Hypothesis/Commentary

Cancer Epidemiology, Biomarkers & Prevention

Converting Epidemiologic Studies of Cancer Etiology to Survivorship Studies: Approaches and Challenges

Amy Berrington de González and Lindsay M. Morton

Abstract

There are nearly 12 million cancer survivors living in the United States, and the number continues to rise with ongoing improvements in treatment and screening. Assuring the long-term health of these patients poses both clinical and public health concerns. Survivorship research covers multiple aspects of life after a cancer diagnosis, including quality of life, acute and late effects of cancer treatment and mortality. Answering these questions requires a wide array of data, including information on the outcomes of interest, treatment history, and lifestyle. One potentially efficient approach to studying late effects and survivorship is to convert or extend existing epidemiologic studies of cancer etiology. In this article, we evaluate the different potential approaches for doing this and the challenges this entails. Our evaluation highlights the combinations of research topic and design most likely to succeed. We show that any question that relates to the existing information including prediagnosis lifestyle factors or genetics (if samples are available) could be efficiently studied, with an appropriate design. On the other hand, most, though not all converted studies would be ill-suited to the evaluation of the effect of treatment and postdiagnosis lifestyle changes. In terms of endpoints, hard outcomes including mortality and second cancers are more likely to be available within the existing study framework than other morbidities or quality of life. In light of the costs and time required to build new cohorts, appropriately leveraging the existing studies offers an important opportunity to gain new insights into cancer survivorship for both clinicians and patients. *Cancer Epidemiol Biomarkers Prev; 21(6); 875–80. ©2012 AACR.*

Introduction

The population of cancer survivors in the United States is approaching 12 million patients (1) and will continue to expand with ongoing improvements in treatment and screening. Assuring the long-term health of these patients poses both clinical and public health concerns. Consequently, a growing body of research addresses numerous related topics, from quality of life and patterns of medical care in these patients to acute and late effects of their cancer treatment and mortality. Answering these questions requires a wide array of data, including information on the specific outcomes of interest, treatment history, and lifestyle, as well as biologic samples. Finding an appropriate study setting in which such data are available or can reasonably be collected can be both challenging and expensive (2). Currently, only a few studies have been developed that provide the array of data described above, and these are focused on childhood cancer survivors (3, 4). The Childhood Cancer Survivor Study is one of the best examples of a large-scale study that has data on all the potential exposures, and the power and information to study a wide variety of outcomes (3). A comparable study of adulthood cancer survivors has been recommended (5), but to date no large-scale cohort has been launched.

One potentially efficient approach to studying late effects and survivorship, particularly for adult cancer survivors, is to convert or extend existing epidemiologic studies of first primary cancer etiology. The use of existing studies provides a number of advantages including lifestyle information, possibly biologic samples, contact details, and consent for follow-up for a large population. Additional information will, however, most likely be required to answer most study questions. We evaluate different potential approaches and challenges this entails, illustrating the potential and the pitfalls with examples from some specific studies. We consider the impact of the design of the existing study (cohort and case control), the key types of exposure information (treatment, lifestyle, and biologic samples), and issues related to outcome ascertainment. The strengths and weaknesses of this approach depend heavily on the outcome of interest and design of the original epidemiologic study. Therefore, we propose the types of questions that can most efficiently and effectively be studied in the setting of cancer survivor studies refitted from etiologic studies.

Authors' Affiliation: Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS

Corresponding Author: Amy Berrington de González, Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, 6120 Executive Boulevard, Bethesda, MD 20892. Phone: 301-594-7201; Fax: 301-402-0207; E-mail: berringtona@mail.nih.gov

doi: 10.1158/1055-9965.EPI-12-0131

Treatment data

Typically, limited availability of treatment data poses the greatest challenge to converting epidemiologic studies of cancer etiology to cancer survivor studies. Careful consideration of the level of detail required for different research questions often can mitigate this problem. If treatment forms the key exposure of interest (e.g., for studies of the acute or late effects of treatment), then high levels of completeness and detail will be essential. Conversely, if treatment is just a potentially weak confounder of the effects of some other exposure of interest (e.g., continued smoking), then far less stringent requirements apply. One can consider 5 different approaches to obtaining treatment data.

