International Journal of Epidemiology, 2016, 713–713f
doi: 10.1093/ije/dyv317
Advance Access Publication Date: 20 December 2015
Cohort Profile

Cohort Profile

Cohort Profile: The MD Anderson Cancer Patients and Survivors Cohort (MDA-CPSC)

Xifeng Wu,[1]* Michelle AT Hildebrandt,[1] Yuanqing Ye,[1] Wong-Ho Chow,[1] Jian Gu,[1] Sonia Cunningham,[1] Hua Zhao,[1] Ernest T Hawk,[2] Elizabeth Wagar,[3] Alma Rodriguez[4] and Stanley R Hamilton[5]

Departments of: [1]Epidemiology, [2]Clinical Cancer Prevention, [3]Laboratory Medicine, [4]Lymphoma and Myeloma and [5]Pathology, University of Texas MD Anderson Cancer Centre, Houston, TX, USA

*Corresponding author. Department of Epidemiology, Unit 1340, University of Texas MD Anderson Cancer Centre, 1515 Holcombe Blvd, Houston, TX 77030, USA. E-mail: xwu@mdanderson.org

Accepted 30 October 2015

Why was the cohort set up?

Cancer patients with similar clinical characteristics differ widely in their outcomes. Recognizing this, the era of personalized medicine strives to better predict clinical outcomes and guide individualized treatment paradigms. In parallel, the trend towards earlier diagnosis along with recent advances in cancer therapy have translated into decreasing mortality.[1] It is predicted that over the next 10 years, the number of Americans alive with a history of cancer will increase over 30% to reach nearly 19 million.[2] Recognizing that cancer survivors have unique medical, lifestyle and psychosocial needs, evidence-based guidelines are urgently required to support this expanding population.

Studies of genetic and molecular profiles can lead to discovery of powerful predictors of clinical outcomes and survivorship endpoints. However, such studies require large sample sizes with well-annotated clinical and pathological characteristics, relatively homogeneous treatment modalities and crucial follow-up data. Compared with the number of well-established prospective cohorts that focus on cancer aetiology,[3–10] there are few large cohorts for assessing clinical outcomes in cancer patients and quality of life (QOL) in survivors.[11,12] There is also a critical need for cancer patient cohorts capable of addressing health disparity issues related to cancer outcomes and QOL among minority and underserved populations.

The MDA-CPSC was developed specifically to support well-powered studies of clinical outcomes and survivorship in a racially diverse and well-characterized population of newly diagnosed patients. The cohort brings together the Patient History Database (PHDB) with core epidemiology data, the Blood Biospecimen Research Resource (BSRR) for biospecimens and our institutional electronic health record (EHR) and Tumour Registry that houses clinical, pathological, laboratory testing, treatment and follow-up data (Figure 1). The cohort is designed to support comprehensive research that will broadly advance the goal of personalized medicine and evidence-based survivorship care. Potential areas of investigation include: (i)discovering, testing and validating promising epidemiological determinants, intermediate clinical phenotypes and blood-based biomarkersfor the prediction of clinical outcomes and survivorship endpoints; (ii) studying the effect of baseline QOL on clinical endpoints; (iii) investigating the prevalence, severity and treatment of symptoms, discovering the underlying mechanisms of these symptoms and improving symptom management through evidence-based clinical trials; (iv) investigating racial disparities after cancer diagnosis and identifying contributors to the observed disparities; (v) studying outcomes, QOL and symptoms of patients with rare cancers; and (vi) developing integrative risk prediction algorithms for clinical outcomes and survivorship endpoints.

