Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2017 June 22.

Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2016 October ; 25(10): 1392–1401. doi: 10.1158/1055-9965.EPI-16-0412.

The Cancer Epidemiology Descriptive Cohort Database: A Tool to Support Population-Based Interdisciplinary Research

Amy E. Kennedy[1], **Muin J. Khoury**[2], **John P.A. Ioannidis**[3,4,5,6], **Michelle Brotzman**[7], **Amy Miller**[7], **Crystal Lane**[8], **Gabriel Y. Lai**[1], **Scott D. Rogers**[1], **Chinonye Harvey**[1], **Joanne W. Elena**[1], and **Daniela Seminara**[9]

[1]Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, NCI, NIH, Rockville, Maryland [2]Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia [3]Department of Medicine, Stanford University, Stanford, California [4]Department of Health Research and Policy, Stanford University, Stanford, California [5]Department of Statistics, Stanford University, Stanford, California [6]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California [7]Westat, Rockville, Maryland [8]Office of Epidemiology and Research, Maternal and Child Health Bureau, Health Resources and Services Administration, Rockville, Maryland [9]Division of Cancer Control and Population Sciences, NCI, NIH, Rockville, Maryland

Abstract

Background—We report on the establishment of a web-based Cancer Epidemiology Descriptive Cohort Database (CEDCD). The CEDCD's goals are to enhance awareness of resources, facilitate interdisciplinary research collaborations, and support existing cohorts for the study of cancer-related outcomes.

Methods—Comprehensive descriptive data were collected from large cohorts established to study cancer as primary outcome using a newly developed questionnaire. These included an inventory of baseline and follow-up data, biospecimens, genomics, policies, and protocols. Additional descriptive data extracted from publicly available sources were also collected. This

Corresponding Author: Daniela Seminara, NCI, 9609 Medical Center Drive, Room 3E336, MSC 9763, Bethesda, MD 20850. Phone: 240-276-6748; Fax: 240-276-7921; seminard@mail.nih.gov.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors' Contributions
Conception and design: A.E. Kennedy, M.J. Khoury, J.P.A. Ioannidis, C. Lane, G.Y. Lai, C. Harvey, D. Seminara
Development of methodology: A.E. Kennedy, M.J. Khoury, J.P.A. Ioannidis, A. Miller, C. Lane, G.Y. Lai, D. Seminara
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): A.E. Kennedy, M. Brotzman, A. Miller, C. Lane, G.Y. Lai, C. Harvey, D. Seminara
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A.E. Kennedy, J.P.A. Ioannidis, C. Lane, G.Y. Lai, C. Harvey, D. Seminara
Writing, review, and/or revision of the manuscript: A.E. Kennedy, M.J. Khoury, J.P.A. Ioannidis, M. Brotzman, A. Miller, C. Lane, G.Y. Lai, S.D. Rogers, C. Harvey, J.W. Elena, D. Seminara
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): A.E. Kennedy, M. Brotzman, C. Lane, G.Y. Lai, C. Harvey, D. Seminara
Study supervision: D. Seminara

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

information was entered in a searchable and publicly accessible database. We summarized the descriptive data across cohorts and reported the characteristics of this resource.

Results—As of December 2015, the CEDCD includes data from 46 cohorts representing more than 6.5 million individuals (29% ethnic/racial minorities). Overall, 78% of the cohorts have collected blood at least once, 57% at multiple time points, and 46% collected tissue samples. Genotyping has been performed by 67% of the cohorts, while 46% have performed whole-genome or exome sequencing in subsets of enrolled individuals. Information on medical conditions other than cancer has been collected in more than 50% of the cohorts. More than 600,000 incident cancer cases and more than 40,000 prevalent cases are reported, with 24 cancer sites represented.

Conclusions—The CEDCD assembles detailed descriptive information on a large number of cancer cohorts in a searchable database.

Impact—Information from the CEDCD may assist the interdisciplinary research community by facilitating identification of well-established population resources and large-scale collaborative and integrative research.

Introduction

Understanding the determinants of cancer and other chronic diseases requires an in-depth knowledge of the complex interactions of genomic, biological, clinical, lifestyle, and societal factors (1). Cohort studies have helped researchers to better understand the complex etiology of cancer, study cancer outcomes, develop risk prediction analyses and models, and improve guidelines for cancer prevention and control policies (2). Combining risk factors and molecular data across cohorts has supported genomic, epigenomic, proteomic, and metabolomics research of unprecedented scope (3–7). It has been proposed that a “synthetic cohort,” achieved through a collaborative approach among the major cohorts funded by the NIH (Bethesda, MD), could expedite an integrative approach by assembling multilevel data collected through the lifespan, in health and disease status, and by providing a framework for interdisciplinary research (8–10). The value of existing and future cohorts has been greatly enhanced by expanding collaborations and studying multiple health outcomes through data sharing and pooling (10–12). Recently, it has been suggested that registration of observational studies may help improve the transparency and quality of epidemiologic research and reduce the extent of publication biases (10, 13–15). In contrast to randomized trials where single protocols are registered, for observational studies, it may be best to register unique cohorts’ infrastructures, so that they can be tapped to produce collaborative studies involving multidisciplinary teams (14, 16). A dynamic, comprehensive, and publicly accessible inventory of cohorts is fundamental to facilitate collaborative scientific efforts and cost-effective assembly and utilization of resources and will assist the research community and funding agencies in the planning of new studies and in maximizing the returns on investments.

The Epidemiology and Genomics Research Program (EGRP), of the NCI’s Division of Cancer Control and Population Sciences, fosters cohort-based research and the establishment of cohorts infrastructures through numerous initiatives and workshops (17–19). The NCI Cohort Consortium (19) was formed in 2000 to foster collaborations across

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

cancer epidemiology cohorts, by initially focusing on the need to assemble large populations for genome-wide association studies to investigate genetic risk factors in cancer (20, 21). The Cohort Consortium expanded its reach to address critical research that could not be addressed by single-research studies (22–24). Similarly, other consortia in the United States and worldwide have combined cohort infrastructures to address the etiology and genomics of complex diseases (25–29).