Medical record abstraction. Although medical records provide the most accurate and comprehensive source of detailed treatment information, obtaining them can be prohibitively time-consuming and expensive, particularly if the original study is a large, multicenter cohort studies with patients diagnosed over a lengthy period of time. Investigators need to recontact patients or their next-of-kin to request consent to abstract their treatment histories, and abstractors would need to visit numerous treatment centers. If the original study was a hospital-based case–control study, then this may have some advantages over population-based studies in that the initial treatment (at least) would have been conducted in a limited number of centers. Cancer etiology studies that have been conducted within the co-operative clinical trial groups could also be another efficient setting, especially if the electronic treatment databases are available for the treatment histories.

For all studies, regardless of their original design, an important consideration is the timing of the data collection with respect to diagnosis. Treatment may not yet be complete for patients who had their first primary cancer diagnosed most recently, necessitating collection of the first course of treatment only, or return at subsequent time points to update treatment histories. Ascertainment of complete treatment histories will be particularly difficult for cancers in which treatment may last many years, such as breast cancer.

Cancer registries. If investigators identify the first primary cancers through cancer registries, they can usually obtain some basic treatment data directly from the registries. However, the treatment data are usually restricted to the first course of treatment according to broad categories (e.g., chemotherapy, radiotherapy, surgery, hormonal therapy, and immunotherapy). Previous assessments of such data in Surveillance, Epidemiology, and End Results (SEER) registries suggest this approach may lead to substantial amounts of missing treatment information (6). As mentioned above, if treatment is only necessary because it may be a weak confounder of the exposure under study, then this level of information may be sufficient. Furthermore, although the treatment data are known to be incomplete, results of overall analyses adjusted for treatment category can be compared with results of sensitivity analyses in subsets of patients known to have received particular treatments.

Alternatively, a combination of registry data and medical record abstraction for a subset of patients can be used for sensitivity analyses and validation. For example, investigations of factors related to overall survival among patients with non-Hodgkin lymphoma (NHL) from a recent population-based case–control study conducted in 4 SEER areas used basic treatment information readily available from the cancer registries for the main analyses, which was supplemented by an assessment of the data quality by medical record abstraction in a single study center (7, 8).

Record linkage. Electronic records offer a great advantage in terms of cost and efficiency for a large population spread across a wide geographical area. The approach and availability will be highly dependent on the health system in the country in which the original epidemiologic study was conducted. In countries without centralized health systems, such as the United States, the need to merge or collect data from multiple health systems will be extremely time consuming, not only because of the need to visit multiple locations but also because of the time required to standardize the data. Electronic medical records are frequently less detailed than paper records, and therefore key treatment details for some research questions (e.g., radiotherapy fields or chemotherapy dose) may not have been captured. For example, in the United States, linkage to Medicare claims data is one approach that has been used in the Iowa Women's Health Study (9) and is under consideration in a number of other epidemiologic cohorts. The obvious limitation of Medicare is that it restricts the study population to those aged 65 and older at the time of treatment and claims data may be subject to biases from billing practices or suffer from lack of sufficiently detailed information (e.g., drug doses). More options may be available in countries with universal health care. In the United Kingdom, the availability of electronic hospital episode statistics for individual patient record linkage studies recently has been expanded from Scotland to England (10). This enables reconstruction of certain aspects of treatment history, such as surgery, but other sources would still be needed for outpatient treatments and prescriptions. In Scandinavia, some disease registries are available with relatively detailed and complete treatment histories, such as the Danish Breast Cancer Co-operative Group (11).

Self report. Patients can provide some limited data on their treatment histories with interviews or questionnaires. Self-report, however, suffers from potential biases due to being limited to patients that had survived to the time of the collection process, unless the collection was initiated at the beginning of the study. In addition, patients–particularly childhood cancer survivors–may not accurately remember their treatment history (12). There is some evidence though, at least for breast cancer patients, that basic self-reported treatment data are reliable (13).

