American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2011.

Vol. 174, No. 7
DOI: 10.1093/aje/kwr146
Advance Access publication:
August 4, 2011

Practice of Epidemiology

Use of Surveillance, Epidemiology, and End Results-Medicare Data to Conduct Case-Control Studies of Cancer Among the US Elderly

Eric A. Engels*, Ruth M. Pfeiffer, Winnie Ricker, William Wheeler, Ruth Parsons, and Joan L. Warren

* Correspondence to Dr. Eric A. Engels, Infections and Immunoepidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Boulevard, EPS 7076, Rockville, MD 20892 (e-mail: engelse@exchange.nih.gov).

Initially submitted October 18, 2010; accepted for publication April 8, 2011.

Cancer is an important cause of morbidity in the elderly, and many medical conditions and treatments influence cancer risk. The Surveillance, Epidemiology, and End Results (SEER)-Medicare database can be used to conduct population-based case-control studies that elucidate the etiology of cancer among the US elderly. SEER-Medicare links data on malignancies ascertained through SEER cancer registries to claims from Medicare, the US government insurance program for people over age 65 years. Under one approach described herein, elderly cancer cases are ascertained from SEER data (1987–2005). Matched controls are selected from a 5% random sample of Medicare beneficiaries. Risk factors of interest, including medical conditions and procedures, are identified by using linked Medicare claims. Strengths of this design include the ready availability of data, representative sampling from the US elderly population, and large sample size (e.g., under one scenario: 1,176,950 cases, including 221,389 prostate cancers, 185,853 lung cancers, 138,041 breast cancers, and 124,442 colorectal cancers; and 100,000 control subjects). Limitations reflect challenges in exposure assessment related to Medicare claims: restricted range of evaluable risk factors, short time before diagnosis/selection for ascertainment, and inaccuracies in claims. With awareness of limitations, investigators have in SEER-Medicare data a valuable resource for epidemiologic research on cancer etiology.

aged; case-control studies; data collection; epidemiologic methods; Medicare; neoplasms; risk factors; SEER Program

Abbreviations: HIV, human immunodeficiency virus; HMO, health maintenance organization; MEDPAR, medical provider analysis and review; NCH, national claims history; OUTPT, outpatient; PEDSF, patient entitlement and diagnosis summary file; SEER, Surveillance, Epidemiology, and End Results; SUMDENOM, summarized denominator.

Cancer is a major cause of morbidity in the United States, with a total of 1.34 million cases reported during 2005 from 49 of the 50 states (1). Cancer incidence typically rises with age, and a disproportionate fraction of cases occur among the elderly. For example, among people aged 65 years or older in the same 49 US states in 2005, there were 738,000 cancers (55% of the US total), including 131,000 lung cancers (67%), 90,000 colorectal cancers (64%), 113,000 prostate cancers (61%), and 79,000 female breast cancers (42%) (1).

Medical conditions are among the factors that are known or suspected to affect cancer risk. Examples of important etiologic associations include viral infections (e.g., human immunodeficiency virus (HIV) with Kaposi sarcoma and non-Hodgkin lymphoma, hepatitis C virus with liver cancer), autoimmune conditions (e.g., rheumatoid arthritis and Sjögren syndrome with non-Hodgkin lymphoma, ulcerative colitis with colon cancer), and metabolic conditions (e.g., obesity with cancers of the colon, esophagus, and uterus) (2–6). In addition, certain medical treatments or procedures can strongly increase subsequent cancer risk (e.g., radiation therapy and sarcomas, solid organ transplant and non-Hodgkin lymphoma) (4, 7). These associations arise because of the

direct effects of the medical conditions or their treatment (e.g., inflammation, metabolic disturbances, or direct DNA damage), or they may be explained by shared genetic traits or environmental exposures that predispose to both the medical condition and cancer. As with cancer, the prevalence of many of these medical conditions increases with age.

Understanding the etiology of cancer in the US elderly is important, especially given the overall aging of the population. The Surveillance, Epidemiology, and End Results (SEER)-Medicare database links data from the National Cancer Institute's SEER cancer registry program with claims data from Medicare, the federally funded insurance program for the US elderly. These data are made available to investigators and have been used extensively in research (details at http://healthservices.cancer.gov/seermedicare/). This resource is valuable for conducting research on cancer in the elderly because, as we describe below, it combines population-based ascertainment of cancer outcomes with insurance claims that can be used to assess the prevalence of medical conditions and treatments. For example, we previously used SEER-Medicare data in several case-control studies that evaluated risk of hematologic malignancies and skin cancers in association with immune-related conditions (8–15).

In the present paper, we describe a general approach to case-control studies of cancer using SEER-Medicare data. We highlight opportunities available to researchers and describe appropriate methods. In particular, we document unique challenges that arise from limitations in exposure assessment related to Medicare claims data, including a restricted range of evaluable risk factors, short time before cancer diagnosis or control selection for ascertainment of medical conditions and treatments, and inaccuracies in claims. These topics were not covered in our previous studies, which were examples of this approach but did not address detailed issues in methodology.

MATERIALS AND METHODS

SEER-Medicare data sources

SEER is a National Cancer Institute-funded program collecting data on cancer incidence and survival from US cancer registries (http://www.seer.cancer.gov). SEER began in 1973 with 9 state and metropolitan area cancer registries. Successive expansions in 1992 and 2001 led to the inclusion in SEER of 17 cancer registries that presently cover approximately 26% of the US population. The number of elderly adults (aged $\geq$65 years) in SEER coverage regions and the number of cancers in the elderly population are shown by calendar year in Table 1. In total, 146 million person-years are covered during 1973–2007, with 3.1 million incident cancers.

Medicare provides federally funded health insurance for approximately 97% of persons aged 65 years or older in the United States (16). Medicare also provides health insurance for individuals under age 65 years who have end-stage renal disease or medical disability. As of 2005, Medicare covered 42.6 million people, of whom 35.8 million (84%) were over the age of 65 years (http://www.cms.hhs.gov/MedicareEnRpts/). All beneficiaries are entitled to Part A coverage, which includes hospital inpatient care. Approximately 96% of participants pay to subscribe to Part B coverage, which covers physician and outpatient services. In January 2006, Medicare began offering voluntary outpatient coverage for medications (Part D); these data have only recently been made available for researchers.