The NCI Cohort Consortium has had success in stimulating more than 50 ongoing and completed collaborative projects; however, many challenges still exist. These include barriers to systematic data harmonization, especially when considering the integration of multilevel datasets; lack of a data-coordinating center (30) to maximize data management and deployment; and consistent, streamlined, and comprehensive data sharing policies and processes (31). Paramount is the overall need for tools facilitating transparency, expediting scientific collaborations, enhancing the quality of reported research, and supporting interdisciplinary approaches within the framework of large epidemiologic studies (10). We report on the establishment and characteristics of the Cancer Epidemiology Descriptive Cohort Database (CEDCD), developed to address some of these challenges and facilitate interdisciplinary research collaborations. We review the descriptive data currently available in the CEDCD and discuss the scope and potential of the combined cancer cohorts' infrastructure.

Materials and Methods

Development of the descriptive data collection instruments

To populate the descriptive data included in the online database, a Data Collection Form and the Biospecimen and Cancer Count Information Spreadsheet (Supplementary Material S1 and S2) were developed. Both tools were assessed for the time and effort required for completion and received OMB clearance. These instruments were designed with consideration for study variation in an effort to promote standardized metrics across cohorts. To maintain participants' privacy given the publicly accessible nature of CEDCD, only descriptive data (i.e., no individual-level data) were collected. The principal investigators from each cohort provided consent to publicly share the cohorts' descriptive information through the CEDCD. The Data Collection Form requested descriptive data in the following categories: basic profile information, such as investigators, websites, and a brief description of the cohort; study design and eligibility criteria; enrollment counts by race/ethnicity/gender; major content domain data collected at baseline and follow-up; types of cancer and other disease outcomes; mortality data, incorporation of mobile health technologies in protocols, and types and counts of biospecimens collected. If accurate information in any of these domains was available from public or NCI sources, the form was prefilled and the cohorts were asked to review the data. The Biospecimen and Cancer Count Information Spreadsheet was included to provide incident and prevalent cancer numbers by gender, with cancers defined by ICD-9 and ICD-10/O coding, and biospecimen counts provided by cancer type. Supplementary information requested included cohorts' policies on data sharing and access, publication policies, and questionnaires. Review and editing of these tools as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

well as a pilot test to assess the time, effort, and ease in completing the forms were conducted by cohort investigators and affiliated staff.

Study selection and data collection

The 57 NCI Cohort Consortium members (as of January 2015) were invited to participate in the CEDCD. These cohorts met the following criteria: (i) focusing on the study of cancer as their primary outcome, and (ii) a minimum of 10,000 study participants currently enrolled, or (iii) cancer patient/survivor cohorts of at least 5,000 participants across multiple cancer sites or 2,000 participants diagnosed with the same or narrowly related cancer sites. The latter criteria was determined to support research addressing determinants of cancer progression, recurrence, mortality, incidence, and other cancer/health–related outcomes (32). The database will be supported by NCI's EGRP, and cohorts will be contacted annually to update existing information. Eligible cohorts can begin the process for inclusion at any time by completing an online request (https://cedcd.nci.nih.gov/contact.aspx).

Completed questionnaires were reviewed for discrepancies and missing data to ensure completeness and accuracy before posting to the web-based database. The website was designed to maximize users' abilities to search for cohorts by name, find detailed information about a single cohort, and compare information across cohorts. On the basis of feedback received from beta-testing, the website design was maximized for ease of navigation, and user-friendly help tips were added throughout the site to highlight different features.

Results

Characteristics of participating cohorts

The CEDCD website was launched in March 2015. As of December 2015, descriptive data from 46 of the 57 invited cohorts (81%) are available on the CEDCD website. Of the 11 not listed, five cohorts did not respond, four requested additional time, and two submitted partially completed forms. Participating cohorts were classified according to three categories: risk, survivor, and hybrid cohorts. Risk epidemiology cohorts, those enrolling healthy participants to be followed over time to detect cancer incidence, represented the majority (83%); hybrid cohorts (15%), which enrolled families including healthy individuals and cancer survivors for prospective follow-up, and one cohort (2%) enrolled only cancer survivors and followed them over time to study cancer-related outcomes and survivorship issues (Table 1). Review of the basic cohort characteristics, including the year enrollment began, and the most recent year data that were collected showed that, overall, this is a group of long-established cohorts, that began enrollment decades ago, with the average (and median) year for the start of enrollment being 1993 (Table 1). Six cohorts (13%) are still actively enrolling participants, while the most recent year of active follow-up ranged from 1974 to 2015. Given that the primary outcome of interest was cancer, most cohorts specifically enrolled adults, with the average minimum and maximum ages at enrollment being 34 and 79.5 years old, respectively. However, four cohorts enrolled participants under the age of 18 (Table 1).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Demographics

More than 6.5 million individuals have been enrolled by the 46 cohorts, with three times as many females enrolled compared with males (Table 2). More than half of the cohorts, 24 in total (52%), have enrolled both genders, while 15 (33%) enroll only females, and 7 (15%) enroll only males (Table 1). Overall, females outnumber males in 34 cohorts. Most cohorts (44) provided data on the race/ethnicities of their participants, with stratified enrollment numbers listed in Table 2. The majority of enrolled participants classify themselves as non-Hispanic whites (65%), followed by Asian (10%). More than 385,000 participants classified themselves as black or African American and 290,000 as Hispanic. In addition, more than 30,000 individuals reported belonging to more than one race.

Participants have been enrolled across three continents, from 17 different countries. The highest recruitment was from North America, with 31 cohorts having catchment areas in the United States and 8 in Canada. There are 13 cohorts with catchment areas in Europe, followed by Asia (6) and Australia (Supplementary Table S1; ref. 3).

Data collection

The database is capable of displaying the types of data collected by each cohort as well as summarizing descriptive data across the cohorts selected for comparison. The tables found under Table 3 show the categories of data collected at baseline, including risk factors, cancer outcomes, comorbidities, and cancer treatment received. The vast majority of cohorts reported collecting data on key risk factors including smoking/cigarette use (91%), alcohol use (84%), physical activity (78%), and dietary intake (78%) (Table 3A). Data collection on other major diseases, which is useful to investigators trying to utilize the cohort data across different complex disease phenotypes, includes diabetes and heart disease in 36 cohorts (78%; Table 3B), representing almost 4 million participants. More than 50% of the cohorts (n = 30 and n = 27, respectively) collected data on digestive and lung disease as well (Table 3B). The cohorts were not asked to specify whether the outcomes were collected through self-report of if they were determined through medical records. Cancer treatment data were collected by 27 cohorts (59%; Table 3C). The most common method to obtain treatment information was from patient-reported questionnaires (55%), followed by medical chart abstraction (44%), abstraction from electronic medical records (18%), and from administrative claims (14%; data not shown).

