.0201034.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.