Proxy variables. Can specific tumor characteristics such as stage, histology, and/or grade serve as surrogates for treatment? Some have proposed that, for certain cancer sites, these characteristics may suffice for adjustment for potential confounding by treatment. This strategy is most appropriate if treatment tightly correlates with the characteristics during the study period–a reasonable assumption for colon cancer, for which treatment is relatively consistent by stage, but not for prostate cancer, which has numerous treatment options for tumors of the same stage. If detailed information about the first primary cancer diagnosis has not been collected as part of the standard outcome ascertainment process, similar approaches to those described for obtaining treatment data could be used to collect these tumor characteristics.

Overcoming the hurdle of obtaining treatment data is of critical importance for considering the value of converting an existing epidemiologic study of cancer etiology to a cancer a survivor study. For questions relating to survival, at least some data on treatment (or a proxy) will probably always be required because of the strong impact of treatment on patient survival. Studies of acute and late effects of treatment that require detailed treatment data likely will be extremely resource intensive. For studies of second cancer etiology more broadly, it may be feasible to design a study question in which treatment is highly unlikely to act as a confounder. However, the potential for unforeseen confounders to be lurking due to collider–stratification bias needs to be assessed, as described further below.

Lifestyle information

Unlike treatment data, behavioral, environmental, and genetic data may be abundant in existing studies of cancer etiology. Indeed, this is a major rationale for leveraging existing studies instead of starting afresh. Attention needs to be paid though to the question of the relevant exposure period (e.g., lifetime exposure vs. current exposure) and whether exposures are modifiable or not. In a study designed from the outset as a survivorship study, the lifestyle data would likely be ascertained sometime soon after study enrollment and close to cancer diagnosis. This questionnaire could ask about current exposures but also about changes in habits since cancer diagnosis. Converting a first primary case–control study into a survivorship study would obviously closely approximate this design, although typically only prediagnosis behavior is collected. In contrast, a cohort study will not have information on lifestyle after the first primary cancer diagnosis unless multiple follow-up questionnaires are administered. It would, however, have the advantage that the prediagnostic information should be more reliable and accurate as it was not recalled after the first primary cancer diagnosis. Although this will not necessarily be a bias in a case–control study, nondifferential misclassification can still dilute the risk estimates, and cancer patients may have particularly poor recall of past behavior.

One challenge in using this prediagnostic information is the length of the interval between baseline and first primary cancer diagnosis, and the variability in this interval. For research questions that require data on current exposures, the availability of repeat questionnaires during follow-up for some cohort studies may make it feasible to reduce the interval between exposure ascertainment and the first primary cancer diagnosis by using the questionnaire closest to that diagnosis, but current exposure may still be highly misclassified. The Health, Eating, Activity and Lifestyle (HEAL) study showed substantial changes in physical activity and weight during the years immediately following a breast cancer diagnosis (14, 15). In contrast, continued tobacco use is persistent, even among patients diagnosed with tobacco-related cancer (16, 17). Thus, the level of exposure misclassification may depend on the exposure of interest and the characteristics of the patient population under study.

For the purpose of counseling cancer survivors in the clinic, the assessment of lifestyle information after the first primary cancer was diagnosed remains essential. The administration of a new questionnaire would resolve the above-mentioned limitations, but greatly reduces the advantages of converting the initial study if the existing information is to be largely disregarded.

Biologic samples

The availability of biospecimens is another potential key benefit that could be obtained from leveraging an existing epidemiologic study for cancer survivor research. In particular, prediagnostic blood samples in many existing cohort studies represent a precious resource for investigation of a range of biomarkers related to various outcomes of interest. In contrast, most existing case–control studies will only have postdiagnostic–and potentially even posttreatment–biospecimens, limiting the questions that can be studied to those biomarkers that are not likely to be affected by the first primary cancer diagnosis or treatment (e.g., germline DNA). Many fewer studies have the capability of obtaining tumor tissue, and tissue retrieval is resource intensive because patients are likely to have been diagnosed in many different hospitals. Nevertheless, it may be possible to obtain tumor tissue for a small number of patients to address specific research questions.