Medicare reimburses providers under a fee-for-service model for specific procedures (e.g., office visit, surgery, radiographic imaging) tied to appropriate medical diagnoses. Alternatively, some Medicare beneficiaries (24% as of 2010) choose to enroll in a health maintenance organization (HMO) that provides capitated care (http://www.statehealthfacts.org/index.jsp). HMOs are not required to submit claims to Medicare for individual services, so there is no information about specific medical conditions or health care provided for Medicare beneficiaries enrolled in HMOs.

The SEER-Medicare database comprises files created during electronic linkage of SEER and Medicare data (http://healthservices.cancer.gov/seermedicare/). The linkage utilizes a deterministic algorithm based on name, Social Security number, sex, and date of birth. The match successfully links 94% of the SEER cancer cases over age 65 years with specific Medicare recipients, the deficit reflecting that 3% of elderly people do not have Medicare and that an additional 3% do not have sufficient or accurate enough information for the linkage. Resulting match files are stripped of identifiers. The SEER-Medicare linkage is updated biennially. This paper utilizes data from the 2008 linkage, which includes SEER cancer cases through 2005 and Medicare claims through 2007; as of December 2010, SEER-Medicare data will include SEER cases through 2007 and Medicare claims through 2009.

Table 2 describes the files included in the SEER-Medicare database. The patient entitlement and diagnosis summary file (PEDSF) contains information on all SEER cancer cases who matched to Medicare records, including demographic data, details from SEER on cancer type (e.g., site, morphology, grade, stage), Medicare eligibility and coverage, and socioeconomic data collected by the US Census for the census tract where the patient resides. The summarized denominator (SUMDENOM) file contains similar demographic and Medicare data for a 5% sample of Medicare beneficiaries living in the SEER areas, randomly selected on the basis of the last 2 digits of their Social Security number. However, the SUMDENOM file excludes people from the 5% sample who were reported to SEER with an incident cancer (i.e., individuals who are in the PEDSF file have been removed from the SUMDENOM file).

The remaining files listed in Table 2 provide Medicare claims data for individuals in the PEDSF and SUMDENOM files. Specifically, the medical provider analysis and review (MEDPAR) file provides hospital claims for all short stay, long stay, and skilled nursing facility care. The national claims history (NCH) file includes claims from physicians and other noninstitutional medical care providers. The outpatient (OUTPT) file contains claims from institutional outpatient providers, including hospital outpatient departments, rural health clinics, renal dialysis facilities, outpatient rehabilitation facilities, and mental health centers.

Table 1. Population and Cancer Counts Among People at Least 65 Years of Age, SEER Cancer Registry Regions, 1973–2007

Calendar Year	Population Size, no.			Cancer Cases, no.			SEER Registries[a]
	Female	Male	Total	Female	Male	Total	
1973	910,592	643,871	1,554,463	11,720	14,238	25,958	SEER9
1974	1,070,498	750,541	1,821,039	14,380	16,831	31,211	SEER9
1975	1,168,250	802,317	1,970,567	15,734	18,407	34,141	SEER9
1976	1,191,572	819,181	2,010,753	16,226	19,356	35,582	SEER9
1977	1,223,470	837,517	2,060,987	16,769	20,017	36,786	SEER9
1978	1,256,377	857,184	2,113,561	17,158	20,582	37,740	SEER9
1979	1,291,060	879,154	2,170,214	18,213	21,583	39,796	SEER9
1980	1,325,095	901,150	2,226,245	19,254	22,490	41,744	SEER9
1981	1,356,438	920,854	2,277,292	20,098	23,263	43,361	SEER9
1982	1,390,489	943,124	2,333,613	20,636	23,841	44,477	SEER9
1983	1,426,712	966,929	2,393,641	21,633	24,887	46,520	SEER9
1984	1,458,762	988,856	2,447,618	22,762	25,652	48,414	SEER9
1985	1,490,180	1,011,253	2,501,433	23,926	26,336	50,262	SEER9
1986	1,523,160	1,036,347	2,559,507	24,553	27,541	52,094	SEER9
1987	1,556,452	1,061,439	2,617,891	25,902	29,827	55,729	SEER9
1988	1,583,809	1,081,481	2,665,290	26,281	30,271	56,552	SEER9
1989	1,613,390	1,103,897	2,717,287	26,888	31,516	58,404	SEER9
1990	1,643,662	1,126,420	2,770,082	27,756	34,089	61,845	SEER9
1991	1,672,571	1,147,746	2,820,317	28,738	37,999	66,737	SEER9
1992	2,353,038	1,610,509	3,963,547	39,769	55,168	94,937	SEER13
1993	2,388,039	1,641,159	4,029,198	40,346	52,328	92,674	SEER13
1994	2,411,502	1,663,884	4,075,386	40,372	49,462	89,834	SEER13
1995	2,439,111	1,692,083	4,131,194	41,255	48,610	89,865	SEER13
1996	2,461,685	1,714,168	4,175,853	42,078	48,927	91,005	SEER13
1997	2,471,807	1,727,413	4,199,220	43,391	50,133	93,524	SEER13
1998	2,476,894	1,739,692	4,216,586	43,935	50,565	94,500	SEER13
1999	2,491,851	1,754,934	4,246,785	43,758	51,441	95,199	SEER13
2000	4,915,497	3,473,990	8,389,487	85,660	101,961	187,621	SEER17
2001	4,947,123	3,512,837	8,459,960	86,694	103,960	190,654	SEER17
2002	4,978,135	3,550,234	8,528,369	86,513	103,465	189,978	SEER17
2003	5,023,423	3,599,456	8,622,879	84,858	100,834	185,692	SEER17
2004	5,057,371	3,641,622	8,698,993	85,841	101,656	187,497	SEER17
2005	4,953,626	3,589,050	8,542,676	83,589	97,132	180,721	SEER17
2006	5,162,172	3,755,460	8,917,632	86,098	101,374	187,472	SEER17
2007	5,247,421	3,839,211	9,086,632	86,337	103,042	189,379	SEER17
Total	85,931,234	60,384,963	146,316,197	1,419,121	1,688,784	3,107,905	

Abbreviation: SEER, Surveillance Epidemiology, and End Results.