The majority of cohorts (76%) have utilized multiple data collection approaches, with most still using mail-in questionnaires (89%), followed by in-person interviews (50%). Interestingly, 41% of cohorts administered questionnaires electronically via the Internet, surpassing the number of studies that contact participants by telephone (24%). Two cohorts have adopted cloud-based approaches for the collection, management or distribution of their study data on the Internet, and another seven (15%) are considering using this technology within their cohort. With the evolution of mobile technologies, cohorts are moving toward using such novel approaches for data collection. Three (6.5%) cohorts have adopted data collection through mobile devices, but another 11 cohorts (24%) are considering the use of mobile devices. Two of the cohorts using mobile technologies reported using tablets, with

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

one specifying its use to obtain consent and collect questionnaires, while the other collected detailed specimen data during in-person blood collections.

Half (n = 23) of the cohorts link their data to other existing databases. The majority of the linkages are between state registries, SEER, Medicare, and other tumor and cancer registries. Almost all of the cohorts confirmed collecting mortality data (44 cohorts, 96%). The majority of cohorts confirm the death of participants by using linkage to the National Death Index (27 cohorts, 59%) or to state death certificates (28 cohorts, 51%) and to other registration systems, such as SEER, cancer registries, obituaries, and the Social Security Death Index.

A total of 634,801 incident cancer cases were reported by the risk epidemiology cohorts and 40,676 prevalent cancer cases by the hybrid and survivors cohorts. A breakdown of the cancer sites reported is listed in Table 4. The top five incident cancers are breast, prostate, lung, colon, and melanoma, while higher prevalent cases included breast, colon, rectum and anus, prostate, and cervical cancers. There were considerable numbers for the less common cancer, as well. For example, there were almost 8,000 incident cases of brain cancer and 22,944 for bladder cancer.

Biobanking, genomics, and other -omics

Blood samples were collected at least once on a subset of cohort participants by 36 cohorts (78%), while 24 cohorts (52%) did so at multiple time points (Table 5A). Types of biospecimens collected included blood, buccal, sputum, feces, lymphocytes, tumor tissue, and urine samples. Tumor tissues were collected by 46% of the cohorts, and normal tissue was available for 30% of the cohorts. For the studies that do not collect tumor tissue, 39% had knowledge of where the tumor tissue was stored for future studies (Table 5B).

The rise in molecular and genomic epidemiology studies is reflected in the biospecimen and molecular data collection numbers (Table 6), with 67% of the cohorts performing genotyping (SNPs) for genome-wide association studies, 46% implementing whole-exome or whole genome sequencing either on blood or tissues, and 50% collecting epigenetic/metabolic marker data. Although not all cohorts perform a systematic molecular and genomic characterization of cohort participants through high-throughput assays, these data reflect the feasibility of comprehensive “omic” characterization within large cohorts, which is currently limited by lack of resources and rapidly changing technologies.

Cohort-based research

A total of 75% of the cohorts (n = 36) report to have participated in cross-cohort data harmonization activities, a fact due at least in part to their participation in the NCI Cohort Consortium, and their involvement in many additional consortia and pooling projects.

CEDCD-based analyses in combination with data linkage to existing databases providing information related to cancer research have also been explored. A list of NIH-funded grants supporting the cohorts’ research was obtained by searching the NIH’s Query, View and Report database (conducted in May–June 2015). A text search, using each of the cohort’s name and acronym, was used to compile the list of grants for each of the cohorts. It should

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

be noted that there are limitations to the text search function when searching grant applications submitted prior to 2008, as a number of applications in the NIH grants database are scanned image files. Those grants that were funded by the NCI were further examined using the NCI's Portfolio Management Application (PMA) 16.1 database and the cancer activity (CA) code to determine the type of grant awarded and the research area addressed.

Thirty-seven cohorts were found to be supported by NIH-funded grants since 1997. Of the nine cohorts without funding, all but one, were international. Of the 37 funded cohorts, 29 were based in the United States, and eight were international. Considering the primary outcome of interest of the cohorts is cancer, it was not surprising that the majority (66%) of the NIH grants were funded by the NCI ($n = 407$). Other major funding institutes included the National Heart, Lung, and Blood Institute (50 grants), the National Institute of Diabetes and Digestive and Kidney Diseases (36 grants), the National Institute of Environmental Health Sciences (30 grants), and the National Institute on Aging (22 grants).

The PMA 16.1 database was used to categorize grants awarded by the NCI by type (Supplementary Fig. S1) and by cancer activity (CA) code (Supplementary Fig. S2). The five most common activity codes contained more than 60% of all of the NCI-funded grants and included Genomic Epidemiology (73 grants), Modifiable Risk Factors (Epidemiology; 70 grants), General Epidemiology (51 grants), Training (29 grants), Nutrition (24 grants), and Early Detection/Biomarkers (21 grants).

Discussion

The CEDCD facilitates collaborative research using data from cancer epidemiology cohort studies. Participation in the CEDCD is open to all eligible cohorts meeting the criteria as described previously. For those investigators interested in having their cohort be included in the database, an inquiry should be made through the contact page (https://cedcd.nci.nih.gov/contact.aspx). When accessing the database, users can select cohorts by searching via cohort name, selecting from the alphabetized list, or by choosing advanced criteria of interest. The "Help" icon enables users to get a step-by-step tutorial on how to explore all of the databases' functions.

Analysis of the CEDCD descriptive data for the major cohorts studying cancer and related outcomes shows a powerful population infrastructure and identifies areas for possible enhancement (33). Pooling projects that combine individual-level data across these cohorts to achieve large sample sizes are feasible, and collaborative studies that no single cohort alone can perform could be undertaken. These include studies involving population subgroups, rare cancers, rare genotypes, exposures, and phenotypes, and evaluating interactions between common genetic and nongenetic risk factors. Research on rare cancers typically relies on case–control studies because the sample sizes required exceed most prospective studies. However, the CEDCD can help investigators identify studies with the rare exposures and endpoints of interest to facilitate pooling projects. Furthermore, with the varied age range of the participants, large-scale research on early-onset cancers among certain ethnic and racial groups is possible. In addition, the assessment for other disease endpoints in the CEDCD cohorts could greatly facilitate the study of comorbidities,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

especially in elderly populations. Collaboration among cancer research teams and experts in other common disease domains (e.g., CVD and aging) is essential for this transdisciplinary work.