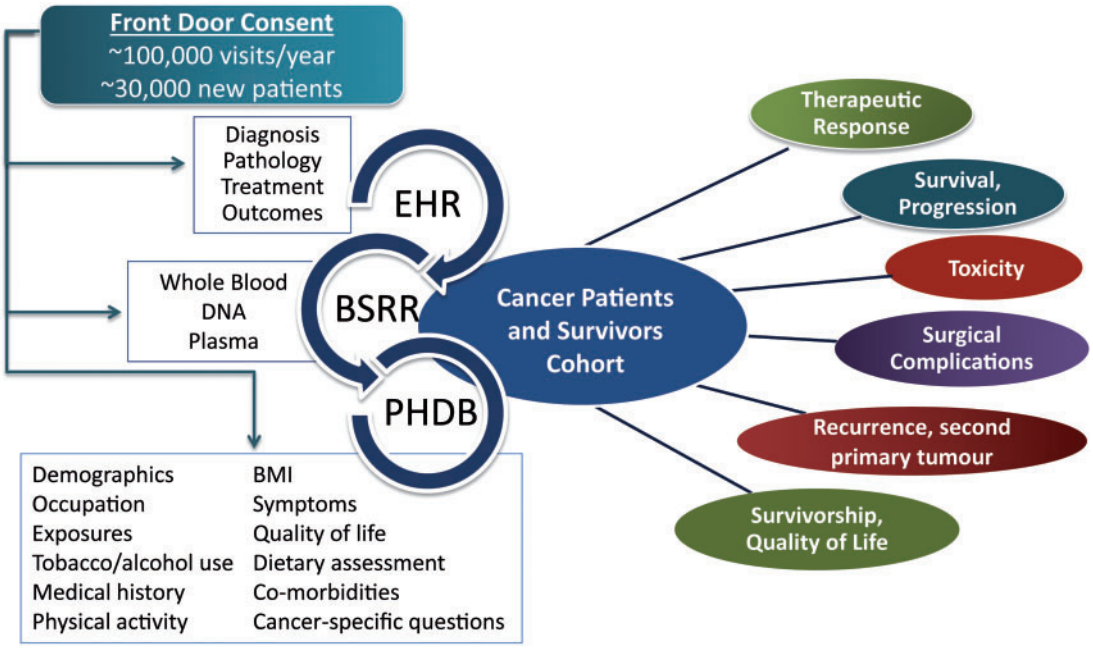

Figure 1. MD Anderson cancer patients and survivors cohort design.

Who is in the cohort?

The MDA-CPSC enrols all qualifying new MD Anderson patients who are 18 years of age and over, diagnosed within 1 year of registration and US residents. Whereas the catchment area for MD Anderson encompasses the state of Texas, many patients are admitted from other states. Of all new patients registered in fiscal year 2013, some 26 020 received a malignant diagnosis. Of these, 58.5% were Texas residents, 38.5% came from other states in the USA and 3% were international.These demographics are reflected in the MDA-CPSCwith the exception of foreign patients who are excluded due to anticipated difficulty with follow-up. As of May 2015, the cohort includes155 155 participants. Based on self-report of race/ethnicity, the majority of patients self-identified as White, with 7.26% Hispanic and 6.97% Black (Table 1).

Currently, biospecimens from 88 998 patients have been banked within the BSRR and are available for research purposes. Of these, 55 974 meet the inclusion criteria of MDA-CPSC (Table 2). Cancers with the highest representation include breast (10.6%), lung (8.0%), prostate (7.7%) and colorectal (6.1%) cancer. This distribution reflects current patterns for cancer incidence across the USA and across Texas.[1] In addition, the MDA-CPSC has good representation of less common cancers. Enrolment will continue at an estimated 10 000 patients per year until the goal of 100 000 participants with biological samples is met.

Table 1. Race/ethnicity distribution of MDA-CPSC participants

	Cohort participants			
Race/ethnicity	*n*	%	SEER[b]	TCR[c]
Am. Indian/Native Am.	587	0.38	0.43	0.09
Asian/Pacific Islander	2196	1.42	5.36	1.36
Black	10808	6.97	9.47	10.73
Hispanic origin	11262	7.26	7.28	17.86
Others	1431	0.92	–	–
Unknown[a]	10206	6.58	0.83	0.65
White	118665	76.48	77.0	69.30
Total	155155	100.00		

Am. American.

[a]Patient left blank/no information collected/unknown.

[b]Surveillance, Epidemiology, and End Results (SEER) Program [www.seer.cancer.gov] SEER*Stat Database: incidence—SEER 18 Regs research data + Hurricane Katrina impacted Louisiana cases, Nov 2013 submission (1973–2011 varying)—linked to county attributes—total USA, 1969–2012 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2014 (updated 7 May 2014), based on the November 2013 submission.

[c]Texas Cancer Registry [www.dshs.state.tx.us/tcr] SEER*Stat Database, 1995–2012 limited-use incidence, Texas statewide, Texas Department of State Health Services, created April 2015, based on NPCR-CSS submission, cut-off 19 November 2014.