[a] SEER9 refers to 9 registries providing data beginning in 1973–1975: Atlanta, GA; Connecticut; Detroit, MI; Hawaii; Iowa; New Mexico; San Francisco, CA; Seattle, WA; and Utah. SEER13 includes the SEER9 registries and 4 additional registries providing data beginning in 1992: Alaska Natives; Los Angeles, CA; rural Georgia; and San Jose-Monterey, CA. SEER17 includes the SEER13 registries and 4 additional registries providing data beginning in 2000: California (remaining areas), Kentucky, Louisiana, and New Jersey.

An overview of the case-control study design

The SEER-Medicare database provides an opportunity to conduct case-control studies utilizing population-based sampling. Specifically, consider the population of all elderly Medicare beneficiaries (aged 65 years or older) living in the SEER registry areas as the source population, that is, a cohort that people enter when they receive Medicare coverage. By selecting cancer cases over age 65 years from the PEDSF file, which consists of cancers identified by the SEER registries, one obtains a complete census of all cancers arising in this source population.

Table 2. Description of SEER-Medicare Data Files, 1973–2007

File Name	Description	Years of Data Available[a]	Comment
PEDSF	SEER cancer cases	1973–2005	One record per person, with up to 10 cancers.
			File includes a flag indicating which SEER cases were in the 5% random sample.
			File includes selected socioeconomic information from the US Census based on census tract.
SUMDENOM	5% random sample of Medicare recipients living in SEER areas, excluding SEER cancer cases	1986–2005	One record per person.
			File includes selected socioeconomic information from the US Census based on census tract.
MEDPAR[b]	Medicare hospital claims for all Part A short stay, long stay, and skilled nursing facility care	1986–2007	One record per hospital stay.
			Each record has up to 10 diagnoses and 6 procedures coded by using the ICD-9.
Carrier claims (NCH)[b]	Medicare claims from physicians and other noninstitutional medical care providers	1991–2007	Multiple records for the same service from different providers.
			Each record can contain multiple services coded by using the HCPCS.
			Each service has a corresponding ICD-9 diagnosis.
OUTPT[b]	Medicare claims from institutional outpatient providers	1991–2007	Multiple records per visit are based on revenue center codes.
			Each record must have an HCPCS code.
			ICD-9 diagnosis and procedure codes may be reported but are not required.

Abbreviations: HCPCS, Healthcare Common Procedure Coding System; ICD-9, *International Classification of Diseases*, Ninth Revision; MEDPAR, medical provider analysis and review; NCH, national claims history; OUTPT, outpatient; PEDSF, patient entitlement and diagnosis summary file; SEER, Surveillance Epidemiology, and End Results; SUMDENOM, summarized denominator.

[a] The heading, "years of data available," refers to the data files provided from the 2008 linkage.

[b] For cancer cases in PEDSF from 2003 to 2005, Medicare claims data are not available before 1998.

Although data on the entire Medicare cohort are not available, it is straightforward to construct a subcohort representing a 5% random sample. To do this, one utilizes the SUMDENOM file and adds back the people who developed cancer. This can be accomplished by using a flag in the PEDSF file that indicates which SEER cases were originally in the 5% sample. Using the 2008 version of the SEER-Medicare data, the authors found that the combined data set created from the SUMDENOM file and the flagged cases from the PEDSF file comprise a 5% subcohort of 812,290 Medicare beneficiaries who were living in SEER areas during some time point of SEER coverage. In a case-control study, one can sample from this 5% subcohort of Medicare recipients to create a representative sample of controls.

For the selected cases and controls, one then uses the linked Medicare claims prior to cancer diagnosis/control selection to identify the presence of medical conditions, treatments, or procedures (i.e., "exposures") possibly related to cancer risk. We emphasize that only exposures reflected in Medicare claims can be evaluated. As discussed below in more detail, this somewhat narrow definition prevents consideration of some important cancer risk factors. In the Discussion, we also review additional issues regarding availability and accuracy of claims that warrant careful attention.

Additional details on sampling of cases and controls

Because exposures are identified through Medicare claims, a key aspect of subject selection is to ensure comparability of the available Medicare data in cases and controls. Claims data are limited by age (data for most people are unavailable before age 65 years) and calendar year (no MEDPAR claims before 1986, no OUTPT or NCH claims before 1991; Table 2). In addition, for PEDSF cases diagnosed in 2003–2005, the most recent linkage does not include Medicare claims before 1998.

The PEDSF and the SUMDENOM files provide additional information on Medicare coverage status for each calendar month. Periods when subjects were covered by both Parts A and B and were not in an HMO are most informative with respect to claims data, and the investigator can use this information to select subjects with a minimum period of Medicare coverage or evaluate for differences in coverage between cases and controls that would lead to differential exposure assessment. For many analyses, we exclude a period prior to diagnosis/selection (several months to a year) from exposure assessment, because during this period cases may have been ill from their incipient cancer and, in comparison to controls, would have been more rigorously evaluated and treated for underlying conditions.

These considerations affect case and control selection. For example, with the exclusion of 1 year of exposure data from Medicare claims immediately prior to diagnosis/selection, we use the following selection criteria for cases: 1) diagnosis in PEDSF of the cancer of interest as a first cancer, where the cancer was not diagnosed first on autopsy or on the death certificate; 2) age at cancer diagnosis of 66–99 years; 3) calendar year of cancer diagnosis of 1987 or after; 4) at least

13 months of Part A, Part B, non-HMO Medicare coverage prior to cancer diagnosis (because Medicare coverage is usually continuous, and exclusion of 1 year of data prior to diagnosis would entail at least 1 earlier month of coverage for assessment of exposures).