Outcome ascertainment

Survivorship studies encompass an extremely wide range of different outcomes, from quality of life to death. Whatever the outcome of interest, a key requirement will be to ascertain complete and comprehensive follow-up for that event plus competing risks for an extended period of time. For this reason, existing etiology studies may be better suited to evaluation of harder endpoints such as mortality and second primary cancers. An exception to this would be other comorbidities if they have been ascertained actively via follow-up questionnaires in an existing prospective cohort study, such as assessment of depression and Parkinson's disease in the NIH–AARP Diet and Health study cohort (18). Many current

cohort studies ascertain all malignancies as well as dates and cause of death for all members of the cohort, so ascertainment of second cancers and mortality among cancer survivors would not require collection of any additional information. In contrast, most existing case–control studies probably would not have these outcome ascertainment systems in place. The level of effort required to obtain this information may be more reasonable for existing case–control studies based in cancer registries, such as the SEER NHL study described previously.

Notably, even if all malignancies are ascertained for all patients in a study, the special care needed for ascertaining second, as opposed to first primary, cancers is often underappreciated. The rules and standards for defining and reporting second cancers vary dramatically, requiring careful development of definitions for multiple primaries. The interval between cancers is usually part of the definition, along with first cancer site, and histology if the second cancer is in the same organ (19). Sites particularly prone to the development of multiple primaries (e.g., breast), the consideration of field cancerization (e.g., bladder; ref. 20), and confusion–even in cancer registry data–between recurrences or metastases and new primary cancers are key challenges that may not be systematically considered in the current infrastructure of studies focused on first primary cancer etiology. In an active reporting system, a specially tailored questionnaire will be necessary to ensure that second primary cancers can be differentiated from recurrence, metastases, or repeated reporting of the initial cancer.

Analytical issues

Selection of a subpopulation–attractive at first glance–can produce substantial collider–stratification bias (21). This often overlooked source of bias in survivorship studies can occur if the risk factor under study for survival (or some other outcome) affects the risk of developing the original disease (i.e., the first primary cancer). Any other risk factors that are also related to the first primary cancer (the collider) and second cancer can become confounders when you stratify on that collider (i.e., select according to it), even if they are not related to the exposure of interest, the usual criteria for a confounder in the traditional setting. There are some interesting paradoxes in survival research that may be due at least in part to collider–stratification bias, such as obesity and cardiovascular disease patients. Paradoxically, studies consistently have shown that patients who are obese seem to survive longer following cardiovascular disease than normal weight patients (22). Causal diagrams can aid the identification of additional confounders that may be lurking due to the selection criteria, and an example is shown in Fig. 1. In this hypothetical example of a study of the relationship between alcohol consumption and second esophageal cancer in breast cancer survivors, age at menarche is a potential confounder. Although it would not usually be related to alcohol consumption in the general population, a relationship is induced between these 2 variables as they are both risk factors for breast cancer (the selected population for the study). Therefore, to correctly estimate the relationship between alcohol and esophageal cancer in this population, age at menarche should be included as an adjustment variable.

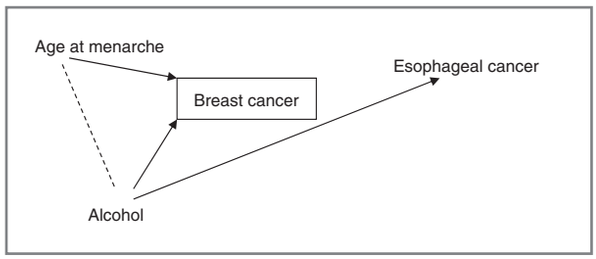

Figure 1. Causal diagram illustrating potential collider stratification bias in studying the risk of second esophageal cancer due to alcohol consumption in breast cancer survivors—age at menarche becomes a potential confounder due to the stratification on breast cancer survivors. The box indicates the stratification criteria (breast cancer survivors) and the dotted line indicates the confounding that is induced by the stratification.