How often have they been followed up?

Follow-up for the MDA-CPSC is enabled by MD Anderson'sTumour Registry, with additional medical chart review. The Tumour Registry collects basic clinical

Table 2. Diagnosis of MDA-CPSC cohort participants

	Cohort participants		Biospecimens	
Diagnosis	*n*	%	*n*	%
Breast	14979	9.65	5914	10.57
Lung/bronchus	14467	9.32	4474	7.99
Prostate	12777	8.23	4310	7.70
Skin	9629	6.21	1825	3.26
Head and neck	7929	5.11	2468	4.41
Colorectal	7251	4.67	3412	6.10
Lymphoma	6422	4.14	3144	5.62
Leukaemia	5242	3.38	2256	4.03
Brain and nervous system	4877	3.14	1288	2.30
Pancreas	4370	2.82	1319	2.36
Kidney	4138	2.67	2007	3.59
Bladder	2173	1.40	1085	1.94
Ovary	2085	1.34	793	1.42
Other malignancies	23807	15.34	9466	16.91
Multiple primary	30160	19.44	10458	18.68
Unknown primary	4849	3.13	1755	3.14
Total	155155	100.00	55974	100

information on cancer site, histology and stage, treatment prior to admission, treatment at MD Anderson, vital status and diagnosis of second primary tumours. Clinical coding specialists abstract patient data from the EHR 4 to 6 months after patient registration. This time lapse allows for definitive staging, pathology and diagnosis and for the initial treatment to be completed. The Tumour Registry checks vital status on all cancer patients on an annual basis, using active and passive approaches. Computer matches with MD Anderson appointment files identify patients who have presented for a recent appointment versus those who have not visited within the past 12–15 months. For the latter group of patients, letters are sent to enquire about the patient's health, diagnosis of second primary cancers and vital status, followed by phone calls if there is no response. In the case of deceased patients, family members or friends may provide date and sometimes cause of death. Date and cause of death are also obtained from the EHR. All those lost to follow-up are reviewed with the Bureau of Vital Statistics of Texas and other states, as well as the Social Security Death Index, to capture data for any additional patients. The estimated loss to follow-up is less than 5%. Additional follow-up information is available from the EHR, including therapeutic response, toxicities, progression and recurrence. For patients continuing to receive follow-up care at MD Anderson, the PHDB core questionnaire with symptom burden assessment is completed every 36 months. As MD Anderson transitions to the Epic EHR system in 2016, it will be possible to deploy additional follow-up questionnaires through the MyChart tool, and 24-month assessments are planned for diet, symptom burden and more extensive QOL measurements.

What has been measured?

The epidemiology data collection component of the MDA-CPSC is the PHDB, which was launched by the Department of Epidemiology in December 1999. As part of the institutional registration process, all new patients are asked to complete the mandatory PHDB questionnaire that forms a component of their primary medical evaluation (Table 3). The standardized questionnaire collects comprehensive baseline information on demographics, occupation, exposures, tobacco and alcohol use history, past medical history, past cancer history, family history, current medical problems and comorbidities, symptom burden using the MD Anderson Symptom Inventory (MDASI[13,14]) and substance misuse (Table 3). In collaboration with clinical colleagues, a series of modules have been developed to collect additional information relevant to specific cancer sites. This questionnaire takes about 30 min to complete. Historically, the PHDB has existed in hard copy with over 300 000 patient records, but in 2014 an electronic version (ePHDB) was fully implemented and some 50 000 records have already been collected using this tool. The ePHDB allows patients to complete the questionnaire at home through MD Anderson's PreCare application or in the clinic by desktop computer or iPad. One of the primary goals of the PHDB is to build a core institutional resource that captures patient data to enhance clinical, translational and population-based research for all newly registered patients. The PHDB has collected and archived core epidemiological risk factor data on new patients at MD Anderson, with a completion rate of about 94%.