The study of cancer health disparities is a current priority for the NCI and NIH. Population-based research on health disparities has been hampered in the past by the lack of sufficient numbers of study participants reflecting diverse ethnic and racial composition (34, 35). The combined data from the CEDCD show that collaborative efforts could greatly help facilitate research addressing underserved ethnic and racial groups, given the collective number of cohorts' participants from the relevant populations, including Hispanics (one of the fastest growing ethnic groups in the United States), African Americans, and Asian populations.

The lifestyle data collected by the majority of the CEDCD cohorts reflect the major risk factors for cancer, as well as for cardiovascular disease and diabetes, which are among the 10 most prominent causes of mortality in high-income countries. The systematic addition of clinical risk factor and treatment data, particularly electronic medical records, is an important feature that could enhance and enrich the usefulness of the cohorts' data. Although the CEDCD data indicate that the process of data integration is under way and present tremendous resource for pooled analyses, it is also clear that a comprehensive standardized effort for multilevel data harmonization, management, and distribution is necessary (36).

In this initial survey of the use or willingness to use m-health technologies, we did not request further details on the specific technologies and approaches that the cohort intended to use to implement m-health. The use of mobile applications for data collection (e.g., physical activity, diet, and medication use), storage, and management will likely increase during the coming years, due in part to methodologic advances and efforts of the precision medicine initiative in the United States and others abroad. Communication across cohorts and a forum to define the most successful m-health protocols to acquire accurate information from each subpopulation and the adoption of cross-cohort standardization for these rapidly evolving tools and protocols are key to support more efficient and cost-effective next-generation collaborative studies. These trends could be monitored in future editions of the CECDC.

Extensive genomic and other 'omics characterization has already been performed on a large subset of the cohorts' biospecimens. A collaborative approach to systematic integration of the existing genomics data has enabled the development of studies large enough to address the complex nature of the genetic factors underlying common diseases. This approach has been adopted by international OncoArray Network (37), an NCI, Genome Canada, and Cancer Research UK initiative, which includes as members many of the CEDCD participating cohorts, enabling the discovery of additional common and rarer susceptibility variants (37) Similarly, the integration of 'omics, lifestyle, functional and clinical data derived from cohorts infrastructure has been initiated by large NCI-sponsored initiatives, such as the Genetic Associations and Mechanisms in Oncology and by multiple NCI-sponsored Consortia (add cohort consortium (17, 19), showing the feasibility of combining

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

epidemiologic, genomics, functional, and clinical data to support a new generation of integrative epidemiology analyses.

The CEDCD cohorts ascertain extremely high numbers of incident common cancers. More than half of the cohorts have collected blood at multiple time points, and most have collected or have the capability to collect tissues, empowering research in disentangling the heterogeneity of molecular heterogeneity within cancer sites and the validation of predictive biomarkers as well as the relationship between somatic and germline genomes. However, the relative lack of systematic tissue and DNA collection across cohorts handicaps the ability to pool molecular data derived from these samples for larger studies.

Given the overall large numbers of incident cancers, the extensive follow-up data, the enrollment of diverse populations and the rich biospecimen collections, these cohorts could provide the infrastructure to enable build-in clinical trials seeking individuals with unique characteristics within a cohort population who may benefit from precision medicine approaches (38, 39). Moreover, they can be used to embed clinical trials of interventions on asymptomatic individuals (40). These cohorts could also support studies of late effects of cancer therapy and related comorbidities, which are not easily explored in the time frame of a clinical trial. To function, this translational pipeline would require the seamless integration of multidisciplinary teams, including epidemiologists, molecular scientists, behavioral scientists, and clinicians, a transition already in process in many of the NCI-funded cancer centers.

In summary, we have created a unique and detailed inventory of existing cohorts with cancer as a primary outcome. Examination of the CEDCD initial data show that the combined cancer cohorts examined could provide access to the broadest range of genotypes, phenotypes, and exposures, thereby accelerating efforts to detect and analyze subtle and important signals with greater accuracy and inform precision medicine. Increasing the number and variety of cohorts enrolled would provide valuable data for investigators from multiple disciplinary domains and facilitate collaborative study of cancer-related endpoints. Continuing this initial effort by maintaining a current and comprehensive descriptive database of large prospective cohorts will facilitate important epidemiologic research by allowing the identification of areas of strength and needs for the current overall cohorts' infrastructure, and by informing the use of resources through cost-effective planning by both investigators and funding agencies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to acknowledge all of the participating cohorts. A special thanks to the program managers and study staff for taking the time to complete the data collection forms and to the principal investigators of each of the cohort studies:

Seventh-day Adventist Cohort Study-2: Gary E. Fraser, Synnove Knutsen, and Karen Jaceldo-Siegl (Loma Linda University School of Public Health), and Michael Orlich (Loma Linda University School of Medicine)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Alpha-Tocopherol Beta-Carotene Cancer Prevention Study: Demetrius Albanes and Stephanie Weinstein (NCI)

British Columbia Generations Project: John J. Spinelli, Angela R. Brooks-Wilson, Tim K. Lee, and Nhu D. Le (BC Cancer Agency)

Breast Cancer Family Registry Cohort: Esther M. John (Cancer Prevention Institute of California), Mary Beth Terry (Columbia University), Irene A. Andrulis (Mount Sinai Hospital), Mary Daly (Fox Chase Cancer Center), Saundra S. Buys (University of Utah Health Sciences Center), and John L. Hopper (The University of Melbourne)

Breast Cancer Surveillance Consortium Research Resource: Ellen O'Meara (Group Health Research Institute)

Breakthrough Generations Study: Anthony Swerdlow (Institute of Cancer Research)

A Follow-up Study for Causes of Cancer in Black Women: Black Women's Health Study: Lynn Rosenberg and Julie Palmer (Boston University), and Lucile Adams-Campbell (Georgetown University)

Carotene and Retinol Efficacy Trial: Gary Goodman, Mark Thornquist, and Marian Neuhouser (Fred Hutchinson Cancer Research Center)

CARTaGENE: Philip Awadalla (CARTaGENE, St-Justine Hospital)

Clue Cohort Study I and II: Kala Visvanathan, Josef Coresh, Corrine Joshu, and Elizabeth Platz (Johns Hopkins University)