The sample size is also an important analytic issue for consideration. For some outcomes of interest, or for any studies of survivors of rarer first primary cancers, even the largest existing cohorts may not have sufficient sample size and therefore a pooling approach may be required. The problem of sample size will be enhanced further if a specific duration of survival after the first primary cancer diagnosis is added as a requirement for entry into the subcohort, for example, if 2 or 5 years survival is required. Most standard methodologic considerations for pooling (e.g., challenges of exposure harmonization, evaluation of interstudy heterogeneity) should apply in this setting, though particular care might be warranted in assessing comparability of outcome ascertainment. In addition, comparability of the study populations may be a concern, if the original etiologic studies were not focused on the general population.

Summary

Survivorship research covers multiple aspects of life after a cancer diagnosis, and conversion of existing etiology studies, regardless of their design, cannot address all the questions in this extremely broad field of research. Our evaluation highlights the combinations of research topic and original study design most likely to succeed, and we summarize the likely availability of the necessary information in Table 1. We show that any question that relates to the existing information including prediagnosis lifestyle factors or genetics (if samples are available) could be efficiently studied, with an appropriate design. On the other hand, most, though not all converted studies would be ill-suited to the evaluation of the effect of treatment and post first cancer diagnosis lifestyle changes. In terms of endpoints, hard outcomes including mortality and second cancers are more likely to be available within the existing

Table 1. Summary of the typical availability of exposure and outcome data when converting epidemiologic studies of cancer etiology to survivorship studies

Original study design	Exposure type and availability		Outcome variable and availability	
Cohort	Treatment	Not routinely available. Detailed information difficult to collect. Basic information may be sufficient for confounding adjustment.	Mortality	Routinely available
	Lifestyle	Prediagnosis routinely available, postdiagnosis depends on availability of follow-up questionnaires.	Second cancers	Usually available, but care needed to define second cancers versus recurrence or metastases.
	Biologic samples	If available usually prediagnosis.	Other morbidities	Possibly available if active outcome ascertainment is routinely used.
Case control	Treatment	Not routinely available. Detailed information difficult to collect. Basic information may be sufficient for confounding adjustment.	Mortality	Not routinely available. Can be obtained with record linkage.
	Lifestyle	Retrospective prediagnosis routinely available, postdiagnosis rarely available.	Second cancers	Not routinely available. Can be obtained with record linkage or via questionnaire.
	Biologic samples	If available always postdiagnosis and possibly posttreatment.	Other morbidities	Not routinely available. Can be obtained with record linkage or via questionnaire.

study framework than other morbidities or quality of life. Despite the limitations described, the use of existing studies provides a number of advantages including lifestyle information, possibly biologic samples, contact details, and consent for follow-up for a large population. In light of the costs and time required to build new cohorts, appropriately leveraging the existing etiology studies offers an important opportunity to gain new insights into cancer survivorship for both clinicians and patients.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: A. Berrington de Gonzalez and L.M. Morton
Development of methodology: A. Berrington de Gonzalez
Writing, review, and/or revision of the manuscript: A. Berrington de Gonzalez and L.M. Morton
Study supervision: A. Berrington de Gonzalez

Received February 2, 2012; revised March 9, 2012; accepted March 9, 2012; published OnlineFirst March 16, 2012.