MD Anderson's EHR allows for abstraction of comprehensive and detailed information for clinical and pathology characteristics, treatment regimens and clinical endpoints (Table 3). Abstraction is possible for all clinical procedures and visits for the duration that the cohort participant receives their cancer care and follow-up at MD Anderson. Data collection includes information obtained from the pathology reports (histology, stage, grade, diagnostic test results), treatment-related metrics (treatment modality, dates, dose, schedule), treatment response (tumour response, changes in symptoms), toxicities (with NCI Common Terminology Criteria for Adverse Events scores) and outcomes (recurrence, progression and survival). Abstraction can also be customized to collect variables based on the research aims of individual projects within

Table 3. MDA-CPSC data collection tools

A. PHDB modules	Variables
Demographics	Date of birth, sex, primary language, education level, religious preference, race/ethnicity and place of birth
Occupation	Selection from a list of occupations
Exposures	Exposures known or suspected to be associated with cancer risk, including: asbestos, dust, vehicle exhaust, solvents, ionizing radiation, radon, excessive sun exposure and selected chemicals
Tobacco and alcohol use history	Frequency, type and duration
Past medical history	Illnesses and conditions by age and year of occurrence. Subsections: heart and blood vessels, brain and nerves, lungs, stomach/intestines, kidney/bladder, blood disorders, immune system, joints/skeleton, liver, endocrine, psychological/psychiatric, skin disorders, genetic diseases and genitourinary
Past cancer history	Age and year for previous cancers; summary of treatments received
Family history	Cancer diagnoses and age/dates for first- and second-degree relatives
Review of systems	Overview of current medical problems and comorbidities by system. Subsections: general, neurological, head and neck, breast, cardiovascular, respiratory, gastrointestinal, genitourinary, musculoskeletal, skin, endocrine, haematology/lymph, psychological, female and male
Substance misuse	Use during the past 6 months, including: amphetamines, cocaine/crack, heroin, methamphetamines, hallucinogens, inhalants, marijuana and barbiturates
Symptom burden (MD Anderson Symptom Inventory; MDASI)	Assesses 13 common cancer-related symptoms: pain, fatigue, nausea, vomiting, dry mouth, shortness of breath, lack of appetite, difficulty remembering, drowsiness, disturbed sleep, sadness, distress and numbness
Quality of Life (Short Form-12; SF-12)	General health-related, 12-items, physical and mental components of health
B. Clinical data available for abstraction from EHR	
Clinical characteristics	Status of primary cancer, performance status, weight loss, comorbidities, laboratory values
Disease characteristics	Clinical stage, pathology/histology, grade, tumour size, tumour markers, regional lymph node metastasis
Treatment	Type: initial and adjuvant therapies, e.g. surgery, chemotherapy, radiation, hormonal, bone marrow / stem cell transplantDose: dates, schedule, administration methods and concurrent medications
Clinical endpoints	Changes in tumour size, change in patient symptoms, emergency room visits and / or hospitalizations, overall response to treatment, duration of response, toxicity evaluation, time to recurrence and progression, survival and second neoplasms

the MDA-CPSC. Furthermore, several other institutional data resources are available to obtain additional clinical variables for research purposes, including clinical laboratory assay measurements and diagnostic test results. To date, we have abstracted detailed clinical data from approximately 50 000 cohort participants. The large sample size allows us to perform outcome studies on histology-, stage- and treatment-''homogeneous'' populations of patients and to address a variety of outcome endpoints that may not otherwise be possible in population-based or smaller cancer patient cohorts.

In the future, we plan to implement several additional collection tools to enrich QOL and energy balance data associated with the MDA-CPSC. For QOL, a panel of validated, established tools will be used to measure cancer-related QOL (Functional Assessment of Cancer Therapy; FACT-G), depression symptoms (Centre for Epidemiologic Studies Depression Scale; CES-D) and perception of stress (Perceived Stress Scale; PSS). For nutritional profiling, the short NCI Diet Screener Questionnaire (DSQ, 26-item) will be deployed and more detailed anthropometry variables related to obesity prior to cancer diagnosis will be collected. Current physical activity assessments will be made using the short form of the International Physical Activity Questionnaire (IPAQ)[15] and the easy-to-administer Godin Leisure Time Activity Questionnaire.[16]

A rapidly growing subset of the cohort participants has blood biospecimens available for research purposes through the BSRR. Launched in 2010, the BSRR leverages MD Anderson's 'front door consent policy' that covers collection and banking of residual tissues and bodily fluids for future research purposes. Typically, about 95% of newly registered patients provide consent. Residual blood samples are held for 48 h by diagnostic laboratory services

prior to release and delivery to the BSRR laboratory. A new protocol was activated in 2014 to collect fresh blood specimens for certain malignancies. This fresh blood is delivered to the BSRR within 2 h of collection. On receipt of all blood, the DNA and plasma are isolated, barcoded and banked. This resource creates an opportunity to link genetic and molecular profiles with the extensive patient data in the MDA-CPSC.