Colon Cancer Family Registry Cohort: Robert W. Haile (Stanford University), Mark Jenkins (University of Melbourne), Noralane Lindor (Mayo School of Medicine, Scottsdale), Steven Gallinger (Mount Sinai Hospital), Loic Le Marchand (University of Hawaii), Polly A. Newcomb (Fred Hutchinson Cancer Research Center), Dennis Ahnen (University of Colorado, Denver), Kristen Anton (Geisel School of Medicine at Dartmouth), Graham Casey (University of Southern California), Iona Cheng (Cancer Prevention Institute of California), James Church (Cleveland Clinic Digestive Disease Institute), and Timothy Church (University of Minnesota)

Cancer Prevention Study II Nutrition Cohort: Susan Gapstur (American Cancer Society)

Canadian Study of Diet, Lifestyle and Health: Thomas Rohan (Albert Einstein College of Medicine) and Vicki Kirsh (Cancer Care Ontario)

California Teachers Study: Leslie Bernstein and James Lacey (City of Hope)

European Prospective Investigation in Cancer and Nutrition: Elio Riboli (Imperial College, London)

Golestan Cohort Study: Christian Abnet and Sanford Dawsey (NCI), Paolo Boffetta (Mount Sinai), Paul Brennan (IARC), Farin Kamangar (Morgan State University), and Reza Malekzadeh (Digestive Diseases Research Institute)

Health Professionals Follow-up Study: Walter Willett, Eric Rimm, Donna Spiegelman, and Meir Stampfer (Harvard School of Public Health)

Iowa Women's Health Study: Kim Robien (George Washington University), DeAnn Lazovich (University of Minnesota)

Janus Serum Bank: Giske Ursin and Hilde Langseth (Cancer Registry of Norway)

Mexican American (Mano a Mano) Cohort: Xifeng Wu, Hua Zhao, and Wong-Ho Chow (MD Anderson Cancer Center)

The Melbourne Collaborative Cohort Study: Graham Giles and Roger Milne (Cancer Council Victoria), Dallas English and John Hopper (University of Melbourne)

Multiethnic Cohort: Loic Le Marchand and Lynne Wilkens (University of Hawaii), Christopher Haiman (University of Southern California)

Mayo Mammography Health Study: Celine M. Vachon (Mayo Clinic)

Nurses' Health Study: Meir Stampfer (Harvard School of Public Health), Wendy Chen, Vincent Carey, and Diane Feskanich (Harvard Medical School)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Nurses' Health Study II: Walter C. Willett and Donna Spiegelman (Harvard School of Public Health), Heather Eliassen and Rulla Tamimi (Brigham and Women's Hospital)

The National Institutes of Health AARP Diet and Health Study: Rashmi Sinha, Charles E. Matthews, Louise Brinton, and Linda Liao (NCI)

NYU Women's Health Study: Anne Zeleniuch-Jacquotte (NYU School of Medicine)

Ontario Health Study: Mark Purdue (Ontario Health Study)

Prostate Cancer Prevention Trial: Catherine Tangen and Michael LeBlanc (Fred Hutchinson Cancer Research Center), Ian Thompson (University of Texas Health Science Center)

Physicians' Health Study I and II: J. Michael Gaziano and Howard D. Sesso (Brigham & Women's Hospital/ Harvard Medical School)

Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial: Paul Pinsky and Neal Freedman (NCI)

RERF Life Span Study: Kotaro Ozasa and Eric Grant (Radiation Effects Research Foundation)

Southern Community Cohort Study: William Blot and Wei Zheng (Vanderbilt University Medical Center), Margaret Hargreaves (Meharry Medical College)

Singapore Chinese Health Study: Jian-Min Yuan and Lesley Butler (University of Pittsburgh), Woon-Puay Koh (National University of Singapore)

Shanghai Cohort Study: Jian-Min Yuan and Lesley Butler (University of Pittsburgh), Yu-Tang Gao (Shanghai Cancer Institute)

Selenium and Vitamin E Cancer Prevention Trial: Catherine Tangen and Michael LeBlanc (Fred Hutchinson Cancer Research Center), Ian Thompson (University of Texas Health Science Center)

Sister Study: Dale P. Sandler, Jack Taylor, Clarice R. Weinberg, (National Institute of Environmental Health Sciences)

Shanghai Men's Health Study: Xiao O. Shu, Wei Zheng, Gong Yang (Vanderbilt University Medical Center), Yong-Bing Xiang (Shanghai Cancer Institute)

Shanghai Women's Health Study: Xiao O. Shu, Wei Zheng, Gong Yang (Vanderbilt University Medical Center), Yu-Tang Gao (Shanghai Cancer Institute)

VITamins And Lifestyle: Emily White and Ulrike Peters (Fred Hutchinson Cancer Research Center)

Women's Health Initiative: Garnet Anderson (Fred Hutchinson Cancer Research Center)

Women's Health Initiative Cancer Survivor Cohort: Garnet Anderson (Fred Hutchinson Cancer Research Center), Electra Paskett (Ohio State University), Bette Caan (Kaiser Permanente Northern California), and Rowan Chlebowski (University of California, Los Angeles)

Women's Health Study: Julie E. Buring, I-Min Lee, Nancy Cook, and Dan Chasman (Brigham and Women's Hospital)

Women's Lifestyle and Health: Elisabete Weiderpass Vainio (Karolinska Institutet, Stockholm, Sweden)

The authors would also like to thank Dr. Kathy Helzlsouer for her comments and edits to the manuscript.