References

1. Parry C, Kent EE, Mariotto AB, Alfano CM, Rowland JH. Cancer survivors: a booming population. Cancer Epidemiol Biomarkers Prev 2011;20:1996–2005.
2. Oeffinger KC, van Leeuwen FE, Hodgson DC. Methods to assess adverse health-related outcomes in cancer survivors. Cancer Epidemiol Biomarkers Prev 2011;20:2022–34.
3. Robison LL, Armstrong GT, Boice JD, Chow EJ, Davies SM, Donaldson SS, 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.
4. Hawkins MM, Lancashire ER, Winter DL, Frobisher C, Reulen RC, Taylor AJ, et al. The British Childhood Cancer Survivor Study: Objectives, methods, population structure, response rates and initial descriptive information. Pediatr Blood Cancer 2008;50:1018–25.
5. Travis LB, Rabkin CS, Brown LM, Allan JM, Alter BP, Ambrosone CB, et al. Cancer survivorship–genetic susceptibility and second primary cancers: research strategies and recommendations. J Natl Cancer Inst 2006;98:15–25.
6. Jagsi R, Abrahamse P, Hawley ST, Graff JJ, Hamilton AS, Katz SJ. Underascertainment of radiotherapy receipt in surveillance, epidemiology, and end results registry data. Cancer 2012;118:333–41.
7. Cerhan JR, Wang S, Maurer MJ, Ansell SM, Geyer SM, Cozen W, et al. Prognostic significance of host immune gene polymorphisms in follicular lymphoma survival. Blood 2007;109:5439–46.
8. Geyer SM, Morton LM, Habermann TM, Allmer C, Davis S, Cozen W, et al. Smoking, alcohol use, obesity, and overall survival from non-Hodgkin lymphoma: a population-based study. Cancer 2010;116:2993–3000.
9. Virnig B, Durham SB, Folsom AR, Cerhan J. Linking the Iowa Women's Health Study Cohort to Medicare data: linkage results and application to hip fracture. Am J Epidemiol 2010;172:327–33.
10. Hospital episode statistics (HES). [accessed November 15, 2011]. Available from: http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937
11. The UK general practice research database, a powerful e-clinical research tool. Clinical Research Focus 2002;13.

12. Kadan-Lottick NS, Robison LL, Gurney JG, Neglia JP, Yasui Y, Hayashi R, et al. Childhood cancer survivors' knowledge about their past diagnosis and treatment: Childhood Cancer Survivor Study. JAMA 2002;287:1832–9.
13. Phillips KA, Milne RL, Buys S, Friedlander ML, Ward JH, McCredie MR, et al. Agreement between self-reported breast cancer treatment and medical records in a population-based Breast Cancer Family Registry. J Clin Oncol 2005;23:4679–86.
14. Irwin ML, Crumley D, McTiernan A, Bernstein L, Baumgartner R, Gilliland FD, et al. Physical activity levels before and after a diagnosis of breast carcinoma. Cancer 2003;97:1746–57.
15. Irwin ML, McTiernan A, Baumgartner RN, Baumgartner KB, Bernstein L, Gilliland FD, et al. Changes in body fat and weight after a breast cancer diagnosis: influence of demographic, prognostic, and lifestyle factors. J Clin Oncol 2005;23:774–82.
16. Walker MS, Vidrine DJ, Gritz ER, Larsen RJ, Yan Y, Govindan R, et al. Smoking relapse during the first year after treatment for early-stage non-small-cell lung cancer. Cancer Epidemiol Biomarkers Prev 2006;15:2370–7.
17. Cooley ME, Sarna L, Kotlerman J, Lukanich JM, Jaklitsch M, Green SB, et al. Smoking cessation is challenging even for patients recovering from lung cancer surgery with curative intent. Lung Cancer 2009;66:218–25.
18. Fang F, Xu Q, Park Y, Huang X, Hollenbeck A, Blair A, et al. Depression and the subsequent risk of Parkinson's disease in the NIH-AARP Diet and Health Study. Mov Disord 2010;25:1157–62.
19. Johnson CH, Peace S, Adamo P, Fritz A, Percy-Laurry A, BK E. The 2007 multiple primary and histology coding rules. Bethesda, MD: National Cancer Institute, Surveillance, Epidemiology and End Results Program; 2007.
20. Dakubo GD, Jakupciak JP, Birch-Machin MA, Parr RL. Clinical implications and utility of field cancerization. Cancer Cell Int 2007;7:2.
21. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol 2010;39:417–20.
22. Zeller M, Steg PG, Ravisy J, Lorgis L, Laurent Y, Sicard P, et al. Relation between body mass index, waist circumference, and death after acute myocardial infarction. Circulation 2008;118:482–90.