What has it found?

The MDA-CPSCis an excellent resource for discovery of risk factors, biomarkers and molecular signatures for prediction of clinical outcomes and survivorship endpoints. Among confirmed malignant diagnoses based on our Tumour Registry, the top 10 preponderant malignancies were breast, lung, prostate, skin, head and neck, colorectal, lymphoma, leukaemia, brain and nervous system, and pancreatic cancer (Table 2). The cohort is currently being drawn upon to support several large next-generation sequencing projects to identify rare cancer susceptibility loci that confer risk in pancreatic, colorectal, head and neck, and ovarian cancers, as well as melanoma.

Data and biospecimens from the MDA-CPSC have been integrated into a number of multi-centre, collaborative analyses evaluating genetic variation in various malignancies through genome-wide association studies (GWAS), including for lung, kidney, bladder, oesophageal and testicular cancers. Several novel findings have also been uncovered for non-Hodgkin's lymphoma through these collaborative efforts. For example, in diffuse large B-cell lymphoma, a meta-analysis of GWAS revealed five novel genetic variants in regions that encompassed credible candidate genes.[17] In follicular lymphoma, five novel non-human leukocyte antigen [HLA] susceptibility loci were identified as significant contributors to disease risk.[18] In the first GWAS of marginal-zone lymphoma, two independent risk loci in the HLA region were uncovered.[19]

Through the PHDB, one of the detailed epidemiology variables collected is smoking. Among the participants, there were 11.23% current smokers, 39.62% former smokers, and 46.95% never smokers (2.56% records had unknown smoking status, patient left blank/no information collected). The availability of data on smoking behaviour among cancer patients since 1999 enabled an analysis of the trend of smoking prevalence over time. We observed that between 2000 and 2012, the percentage of never smokers increased from 46% to 57%, whereas the percentage of current smokers, recent quitters and former smokers decreased from 13% to 10%, 7% to 5% and 33% to 28%, respectively (Figure 2). This analysis and similar approaches provide vital information that can be used to guide interventions among cancer patients and also cancer control recommendations. Furthermore, the depth of data available demonstrates the potential power of the MDA-CPSC for investigation of the role of smoking, as well as other epidemiology variables, in clinical outcomes.

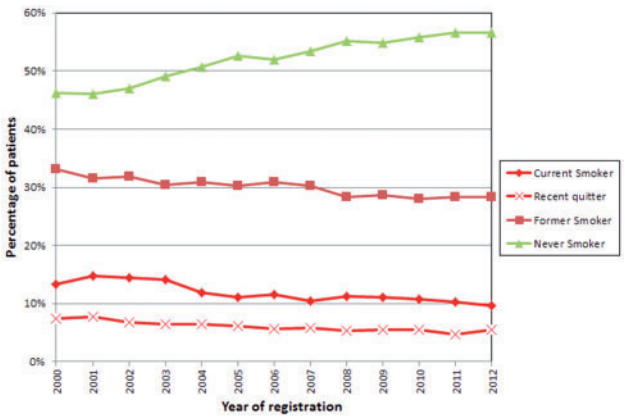

Figure 2. Temporal trend of smoking prevalence for MD Anderson patients.

Major efforts are being focused towards prediction of clinical outcomes and prognosis within the MDA-CPSC. The availability of diagnostic and clinical testing results have enabled the identification of serum albumin, bilirubin and lactate dehydrogenase (LDH) as possible prognostic factors for overall survival in non-metastatic breast cancer.[20] Using prostate cancer biospecimens, we have also reported on the ability of a genetic variant to stratify prostate cancer patients into a higher-risk subgroup at diagnosis.[21]

The PHDB captures baseline QOL data that are being analysed to better understand the mediators of poor QOL at diagnosis, associations between QOL with survival, and racial disparities in QOL. A consistent trend has been observed—cancer patients with poor QOL at diagnosis experience poorer overall survival compared with patients with better QOL. Racial differences in both QOL and survival have also been observed across several cancer sites. Ongoing studies are investigating the potential contributors and mechanisms for this relationship.