References

1. Lauer MS. Time for a creative transformation of epidemiology in the United States. JAMA. 2012; 308:1804–5. [PubMed: 23117782]
2. Potter JD. Epidemiology informing clinical practice: from bills of mortality to population laboratories. Nat Clin Pract Oncol. 2005; 2:625–34. [PubMed: 16341118]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, et al. Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet. 2008; 40:310–5. [PubMed: 18264096]
4. Cox DG, Blanche H, Pearce CL, Calle EE, Colditz GA, Pike MC, et al. A comprehensive analysis of the androgen receptor gene and risk of breast cancer: results from the National Cancer Institute Breast and Prostate Cancer Cohort Consortium (BPC3). Breast Cancer Res. 2006; 8:R54. [PubMed: 16987421]
5. Lynch SM, Vrieling A, Lubin JH, Kraft P, Mendelsohn JB, Hartge P, et al. Cigarette smoking and pancreatic cancer: a pooled analysis from the pancreatic cancer cohort consortium. Am J Epidemiol. 2009; 170:403–13. [PubMed: 19561064]
6. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet. 2007; 39:645–9. [PubMed: 17401363]
7. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet. 2009; 41:986–90. [PubMed: 19648918]
8. Boerwinkle E. Translational genomics is not a spectator sport: a call to action. Genet Epidemiol. 2012; 36:85–7. [PubMed: 22851471]
9. Willett WC, Blot WJ, Colditz GA, Folsom AR, Henderson BE, Stampfer MJ. Merging and emerging cohorts: not worth the wait. Nature. 2007; 445:257–8. [PubMed: 17230171]
10. Khoury MJ, Lam TK, Ioannidis JP, Hartge P, Spitz MR, Buring JE, et al. Transforming epidemiology for 21st century medicine and public health. Cancer Epidemiol Biomarkers Prev. 2013; 22:508–16. [PubMed: 23462917]
11. Khoury, MJ., Wei, G. The future of epidemiology in the age of precision medicine: cancer, cardiovascular disease, and beyond. Atlanta, GA: CDC Genomics and Health Impact Blog; 2015. Available from: http://blogs.cdc.gov/genomics/2015/08/11/the-future-of-epidemiology/
12. Roger VL, Boerwinkle E, Crapo JD, Douglas PS, Epstein JA, Granger CB, et al. Strategic transformation of population studies: recommendations of the working group on epidemiology and population sciences from the National Heart, Lung, and Blood Advisory Council and Board of External Experts. Am J Epidemiol. 2015; 181:363–8. [PubMed: 25743324]
13. Bracken MB. Preregistration of epidemiology protocols: a commentary in support. Epidemiology. 2011; 22:135–7. [PubMed: 21293203]
14. Dal-Re R, Ioannidis JP, Bracken MB, Bufffer PA, Chan AW, Franco EL, et al. Making prospective registration of observational research a reality. Sci Transl Med. 2014; 6:224cm1.
15. Samet JM. To register or not to register. Epidemiology. 2010; 21:610–1. [PubMed: 20657292]
16. Rifai N, Bossuyt PM, Ioannidis JP, Bray KR, McShane LM, Golub RM, et al. Registering diagnostic and prognostic trials of tests: is it the right thing to do? Clin Chem. 2014; 60:1146–52. [PubMed: 24855278]
17. National Cancer Institute. Genetic Associations and Mechanisms in Oncology (GAME-ON): a network of consortia for Post-Genome Wide Association (Post-GWA) research. Available from: http://epi.grants.cancer.gov/gameon/
18. National Cancer Institute. Cancer Epidemiology Consortia. Available from: http://epi.grants.cancer.gov/Consortia
19. National Cancer Institute. NCI Cohort Consortium. Available from: http://epi.grants.cancer.gov/Consortia/cohort.html
20. Hunter DJ, Riboli E, Haiman CA, Albanes D, Altshuler D, Chanock SJ, et al. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. Nat Rev Cancer. 2005; 5:977–85. [PubMed: 16341085]
21. Kraft P, Pharoah P, Chanock SJ, Albanes D, Kolonel LN, Hayes RB, et al. Genetic variation in the HSD17B1 gene and risk of prostate cancer. PLoS Genet. 2005; 1:e68. [PubMed: 16311626]
22. Gallicchio L, Helzlsouer KJ, Chow WH, Freedman DM, Hankinson SE, Hartge P, et al. Circulating 25-hydroxyvitamin D and the risk of rarer cancers: design and methods of the Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. Am J Epidemiol. 2010; 172:10–20. [PubMed: 20562188]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

23. Birmann BM, Neuhouser ML, Rosner B, Albanes D, Buring JE, Giles GG, et al. Prediagnosis biomarkers of insulin-like growth factor-1, insulin, and interleukin-6 dysregulation and multiple myeloma risk in the Multiple Myeloma Cohort Consortium. Blood. 2012; 120:4929–37. [PubMed: 23074271]
24. Trabert B, Ness RB, Lo-Ciganic WH, Murphy MA, Goode EL, Poole EM, et al. Aspirin, nonaspirin nonsteroidal anti-inflammatory drug, and acetaminophen use and risk of invasive epithelial ovarian cancer: a pooled analysis in the Ovarian Cancer Association Consortium. J Natl Cancer Inst. 2014; 106:djt431. [PubMed: 24503200]
25. GIANT consortium. Available from: https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium
26. The Asia Cohort Consortium. Available from: https://www.asiacohort.org/
27. European Network of Genomic and Genetic Epidemiology (ENGAGE). Available from: http://www.euengage.org/
28. Type 1 Diabetes Genetics Consortium (T1DGC). Available from https://www.niddkrepository.org/studies/t1dgc/
29. Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO). Available from: https://www.fredhutch.org/en/labs/phs/projects/cancer-prevention/projects/gecco.html
30. National Cancer Institute. Novel approaches and challenges to data harmonization: maximizing the use of multi-level data in collaborative studies. Available from: http://epi.grants.cancer.gov/events/data-harmonization/
31. National Cancer Institute. NCI Workshop on broadening epidemiologic data sharing. Available from: http://epi.grants.cancer.gov/events/datasharing2014/
32. National Institutes of Health. Core Infrastructure and Methodological Research for Cancer Epidemiology Cohorts (U01). Available from: http://grants.nih.gov/grants/guide/pa-files/PAR-15-104.html
33. National Institutes of Health Precision Medicine Initiative Working Group. The Precision Medicine Initiative Cohort Program – Building a research foundation for 21st century medicine. Available from: https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf
34. Khoury MJ, Iademarco MF, Riley WT. Precision public health for the era of precision medicine. Am J Prev Med. 2016; 50:398–401. [PubMed: 26547538]
35. Khoury MJ, Evans JP. A public health perspective on a national precision medicine cohort: balancing long-term knowledge generation with early health benefit. JAMA. 2015; 313:2117–8. [PubMed: 26034952]
36. Rolland B, Reid S, Stelling D, Warnick G, Thornquist M, Feng Z, et al. Toward rigorous data harmonization in cancer epidemiology research: one approach. Am J Epidemiol. 2015; 182:1033–8. [PubMed: 26589709]
37. National Cancer Institute. OncoArray Network. Available from: http://epi.grants.cancer.gov/oncoarray/
38. Lawler M, Sullivan R. Personalised and precision medicine in cancer clinical trials: panacea for progress or Pandora's box? Public Health Genomics. 2015; 18:329–37. [PubMed: 26555236]
39. Schwaederle M, Zhao M, Lee JJ, Eggermont AM, Schilsky RL, Mendelsohn J, et al. Impact of precision medicine in diverse cancers: a meta-analysis of phase II clinical trials. J Clin Oncol. 2015; 33:3817–25. [PubMed: 26304871]
40. Ioannidis JP, Adami HO. Nested randomized trials in large cohorts and biobanks: studying the health effects of lifestyle factors. Epidemiology. 2008; 19:75–82. [PubMed: 18090999]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Characteristics of cohorts included in database (as of August 19, 2015)[a]