Several variations are possible. First, the investigator may include all cases of the cancer of interest, not just as a first cancer; include cases diagnosed at autopsy or on death certificate; or not exclude cases based on a maximum age at cancer diagnosis. Second, depending on the importance of capturing outpatient claims documenting the exposure of interest, one may require that cases be diagnosed in 1992 or after to ensure availability of NCH and OUTPT claims data. Third, the investigator can specify that the minimum duration of Medicare coverage be continuous or that coverage be obtained over specific time windows prior to diagnosis.

Control selection from the 5% random sample of Medicare recipients in the SEER areas mirrors the above criteria for cases. For each calendar year from which cases were sampled, we enumerate individuals in the 5% random sample who were cancer free as of July 1 (the midpoint) of that year and who meet the specified Medicare coverage requirement (e.g., at least 13 months of prior Part A, Part B, non-HMO coverage). From the eligible group, we randomly select controls for each calendar year who are frequency matched to cases by sex and age as of July 1 of that year. Controls can be sampled only once in a calendar year, but they can be sampled repeatedly across multiple years, and they can later become cancer cases. However, cancer cases diagnosed in 2003–2005 cannot be used as controls before 2003, because unlike cancer cases before 2003, they lack claims data prior to 1998.

Additional details on exposure assessment using Medicare claims data

For the selected cases and controls, the investigator assesses the presence of exposures of interest using the Medicare claims submitted prior to the diagnosis/selection date. Table 2 provides some relevant characteristics of the Medicare files; a more detailed description of Medicare claims files is beyond the scope of this article, and readers are referred elsewhere (http://www.resdac.umn.edu).

For many conditions, an inpatient diagnosis in MEDPAR may be considered to indicate a more severe manifestation than claims present only in the NCH or OUTPT files. In addition, for many conditions, inpatient diagnoses may be more reliable than other claims, because hospitals are more thoroughly audited for accuracy of claim diagnoses than individual providers (17–20). For this reason, we often consider a medical condition to be present if there is either a single MEDPAR diagnosis or 2 NCH or OUTPT diagnoses separated by at least 30 days. As noted above, we usually exclude a period prior to diagnosis/selection to avoid differential ascertainment of exposures between cases and controls. Additional measures of exposure that can be used to assess associations with cancer include latency (time from first Medicare claim for the exposure until diagnosis/selection), inpatient (MEDPAR) vs. outpatient-only (NCH or OUTPT) diagnoses, and a ''dose-response'' relation using the number of claims for the condition as a measure of severity.

Statistical analysis

The prevalence of the exposure of interest is compared between cases and controls by using contingency tables and unconditional logistic regression. In the logistic regression models, the investigator adjusts for the matching factors such as calendar year, sex, and age. Polytomous logistic regression is used when more than one type of cancer case is analyzed (e.g., subtypes of non-Hodgkin lymphoma).

Under the approach we have outlined, the variance of the odds ratios from these models needs to be adjusted for the multiple sampling of controls across calendar years and the inclusion of some controls as subsequent cases (12). Further statistical details are given in the Appendix.

Upon request of the corresponding author, we will provide the following macros for SAS software (SAS Institute, Inc., Cary, North Carolina) that assist investigators in the selection of cases and controls and in statistical analyses, using the above approach.

1. ALLCANCER.FILE.SAS: Selects cancer cases from PEDSF.
2. SUMDENOM.ALLCANCERS.SAS: Selects matched controls from the 5% random sample of Medicare beneficiaries in SEER areas.
3. ROBUSTVARIANCE: Performs polytomous logistic regression accounting for the sampling design described in this paper.

RESULTS

Table 3 presents the number of SEER cancers in 1992–2005 selected as cases by using the criteria described above. Overall, there are 1,176,950 cancer cases, including 221,389 prostate cancers, 185,853 lung cancers, 138,041 breast cancers, and 124,442 colorectal cancers.

Table 4 compares these cases with 100,000 cancer-free controls selected as described above. By design, the cases are frequency matched perfectly by sex, age category, and calendar year. The cases and controls are also similar in terms of race/ethnicity and duration of Medicare claims data. These controls represent 86,336 unique individuals, with 74,249 selected once, 10,709 selected twice, and 1,378 selected 3 or more times. Also, 7,125 controls (7.1%) subsequently developed cancer.

The prevalence of some example medical conditions and procedures among the controls is presented in Table 5. As expected, some chronic viral infections (e.g., hepatitis C virus and HIV) and medical conditions (e.g., organ transplantation) that are strongly associated with cancers are quite rare in this population. A higher prevalence is seen for additional medical conditions and treatments of potential interest (e.g., rheumatoid arthritis, blood transfusion), and other conditions that may not be linked to cancer are also very common, as expected (e.g., depression, essential hypertension). The apparent prevalence of these conditions decreases with use of more stringent criteria, such as requiring multiple supporting claims (Table 5).

Table 3. Cancers Selected as Cases[a] Among Elderly US Medicare Beneficiaries ($n = 1,176,950$), 1992–2005