With the wealth of patient data in the MDA-CPSC, a major research effort has been devoted to the development of risk prediction models for clinical outcomes. For prostate cancer, MDA-CPSC data were used to develop a risk prediction model for cancer aggressiveness through analysis of a number of potential risk factors including epidemiology variables, comorbidities and laboratory variables. We showed that compared with a model based on age and smoking status, the incorporation of BMI and cardiac comorbidity improved model performance (area under the curve, AUC) by 4%, and further incorporation

of testosterone levels improved the AUC by 10%. Risk prediction models for breast cancer recurrence and survival have been developed based on data from approximately 15 000 women with stage I–III invasive primary breast cancer. In addition to known clinical variables, our models showed that lifestyle factors, such as alcohol consumption, cigarette smoking and QOL, may also influence survival and recurrence for breast cancer and were able to improve risk prediction. Furthermore, the distribution of these factors varied by race.

What are the main strengths and weaknesses?

Whereas MD Anderson is at the forefront of cancer genomic, population, translational and clinical research, the goal of the MDA-CPSC is to meld these strengths to specifically address the determinants of cancer-related outcomes and survivorship. The MDA-CPSC leverages MD Anderson's extensive pre-existing research framework including the large patient population, comprehensive EHR system and significant institutional resources.

The major advantages of the MDA-CPSC include extensive epidemiological and clinical data collection, a large patient population from a single institution that minimizes treatment heterogeneity, comprehensive clinical and follow-up information, and a diverse patient population with substantial numbers of minorities and patients with rare cancers. Moreover, the well-annotated biorepository is rapidly growing, adding additional opportunities for genetic and molecular profiling to better define predictors of clinical outcomes and survivorship endpoints.

The cohort is also strengthened by access to clinical laboratory and diagnostic test results through linkage with the institutional laboratory medicine database and the availability of tumour specimens banked within the institutional Tissue Biospecimen Pathology Resource (TBPR). The TBPR banks somatic tissues removed for biopsy or therapeutic resection. Access to these biospecimens enables integration of germline and somatic genetic information.

A potential limitation is that this is a single-institution patient population and the results may not be generalizable to all patients. However, the clinical and demographic characteristics of MDA-CPSC participants are compared with data from the Texas Cancer Registry and U.S. SEER data on an annual basis to inform analytical strategies to adjust for potential sampling differences. As shown in Table 1, the racial/ethnic distribution of MD Anderson patients is similar to that of the U.S. SEER population, alhough the cohort has a lower representation of Asian/Pacific Islanders and Black cancer patients. When compared with the Texas Cancer Registry, the percentage of Black and Hispanic cancer patients is fewer in the MDA-CPSC.

This cohort is poised to make significant impact in the areas of personalized medicine, survivorship and QOL, racial disparity after cancer diagnosis and rare cancers. The MDA-CPSC also provides exceptional opportunities for cross-racial comparisons with identification of behavioural, socioeconomic and biological differences that may explain the racial disparities in cancer outcomes and QOL. Integration of all data sources to generate comprehensive risk prediction models will aid in development of evidence-based guidelines across the cancer continuum.

Can I get hold of the data? Where can I find out more?

Biospecimens, study documents, summary-level information and individual-level data are available through a data sharing agreement. Requests for access should be addressed to Dr Xifeng Wu [xwu@mdanderson.org] and are evaluated by a Data and Biospecimen Access Committee on a case-by-case basis.

Profile in a nutshell

- The MDA-CPSC was designed to fill a tremendous need for a large, racially diverse and well-characterized cancer patient cohort, by integrating a rich biorepository with epidemiological, quality of life (QOL), clinical, pathological, treatment and follow-up data.
- Cohort participants comprise MD Anderson Cancer Centre patients who are newly diagnosed within 1 year of enrolment, age 18 years or older, and US residents.
- The core baseline epidemiology, demographics, exposure, QOL and medical/family history information collection for the MDA-CPSC was established in 1999, with blood collection launched in 2010.
- This ongoing cohort currently includes 155 155 participants with core patient data available and 55 974 participants with banked biospecimens.
- Access to cohort resources for collaborative research may be requested through the Data and Biospecimen Access Committee that reviews and prioritizes all research projects for approval.