Cohort name[b]	Abbreviation	Cohort design type[b]	Year enrollment began	Currently enrolling (Y/N)	Most recent year data collected	Enrollment		Participant ages		Website
						Male	Female	Minimum	Maximum	
Agricultural Health Study	AHS	Risk	1993	N	2014	55,967	33,689	30	64	Yes
Seventh-day Adventist Cohort Study-2	AHS-2	Risk	2002	N	2015	33,643	62,311	30	111	Yes
Alpha-Tocopherol Beta-Carotene Cancer Prevention Study	ATBC	Risk	1985	N	1993	29,133	N/A	50	69	Yes
British Columbia Generations Project	BC Generations	Hybrid	2009	Y	2015	9,298	20,306	35	69	Yes
Breast Cancer Family Registry Cohort	BCFR Cohort	Hybrid	1992	N	2014	6,094	30,126	15	101	Yes
Breast Cancer Surveillance Consortium Research Resource	BCSC	Risk	1994	N	2009	N/A	2,326,132	18	109	Yes
Breakthrough Generations Study	BGS	Risk	2003	NR[a]	2015	N/A	113,000	16	102	Yes
A Follow-up Study for Causes of Cancer in Black Women: Black Women's Health Study	BWHS	Risk	1995	Y	2013	N/A	59,000	21	69	Yes
Carotene and Retinol Efficacy Trial	CARET	Risk	1985	N	2005	12,025	6,289	45	74	Yes
CARTaGENE	CARTaGENE	Hybrid	2008	N	2015	19,250	22,992	40	69	Yes
Clue Cohort Study- Clue I	CLUE I	Risk	1974	N	1974	10,952	15,193	12	97	Yes
Clue Cohort Study- Clue II	CLUE II	Risk	1989	N	2007	14,171	18,723	2	99	Yes
Colon Cancer Family Registry Cohort	Colon CFR	Hybrid	1997	N	2015	16,862	20,698	18	75	Yes
Cancer Prevention Study II Nutrition Cohort	CPS-II Nutrition	Risk	1992	N	2013	86,402	97,783	40	93	Yes
Canadian Study of Diet, Lifestyle and Health	CSDLH	Risk	1992	N	1999	34,291	39,618	21	100+	No
California Teachers Study	CTS	Risk	1995	N	2014	N/A	133,479	22	104	Yes
European Prospective Investigation in Cancer and Nutrition	EPIC	Risk	1992	N	2015	153,457	366,521	35	70	Yes
Golestan Cohort Study	GCS	Risk	2004	N	2012	21,234	28,811	40	75	Yes

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Cohort name[b]	Abbreviation	Cohort design type[b]	Year enrollment began	Currently enrolling (Y/N)	Most recent year data collected	Enrollment		Participant ages		Website
						Male	Female	Minimum	Maximum	
Health Professionals Follow-up Study	HPFS	Risk	1986	N	2014	51,529	N/A	40	75	Yes
Iowa Women's Health Study	IWHS	Risk	1986	N	2004	N/A	41,836	55	69	Yes
Janus Serum Bank	Janus	Risk	1972	N	2004	166,137	152,491	18	68	Yes
Mexican American (Mano a Mano) Cohort	MAC	Hybrid	2001	Y	2014	5,540	18,898	18	94	Yes
The Melbourne Collaborative Cohort Study	MCCS	Risk	1990	N	2007	17,045	24,469	27	76	Yes
Multiethnic Cohort	MEC	Risk	1993	N	2015	96,810	118,441	45	75	Yes
Mayo Mammography Health Study	MMHS	Risk	2003	N	2006	N/A	17,639	35	94	Yes
Nurses' Health Study	NHS	Risk	1976	N	2014	N/A	121,701	30	55	Yes
Nurses' Health Study II	NHSII	Risk	1989	N	2013	N/A	116,430	25	42	Yes
The National Institutes of Health AARP Diet and Health Study	NIH-AARP	Risk	1995	N	2006	339,666	226,732	50	71	Yes
NYU Women's Health Study	NYUWHS	Risk	1985	N	2010	N/A	14,274	35	65	Yes
Ontario Health Study	OHS	Hybrid	2010	Y	2015	84,505	127,227	18	99	Yes
Prostate Cancer Prevention Trial	PCPT	Risk	1994	N	2001	18,880	N/A	55	86	Yes
Physicians' Health Study I and II	PHS I & II	Risk	1982	N	2014	29,071	N/A	40	88	Yes
Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial	PLCO	Risk	1993	N	2014	75,521	73,161	55	74	Yes
RERF Life Span Study (Adult Health Study Clinical Subcohort) & F1 Cohort (Hiroshima and Nagasaki)	RERF	Risk	1950	N	2010	50,175	70,146	0	>60	Yes
Southern Community Cohort Study	SCCS	Risk	2002	N	2015	34,203	50,366	40	79	Yes
Singapore Chinese Health Study	SCHS	Risk	1993	Y	2013	27,954	35,303	45	74	No
Shanghai Cohort Study	SCS	Risk	1986	Y	2013	18,244	N/A	45	64	No
Selenium and Vitamin E Cancer Prevention Trial	SELECT	Risk	2001	N	2014	34,897	N/A	50	93	Yes
Sister Study	SIS	Risk	2003	N	2014	N/A	50,884	35	74	Yes
Shanghai Men's Health Study	SMHS	Risk	2001	N	2014	61,491	N/A	40	74	Yes

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2017 June 22.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Cohort name[b]	Abbreviation	Cohort design type[b]	Year enrollment began	Currently enrolling (Y/N)	Most recent year data collected	Enrollment		Participant ages		Website
						Male	Female	Minimum	Maximum	
Shanghai Women's Health Study	SWHS	Risk	1997	N	2014	N/A	74,941	40	70	Yes
VITamins And Lifestyle	VITAL	Risk	2000	N	2002	37,393	40,345	50	76	Yes
Women's Health Initiative	WHI	Risk	1994	N	2014	N/A	161,808	50	79	Yes
Women's Health Initiative Cancer Survivor Cohort	WHI-CSC	Survivor	2013	N	2013	N/A	8,992	50	94	Yes
Women's Health Study	WHS	Risk	1992	N	2014	N/A	39,876	39	90	Yes
Women's Lifestyle and Health	WLH	Hybrid	1991	N	2003	N/A	49,258	30	50	Yes

Abbreviations: N/A, not applicable; cohort not enrolling gender; NR, not reported.