Cancer Site[b]	Count	% of Total
Lip	2,432	0.21
Tongue	4,619	0.39
Salivary gland	2,593	0.22
Floor of mouth	1,436	0.12
Gum and other mouth	3,907	0.33
Nasopharynx	797	0.07
Tonsil	1,620	0.14
Oropharynx	564	0.05
Hypopharynx	1,696	0.14
Other oral cavity and pharynx	527	0.04
Esophagus	11,829	1.01
Stomach	23,604	2.01
Small intestine	3,825	0.32
Cecum	29,916	2.54
Appendix	711	0.06
Ascending colon	21,333	1.81
Hepatic flexure	6,802	0.58
Transverse colon	10,617	0.90
Splenic flexure	4,122	0.35
Descending colon	6,378	0.54
Sigmoid colon	31,576	2.68
Large intestine NOS	5,946	0.51
Rectosigmoid junction	12,113	1.03
Rectum	25,555	2.17
Anus, anal canal, and anorectum	2,717	0.23
Liver	10,662	0.91
Intrahepatic bile duct	2,058	0.17
Gallbladder	3,885	0.33
Other biliary	5,102	0.43
Pancreas	34,402	2.92
Retroperitoneum	791	0.07
Peritoneum, omentum, and mesentery	1,482	0.13
Other digestive organs	1,350	0.11
Nose, nasal cavity, and middle ear	1,500	0.13
Larynx	8,447	0.72
Lung and bronchus	185,853	15.79
Pleura	84	0.01
Trachea, mediastinum, and other respiratory organs	249	0.02
Bones and joints	787	0.07
Soft tissue including heart	4,909	0.42
Melanoma of the skin	28,364	2.41
Other nonepithelial skin	4,253	0.36
Breast	138,041	11.73
Cervix	4,131	0.35
Uterine corpus	27,530	2.34
Uterus NOS	645	0.05
Ovary	16,621	1.41
Vagina	958	0.08
Vulva	3,404	0.29
Other female genital organs	762	0.06
Prostate	221,389	18.81
Testis	206	0.02
Penis	869	0.07
Other male genital organs	259	0.02
Urinary bladder	63,951	5.43
Kidney and renal pelvis	25,484	2.17
Ureter	1,533	0.13
Other urinary organs	770	0.07
Eye and orbit	1,537	0.13
Brain	9,860	0.84
Cranial nerves and other nervous system	461	0.04
Thyroid	6,082	0.52
Other endocrine including thymus	756	0.06
Hodgkin lymphoma—nodal	1,902	0.16
Hodgkin lymphoma—extranodal	87	0.01
Non-Hodgkin lymphoma—nodal	31,191	2.65
Non-Hodgkin lymphoma—extranodal	14,684	1.25
Myeloma	15,993	1.36
Acute lymphocytic leukemia	758	0.06
Chronic lymphocytic leukemia	11,757	1.00
Other lymphocytic leukemia	919	0.08
Acute myeloid leukemia	8,786	0.75
Acute monocytic leukemia	538	0.05
Chronic myeloid leukemia	3,788	0.32
Other myeloid/monocytic leukemia	445	0.04
Other acute leukemia	1,228	0.10
Aleukemic, subleukemic, and NOS leukemia	985	0.08
Mesothelioma	3,460	0.29
Kaposi sarcoma	684	0.06
Miscellaneous	43,061	3.66
Invalid	22	0.00
Total	1,176,950	100.00

Abbreviations: NOS, not otherwise specified; SEER, Surveillance Epidemiology, and End Results.

[a] Cases were individuals reported to SEER with a first invasive cancer at ages 66–99 years during calendar years 1992–2005, excluding cancers diagnosed at autopsy or by death certificate, and had a minimum of 13 months of Part A, Part B, non-health maintenance organization Medicare coverage preceding cancer diagnosis.

[b] Cases were classified by using the "SEER site recode with Kaposi sarcoma and mesothelioma." Refer to http://seer.cancer.gov/siterecode/icdo3_d01272003/ for details.

Table 4. Characteristics of Cases and Controls Sampled From SEER-Medicare, 1992–2005[a]

Characteristic	Cases (n = 1,176,950)		Controls (n = 100,000)	
	No.	%	No.	%
Sex				
Male	624,464	53.1	53,056	53.1
Female	552,486	46.9	46,944	46.9
Age at diagnosis/selection, years				
66–69	193,397	16.4	16,431	16.4
70–74	300,368	25.5	25,520	25.5
75–79	298,807	25.4	25,388	25.4
80–84	217,667	18.5	18,496	18.5
85–99	166,711	14.2	14,165	14.2
Median age, years	76		76	
Calendar year at diagnosis/selection				
1992–1996	291,790	24.8	24,793	24.8
1997–2001	416,327	35.4	35,371	35.4
2002–2003	239,179	20.3	20,323	20.3
2004–2005	229,654	19.5	19,513	19.5
Median calendar year	2001		2001	
Race				
White	1,005,547	85.4	83,374	83.4
Black	92,559	7.9	6,966	7.0
Hispanic	19,463	1.7	2,717	2.7
Asian	30,384	2.6	4,070	4.1
Other/unknown	28,994	2.5	2,873	2.9
Duration of Medicare coverage, years[b]				
1.1–3.9	231,530	19.7	20,651	20.7
4.0–7.5	312,476	26.6	27,668	27.7
7.6–10.9	269,017	22.9	21,571	21.6
11.0–19.5	363,927	30.9	30,110	30.1
Median duration	8.1		8.0	

Abbreviation: SEER, Surveillance, Epidemiology, and End Results.

[a] Cases were individuals reported to SEER with a first invasive cancer at ages 66–99 years during calendar years 1992–2005, excluding cancers diagnosed at autopsy or by death certificate, and had a minimum of 13 months of Part A, Part B, non-health maintenance organization Medicare coverage preceding cancer diagnosis. Controls were selected from the 5% random sample of Medicare beneficiaries living in SEER areas. Controls were alive and cancer free as of July 1 in the calendar year of selection, and they were frequency matched to cases according to sex, age group (categories shown in table), and calendar year. Controls were also required to have a minimum of 13 months of Part A, Part B, non-health maintenance organization Medicare coverage preceding the selection date (July 1).

[b] Duration includes time when subjects were covered by Parts A and B of Medicare, were not enrolled in a health maintenance organization, were at least 65 years of age, and for whom claims data are included in the SEER-Medicare database. Medicare coverage does not have to be continuous.

DISCUSSION

We describe a general approach for conducting population-based case-control studies of cancer among the US elderly using SEER-Medicare data. Cases and controls are drawn from the population of Medicare beneficiaries over age 65 years who reside in SEER catchment areas. Exposures are assessed by using linked Medicare claims.