Funding

This work was supported in part by: the Centre for Translational and Public Health Genomics, Duncan Family Institute for Cancer Prevention and Risk Assessment, the University of Texas MD Anderson Cancer Centre; the State of Texas Tobacco Settlement

Funds; and MD Anderson's Cancer Centre Support Grant (CA016672) from NIH/NCI.

Conflict of interest: All authors declare no conflict of interest.

References

1. American Cancer Society. *Cancer Facts and Figures 2015*. Atlanta, GA: American Cancer Society, 2015.
2. DeSantis CE, Lin CC, Mariotto AB *et al*. Cancer treatment and survivorship statistics, 2014. *CA Cancer J Clin* 2014;**64:**252–71.
3. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. *Nat Gene* 2006;**38:**500–01.
4. Colditz GA, Hankinson SE. The Nurses' Health Study: lifestyle and health among women. *Nat Rev Cancer* 2005;**5:**388–96.
5. Giovannucci E, Rimm EB, Colditz GA *et al*. A prospective study of dietary fat and risk of prostate cancer. *J Natl Cancer Inst* 1993;**85:**1571–79.
6. Alavanja MC, Akland G, Baird D *et al*. Cancer and noncancer risk to women in agriculture and pest control: the Agricultural Health Study. *J Occup Med* 1994;**36:**1247–50.
7. Kolonel LN, Henderson BE, Hankin JH *et al*. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol* 2000;**151:**346–57.
8. Schatzkin A, Subar AF, Thompson FE *et al*. Design and serendipity in establishing a large cohort with wide dietary intake distributions: the National Institutes of Health-American Association of Retired Persons Diet and Health Study. *Am J Epidemiol* 2001;**154:**1119–25.
9. Bernstein L, Allen M, Anton-Culver H *et al*. High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). *Cancer Causes Control* 2002;**13:**625–w35.
10. White E, Patterson RE, Kristal AR *et al*. VITamins And Lifestyle cohort study: study design and characteristics of supplement users. *Am J Epidemiol* 2004;**159:**83–93.
11. Kwan ML, Ambrosone CB, Lee MM *et al*. The Pathways Study: a prospective study of breast cancer survivorship within Kaiser Permanente Northern California. *Cancer Causes Control* 2008;**19:**1065–76.
12. Robison LL, Armstrong GT, Boice JD *et al*. The Childhood Cancer Survivor Study: a National Cancer Institute-supported resource for outcome and intervention research. *J Clin Oncol* 2009;**27:**2308–18.
13. Cleeland CS, Mendoza TR, Wang XS *et al*. Assessing symptom distress in cancer patients: the M.D. Anderson Symptom Inventory. *Cancer* 2000;**89:**1634–46.
14. Cleeland CS, Zhao F, Chang VT *et al*. The symptom burden of cancer: Evidence for a core set of cancer-related and treatment-related symptoms from the Eastern Cooperative Oncology Group Symptom Outcomes and Practice Patterns study. *Cancer* 2013;**119:**4333–40.
15. Craig CL, Marshall AL, Sjostrom M *et al*. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc* 2003;**35:**1381–95.
16. Godin G, Shephard RJ. A simple method to assess exercise behavior in the community. *Can J Appl Sport Sci* 1985;**10:**141–46.
17. Cerhan JR, Berndt SI, Vijai J *et al*. Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nat Genet* 2014;**46:**1233–38.
18. Skibola CF, Berndt SI, Vijai J *et al*. Genome-wide Association Study Identifies Five Susceptibility Loci for Follicular Lymphoma outside the HLA Region. *Am J Hum Genet* 2014;**95:**462–71.
19. Vijai J, Wang Z, Berndt SI *et al*. A genome-wide association study of marginal zone lymphoma shows association to the HLA region. *Nat Commun* 2015;**6:**5751.
20. Liu X, Meng QH, Ye Y, Hildebrandt MA, Gu J, Wu X. Prognostic significance of pretreatment serum levels of albumin, LDH and total bilirubin in patients with non-metastatic breast cancer. *Carcinogenesis* 2015;**36:**243–48.
21. He Y, Gu J, Strom S, Logothetis CJ, Kim J, Wu X. The prostate cancer susceptibility variant rs2735839 near KLK3 gene is associated with aggressive prostate cancer and can stratify gleason score 7 patients. *Clin Cancer Res* 2014;**20:**5133–39.