[a] Data retrieved from The Cancer Epidemiology Descriptive Cohort Database. The Epidemiology and Genomics Research Program. NCI (https://cedcd.nci.nih.gov; updated July 6, 2015; accessed December 17, 2015).

[b] Cohort design types defined as follows: Risk epidemiology cohort is defined as a cohort enrolling healthy participants and follows them over time to see whether they acquire a disease/outcome; Survivor cohort is defined as a cohort enrolling participants diagnosed with cancer and follows them over time; Hybrid cohort is defined as a cohort that enrolls both healthy individuals and cancer survivors and follows them over time

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Total cohort enrollment numbers[a]

Gender (n = 46)	
Female	5,029,889
Male	1,651,840
Race/ethnicity (n = 44)[†]	
White (non-Hispanic)	3,874,651
Asian	586,781
Unknown or not reported	506,452
Black or African American	385,813
White (unknown ethnicity)	231,181
Hispanic (white)	207,870
Hispanic (other)	86,988
American Indian/Alaska Native	45,060
More than one race	33,935
Native Hawaiian or Other Pacific Islander	18,444

[a]Two cohorts only provided gender enrollment numbers and did not provide race/ethnicity numbers.

[†]Race and ethnicity standards are set by the Office of Management and Budget (OMB).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Cohort's data domains

A. Risk factor data collected at baseline	
	Percent of cohorts collecting data points[a]
Cigarette use/smoking status	91%
Alcohol consumption	84%
Family history of cancer	80%
Dietary supplement use	78%
Physical activity	78%
Dietary intake	78%
Reproductive history	70%
Nonprescription medication use	61%
Prescription medication use	60%
Environmental or occupational exposures	39%
B. Non-cancer outcome data	
Diabetes	78%
Heart and vascular diseases	78%
Digestive diseases	65%
Lung diseases	59%
Osteoporosis/bone-related conditions	39%
Autoimmune diseases	37%
Neurodegenerative disorders/mental health illnesses	32%
C. Treatment data	
Cancer treatment data (any type)	59%
Surgery	52%
Radiation	52%
Chemotherapy	52%
Hormonal therapy	43%
Bone marrow/stem cell transplant	11%

[a]Cohorts without a response to data collection queries were assumed to be missing these data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Incident and prevalent cancer counts[†]

Cancer type	Incident cancer counts[z]		Prevalent cancer counts[z]
Breast	210,465	Breast	16,034
Prostate	97,727	Colon	7,459
Lung (trachea, bronchus)	62,035	Rectum and anus	3,246
Colon	43,395	Prostate	2,954
Melanoma	25,672	Cervix	2,479
Bladder	22,944	Thyroid	1,219
Lymphoma (HL and NHL)	20,829	Melanoma	1,084
Corpus, body of uterus	18,081	Lymphoma (HL and NHL)	1,066
Rectum and anus	16,638	Corpus, body of uterus	962
Ovary, fallopian tube, broad ligament	15,820	Ovary, fallopian tube, broad ligament	846
Pancreas	14,536	Bladder	668
Kidney (including renal pelvis, ureter, urethra)	12,730	Lung (trachea, bronchus)	640
Leukemia	12,185	Kidney (including renal pelvis, ureter, urethra)	518
Stomach	10,566	Leukemia	434
Oropharyngeal	8,852	Brain	261
Brain	7,962	Oropharyngeal	201
Thyroid	7,261	Liver and intrahepatic bile ducts	168
Myeloma	6,375	Stomach	134
Liver and intrahepatic bile ducts	6,274	Bone	82
Esophagus	4,985	Pancreas	75
Cervix	3,874	Esophagus	64
Gall bladder and extrahepatic bile duct	3,147	Myeloma	48
Small intestine	1,924	Small intestine	30
Bone	524	Gall bladder and extrahepatic bile duct	4
Total number of cancer cases	634,801	**Total number of cancer cases**	40,676

Abbreviations: HL, Hodgkin lymphoma; NHL, non-Hodgkin lymphoma.

[a] Total counts, from all cohorts providing data.

[†] An incident cancer was defined as a cancer diagnosed after enrollment into a cohort.

[‡] A prevalent cancer was defined as a cancer diagnosed prior to enrollment into a cohort.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

Specimen data

A. Specimen collection

Specimen type	Percent of cohorts collecting specimen type
Blood collected (once)	78%
Blood collected (at multiple time points)	52%
Saliva/buccal collected (once)	46%
Saliva/buccal (at multiple time points)	4%
Tumor tissue collected	46%
Able to potentially retrieve samples for future studies (if not currently collecting tissue)	39%
Normal tissue collected	30%

B. Specimen counts[a]

Specimen type[†]	Prevalent and incident cancer cases	Healthy individuals
Blood	293,859	1,184,636
Buccal[b]	44,715	246,369
Lymphocytes	17,501	83,120
Tumor tissue: fresh/frozen	485	—
Tumor tissue: FFPE	41,181	—
Sputum	70	3,025
Urine	46,362	344,489
Buffy coat	4,175	49,108
Guthrie card	0	29,378
DNA	249,422	640,707

Abbreviation: FFPE, formalin fixed paraffin embedded.

[a] Of the 46 cohorts included in the database, 5 reported they had no biospecimens and 2 did not respond to this query,

[†] Individual specimen type and DNA are not mutually exclusive categories.

[b] Buccal includes all samples reported as buccal, mouthwash/buccal, and cheek cells.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 6

Molecular data

	Percent of cohorts performing molecular test
Genotyping data (SNP)	67%
Epigenetic or metabolic markers	50%
Exome sequencing data	22%
Whole-genome sequencing data	26%
Other “omics” data	28%