A major strength of such case-control studies is the essentially complete ascertainment of cancer cases from the source population. Cancer registries participating in the SEER program are required to meet strict standards with respect to

Table 5. Prevalence of Selected Medical Conditions and Procedures Among 100,000 US Medicare Controls, 1973–2007

Medical Condition or Procedure[a]	No.	%
Medical conditions where diagnosis is based on at least 1 claim in MEDPAR, NCH, or OUTPT files[a]		
Hepatitis B virus infection	230	0.2
Hepatitis C virus infection	298	0.3
Human immunodeficiency virus infection	149	0.1
Rheumatoid arthritis	5,004	5.0
Depression	9,363	9.4
Essential hypertension	37,836	37.8
Medical conditions where diagnosis is based on at least 1 MEDPAR claim or 2 NCH and/or OUTPT claims separated by 30 days or more[a]		
Hepatitis B virus infection	117	0.1
Hepatitis C virus infection	191	0.2
Human immunodeficiency virus infection	27	0.0
Rheumatoid arthritis	2,427	2.4
Depression	5,926	5.9
Essential hypertension	28,020	28.0
Other conditions or procedures		
End-stage renal disease[b]	339	0.3
Kidney transplantation[c]	18	0.0
Blood transfusion[d]	5,290	5.3

Abbreviations: MEDPAR, medical provider analysis and review; NCH, national claims history; OUTPT, outpatient.

[a] For these conditions, diagnosis was based on claims excluding the 12-month period prior to control selection.

[b] Diagnosis of end-stage renal disease was based on at least 1 NCH or OUTPT claim for this condition or administration of dialysis in each of the 3 months prior to selection.

[c] Diagnosis of kidney transplant is based on a claim in MEDPAR with a diagnosis-related group code indicating that the procedure was performed during a hospitalization.

[d] Diagnosis of blood transfusion was based on claims in MEDPAR indicating the transfusion of packed red blood cells as a procedure or an indication that the number of transfused units (BLDPNTS variable) was greater than 0. The 12-month period prior to control selection was excluded.

case ascertainment and data quality (http://seer.cancer.gov). PEDSF data derived from SEER include information on tumor histology, grade, and stage, allowing analysis by cancer subtype. In parallel, the availability of a 5% random subcohort of Medicare beneficiaries provides an opportunity to select controls who appropriately reflect the source population. Individuals in this subcohort are eligible to be selected as controls for as long as they remain cancer free.

Cases and controls are thus fully representative of the elderly Medicare population living in SEER areas. These samples can be generalized to the entire US elderly population with 2 caveats. First, 3% of people over age 65 years do not have Medicare. Medicare eligibility depends on having Social Security benefits, or being married to someone with benefits, which in turn depends on documentation of work history. Although the proportion of elderly who do not qualify is very small, presumably poor people and recent immigrants would be overrepresented. The second caveat is that SEER areas are not entirely representative of the overall US population. SEER areas were selected to include a relatively large fraction of racial/ethnic minorities (refer to http://seer.cancer.gov/). SEER areas also overrepresent urban areas and higher income persons (16).

As we illustrate in Table 3, an added strength of the described approach is the very large number of cancer cases and controls that can be evaluated. The sample size is substantial even for some less common cancers, and these large numbers enhance the investigator's ability to examine rare medical conditions and procedures as cancer risk factors.

Importantly, researchers should be cautioned regarding several limitations. Because Medicare coverage is largely restricted to elderly people, the SEER-Medicare data cannot be used to evaluate risk factors that arise earlier in life. Likewise, the vast majority of cancer cases who can be included in a case-control study (i.e., with antecedent Medicare claims data for exposure assessment) are over age 65 years. One must be cognizant that results from studies of the elderly may not be generalizable to younger populations. Nonetheless, because risk of most cancers increases steeply with age, such studies are directly informative for a substantial fraction of cancer cases.

The major issues with use of SEER-Medicare to conduct case-control studies concern the completeness and accuracy of Medicare claims to evaluate risk factors of interest (i.e., exposure assessment). First, only conditions diagnosed and recorded by a health-care provider, or related procedures, can be evaluated. If a medical condition is asymptomatic or underdiagnosed in the elderly (e.g., possible examples include hepatitis C virus infection, depression, and alcoholism), then reliance on Medicare claims will lack sensitivity. In addition, as described earlier, Medicare claims (particularly in NCH) may falsely document the presence of a condition when it is not actually present. This nonspecificity can be reduced by requiring multiple claims or a MEDPAR claim for the condition.

Furthermore, the claims files described in Table 2 do not provide data on some classical exposures of interest to cancer epidemiologists. For instance, there are limited data on tobacco or alcohol use, except indirectly as indicated by the presence of medical conditions that arise from smoking (e.g., emphysema) or drinking (e.g., alcoholic hepatitis), or laboratory test results, except when abnormalities trigger a medical diagnosis (e.g., anemia). Similarly, data on physical activity and body mass index are not available, although obesity itself can be evaluated as a claims diagnosis. Without Part D data, researchers have no information on medication use, except for certain drugs administered as infusions or injections (e.g., chemotherapy). These restrictions limit the range of conditions that can be evaluated as risk factors.

As noted above, we typically exclude a period prior to cancer diagnosis/control selection from exposure assessment, because medical evaluation of cases likely leads to heightened ascertainment of medical conditions. This bias could be quite severe, since cases, as they develop early signs of cancer, would be expected to increasingly visit their

Table 6. US Medicare Claims Documenting HIV Infection Among Kaposi Sarcoma Cases and Controls, 1973–2007[a]

	First Medicare Claim for HIV Infection, According to Time Period Relative to Cancer Diagnosis or Control Selection							
	Month −3 or Earlier		Months −2 to −1		Months 0 to 11		Month 12 or After	
	No.[b]	%	No.	%	No.	%	No.	%
Kaposi sarcoma cases ($n = 602$)	14	2.33	12	1.99	42	6.98	1	0.17
Controls ($n = 119{,}704$)	158	0.13	10	0.01	41	0.03	33	0.03

Abbreviation: HIV, human immunodeficiency virus.
[a] This table presents unpublished data from a case-control study of skin cancer among elderly US adults (*Int J Cancer.* 2010;126(7):1724–1731) (15).
[b] Number of HIV diagnoses.

health-care providers. Both nonspecific health complaints and symptoms related to the organ system in which the incipient cancer is situated would prompt added testing and diagnoses in cases.

An example that supports exclusion of a period prior to diagnosis/selection is provided in Table 6, using unpublished data from a case-control study of skin cancer in the elderly (15). HIV infection is an established strong risk factor for Kaposi sarcoma, and results using the Medicare claims data for the period before 3 months prior to case diagnosis/control selection support this conclusion (2.33% of cases with a claim for HIV vs. 0.13% of controls, yielding a crude odds ratio of 18). However, these prevalence estimates based on antecedent Medicare claims substantially underestimate the true HIV prevalence. Notably, for the cases, many additional HIV diagnoses are present in the Medicare claims data in the months at or after Kaposi sarcoma diagnosis, reflecting both newly diagnosed infections (prompted by HIV testing after recognition of the cancer) and initial claims for previously recognized HIV infection. Indeed, the 14 HIV claims prior to diagnosis represent only 20% of all such claims among the cases. In contrast, controls have few additional claims documenting HIV infection at or after their selection date, reflecting an absence of specific medical attention and testing relative to their arbitrary selection date. To avoid differential exposure assessment between cases and controls, one must utilize only the HIV diagnoses prior to case diagnosis/control selection, even though this approach results in a marked underascertainment of HIV (particularly for the cases). In turn, this nondifferential underascertainment leads to a bias toward the null in the magnitude of association with exposures of interest.

Another important limitation in exposure assessment arises from the restricted window available before diagnosis/selection in which to assess Medicare claims. Specifically, claims data are not available prior to age 65 years (rarely, data are available from younger ages if the person was covered due to end-stage renal disease or disability) or 1986 for inpatient data in MEDPAR (NCH and OUTPT data begin in 1991). In addition, for cases and controls from 2003 to 2005, only claims in 1998 and after can be assessed. One may evaluate associations with time since first Medicare claim as a proxy for duration of exposure (i.e., latency), and increasing risk with increasing duration can be taken as evidence for an etiologic relation. However, for many exposures, the interval based on claims data is only a rough proxy, because the data cover only a limited time period, and it is usually not possible to determine when the exposure was first present. Furthermore, if the effect of an exposure on cancer risk is greatest soon after onset of the exposure (e.g., a new user effect for medication), evaluation of claims data that capture mostly long-term exposures will lead to an underestimate of the association (21).

While reliance on Medicare data somewhat restricts the duration over which associations can be assessed, the time window is often quite long. For example, among the controls shown in Table 4, the median duration for which claims were available prior to selection was 8.0 years, and 30.1% had 11 years or more. Nonetheless, cases and controls selected at young ages or in early calendar years will have more limited claims data, and less opportunity to be identified as exposed to the risk factor of interest, than subjects selected at older ages or later in calendar time. Depending on the exposure of interest, there may be a minimum duration of available claims data required to be reasonably certain of capturing the exposure, which would then entail eliminating subjects who are younger or from earlier calendar years. Our approach to control selection matches them to the cases according to age and calendar year, so that the lack of sensitivity in exposure assessment that arises from the limited duration of claims data is nondifferential.

Although we focused on a case-control design, other options can be utilized with the SEER-Medicare database. One is a case-cohort study design, considering all cancer cases in the Medicare population along with the reconstructed 5% random subcohort. Using incidence density sampling, it would also be possible to individually match controls to cases to create a ''nested'' case-control study (i.e., nested in the Medicare cohort). However, given the extremely large number of subjects, both approaches are computationally challenging. The case-cohort design requires repeated evaluation of exposure information (i.e., prior medical conditions or treatments based on Medicare claims) in each successive risk set. For the nested case-control study, the computation burden associated with individual control selection and analyses using conditional logistic regression could be substantial, and this approach would only be feasible for cancers where the number of cases is not too large. Nonetheless, these case-cohort and case-control approaches would be expected to yield equivalent measures of association.

In closing, we encourage investigators to utilize SEER-Medicare data, which can be readily obtained, to conduct studies evaluating risk factors for cancer. Such studies have compelling strengths, including the availability of large population-based samples of cancer cases and representative controls. There are also important challenges, particularly related to the limitations and complexities of claims data, and we hope our discussion will facilitate appropriate study design and analysis.

ACKNOWLEDGMENTS

Author affiliations: Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland (Eric A. Engels, Ruth M. Pfeiffer); Information Management Services, Rockville, Maryland (Winnie Ricker, William Wheeler, Ruth Parsons); and Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, Maryland (Joan L. Warren).

This research was supported by the National Cancer Institute.

The authors acknowledge the efforts of the Applied Research Program, National Cancer Institute; the Office of Research, Development, and Information, Centers for Medicare and Medicaid Services; Information Management Services, Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER-Medicare database.

The interpretation and reporting of these data are the sole responsibility of the authors.

Conflict of interest: none declared.

REFERENCES

1. CDC WONDER. United States cancer statistics, 1999–2005 mortality archive request. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2008. (http://wonder.cdc.gov/CancerMort-v2005.html). (Accessed September 2, 2010).
2. Renehan AG, Tyson M, Egger M, et al. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet*. 2008;371(9612):569–578.
3. Tucker MA, D'Angio GJ, Boice JD Jr, et al. Bone sarcomas linked to radiotherapy and chemotherapy in children. *N Engl J Med*. 1987;317(10):588–593.
4. Grulich AE, van Leeuwen MT, Falster MO, et al. Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. *Lancet*. 2007;370(9581):59–67.
5. Ekström Smedby K, Vajdic CM, Falster M, et al. Autoimmune disorders and risk of non-Hodgkin lymphoma subtypes: a pooled analysis within the InterLymph Consortium. *Blood*. 2008;111(8):4029–4038.
6. Saito I, Miyamura T, Ohbayashi A, et al. Hepatitis C virus infection is associated with the development of hepatocellular carcinoma. *Proc Natl Acad Sci U S A*. 1990;87(17):6547–6549.
7. Ekbom A, Helmick C, Zack M, et al. Ulcerative colitis and colorectal cancer. A population-based study. *N Engl J Med*. 1990;323(18):1228–1233.
8. Anderson LA, Pfeiffer R, Warren JL, et al. Hematopoietic malignancies associated with viral and alcoholic hepatitis. *Cancer Epidemiol Biomarkers Prev*. 2008;17(11):3069–3075.
9. Anderson LA, Landgren O, Engels EA. Common community acquired infections and subsequent risk of chronic lymphocytic leukaemia. *Br J Haematol*. 2009;147(4):444–449.
10. Anderson LA, Gadalla S, Morton LM, et al. Population-based study of autoimmune conditions and the risk of specific lymphoid malignancies. *Int J Cancer*. 2009;125(2):398–405.
11. Anderson LA, Pfeiffer RM, Landgren O, et al. Risks of myeloid malignancies in patients with autoimmune conditions. *Br J Cancer*. 2009;100(5):822–828.
12. Quinlan SC, Morton LM, Pfeiffer RM, et al. Increased risk for lymphoid and myeloid neoplasms in elderly solid-organ transplant recipients. *Cancer Epidemiol Biomarkers Prev*. 2010;19(5):1229–1237.
13. Chang CM, Quinlan SC, Warren JL, et al. Blood transfusions and the subsequent risk of hematologic malignancies. *Transfusion*. 2010;50(10):2249–2257.
14. Lanoy E, Engels EA. Skin cancers associated with autoimmune conditions among elderly adults. *Br J Cancer*. 2010;103(1):112–114.
15. Lanoy E, Costagliola D, Engels EA. Skin cancers associated with HIV infection and solid-organ transplantation among elderly adults. *Int J Cancer*. 2010;126(7):1724–1731.
16. Warren JL, Klabunde CN, Schrag D, et al. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med Care*. 2002;40(8 suppl):IV-3–IV-18.
17. Katz JN, Barrett J, Liang MH, et al. Sensitivity and positive predictive value of Medicare Part B physician claims for rheumatologic diagnoses and procedures. *Arthritis Rheum*. 1997;40(9):1594–1600.
18. Fowles JB, Lawthers AG, Weiner JP, et al. Agreement between physicians' office records and Medicare Part B claims data. *Health Care Financ Rev*. 1995;16(4):189–199.
19. Kiyota Y, Schneeweiss S, Glynn RJ, et al. Accuracy of Medicare claims-based diagnosis of acute myocardial infarction: estimating positive predictive value on the basis of review of hospital records. *Am Heart J*. 2004;148(1):99–104.
20. Klabunde CN, Harlan LC, Warren JL. Data sources for measuring comorbidity: a comparison of hospital records and Medicare claims for cancer patients. *Med Care*. 2006;44(10):921–928.
21. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003;158(9):915–920.

APPENDIX

Variance Calculation for Polytomous Logistic Regression

Our variance calculation was previously presented in Quinlan et al. (12) and modifies an approach first described in Anderson et al. (8). Let $Y = (Y_0, Y_1, Y_2, \ldots. Y_K)$ to denote the outcome variable in a nested case-control study comprising one control group and K case groups. We use indicator notation, that is, $Y_0 = 1$ if the person is a control and

0 otherwise; and $Y_i = 1$ if the person is a case of type i and 0 otherwise, $i = 1, \ldots, K$. We use polytomous logistic regression to compare each case group with the controls, by modeling:

$$P(Y_i = 1 \mid X) = p(X, \theta_i) = \exp(X'\theta_i) / \sum_{s=1}^{K} \{1 + \exp(X'\theta_s)\},$$

for the covariate vector $X = [1, X_1, \ldots, X_m]$, that includes a one for the intercept term. As $\sum_{i=1}^{K} P(Y_i = 1) = 1$, we assume $\theta_0 = [0, \ldots, 0]$. We then use maximum likelihood estimation to obtain the log odds ratio estimates $\theta_j = [\theta_{j1}, \theta_{j2}, \ldots, \theta_{jm}]$, $j = 1, \ldots, K$, for the jth case type in the polytomous logistic model.

Although the corresponding covariance estimator accounts for the fact that the same control group is used for each disease subtype comparison, we additionally need to consider that, due to constraints in our subcohort, a substantial number of individuals were sampled multiple times as controls, and that some case individuals were sampled as controls prior to developing disease and becoming a case. Let the covariance matrix of the maximum likelihood estimates of the log odds ratio parameters be denoted by Σ. For each study subject, we obtain the scores $S_i = (S_{i1}, \ldots, S_{ik})$, from each of the K polytomous logistic regression models. For example, for subject l, the score for model j, or, equivalently, θ_j, is given by $S_{ij} = -X_{ij}[Y_{ij} - P(Y_{ij} = 1 | X_{ij}, \theta_j)]$. We define the matrix of scores for n study subjects as

$$S = \begin{pmatrix} S_{11} & S_{12} & \ldots & S_{1K} \\ S_{21} & S_{22} & .. & S_{2K} \\ S_{(n-1)1} & .. & S_{(n-1)(K-1)} & S_{(n-1)K} \\ S_{n1} & S_{n2} & S_{n(K-1)} & S_{nK} \end{pmatrix}.$$

Control subjects have entries in every column of the score matrix S, as they contribute to all logistic models. Individuals who served as controls before they were selected as cases also contribute to several logistic models. By use of the above notation, the asymptotic variance of the estimates $(\theta_1, \ldots, \theta_k)$ is given by $\Sigma B \Sigma$. B is estimated by the following equation:

$$\hat{B} = \sum_i \left(\sum S_{ik} \right) \left(\sum S_{ik} \right)',$$

where i denotes the sum over individuals, and the second sum inside refers to the repeated measurements on the same person.