Commentary and Review Series

RERF CR 1-92

Correcting for Catchment Area Nonresidency in Studies Based on Tumor-registry Data

Richard Sposto, Dale L. Preston

放影研 RERF
Radiation Effects Research Foundation
A Cooperative Japan–United States Research Organization

RERF Commentary and Review Series

Reports in the Commentary and Review Series are published to rapidly disseminate ideas, discussions, comments, and recommendations on research carried out by RERF scientists. This series also includes working papers prepared for national and international organizations, discussion of research concerning atomic-bomb survivors carried out elsewhere, and, in general, materials of lasting importance to RERF and atomic-bomb-survivor research. Unlike the RERF Technical Report Series, which conveys the results of original research carried out at the Foundation, reports in this series will receive only internal peer review. These reports may be submitted for publication in the scientific literature, in part or in toto. Copies are available upon request from Publication and Documentation Center, RERF, 5-2 Hijiyama Park, Minami-ku, Hiroshima, 732 Japan.

The Radiation Effects Research Foundation (formerly ABCC) was established in April 1975 as a private nonprofit Japanese foundation, supported equally by the Government of Japan through the Ministry of Health and Welfare, and the Government of the United States through the National Academy of Sciences under contract with the Department of Energy.

解説・総説シリーズ

腫瘍登録に基づくコホート調査における連絡地域外居住者についての補正§

Correcting for Catchment Area Nonresidency in Studies Based on Tumor-registry Data

Richard Sposto[1,2] Dale L. Preston[1]

要　約

腫瘍登録に基づくコホート調査で得られた癌発生率の推定値に対して連絡地域外居住者の及ぼす影響を考察し，各個人の居住歴または地域内居住の確率が分かっている場合に，ポアソン回帰分析により比較的簡単に補正が可能であることを示す。シミュレートしてできた大きなデータセットにこの簡易補正法が十分有効であったことを，完全なデータへの最尤法による解析といくつかのポアソン回帰分析との比較により示す。放射線影響研究所の腫瘍登録から得た胃癌発生率について，補正を行った場合と行わない場合の解析を比較する。また，死亡診断書のみに基づいて確認された症例を含める場合に，このことの持つ意義についても考察する。

§本論文にはこの要約以外に訳文はない。承認1992年1月29日。印刷1993年5月。

[1]放影研統計部，[2]現在，メリーランド州ポトマックEMMES会社。

Commentary and Review Series

Correcting for Catchment Area Nonresidency in Studies Based on Tumor-registry Data§

Richard Sposto,[1,2] Dale L. Preston[1]

Summary

We discuss the effect of catchment area nonresidency on estimates of cancer incidence from a tumor-registry-based cohort study and demonstrate that a relatively simple correction is possible in the context of Poisson regression analysis if individual residency histories or the probabilities of residency are known. A comparison of a complete data maximum likelihood analysis with several Poisson regression analyses demonstrates the adequacy of the simple correction in a large simulated data set. We compare analyses of stomach-cancer incidence from the Radiation Effects Research Foundation tumor registry with and without the correction. We also discuss some implications of including cases identified only on the basis of death certificates.

Introduction

The Radiation Effects Research Foundation (RERF) tumor registry is unique in that it allows one to perform cancer incidence studies within the well-defined cohort of the RERF Life Span Study (Thompson et al, in press). The Life Span Study (LSS) cohort comprises approximately 120,000 persons who were residents of Hiroshima and Nagasaki, Japan, in 1950, 93,000 of whom were in the cities at the time of the bombings (ATB) (Beebe and Ishida, 1959). The cohort has been followed since 1950 to study the effect on mortality of exposure to atomic-bomb (A-bomb) radiation. The LSS provides essentially complete follow-up for mortality, since it takes advantage of the Japanese population registry (*koseki*) system. Through this system, the date and cause of death are known rapidly (within 2 yr) for everyone who dies, and, by implication, the last follow-up time for individuals who are alive can be assumed to be within 2 yr. In the case of cancer incidence studies based on the RERF tumor registry, however, the LSS follow-up time does not take into account the fact that cancers diagnosed in individuals who are no longer residents of the tumor-registry catchment area will usually not be recorded in the tumor registry. Ignoring the problem of nonresidency will lead to

§*The complete text of this report will not be available in Japanese. Approved 29 January 1992; printed May 1993.*

[1]*Department of Statistics, RERF;* [2]*presently the EMMES Corporation, Potomac, Maryland.*

underestimates of cancer incidence because individuals who are not at risk for having a cancer recorded in the tumor registry are counted as if they are.

In this report we investigate the effect of nonresidency on estimates of cancer incidence. We show analytically that, although it may not be simple in general to correct exactly for nonresidency even if residency status is known, in the case of cancer, which is a relatively rare disease, a simple approximate correction to the usual Poisson regression analysis is possible. To demonstrate the adequacy of the simple correction, we compare five types of analyses in a large simulated data set based on total cancer incidence in Denmark (Muir et al, 1987). In addition, we compare two analyses of stomach-cancer incidence in Hiroshima and Nagasaki—one without any correction for nonresidency and one using a correction based on residency probabilities. We also discuss the implications of including death-certificate-only cases as incident cases in the analysis. In the Appendix, we outline the method used to estimate residency probabilities using patient-contact data from the RERF Adult Health Study (Hollingsworth and Beebe, 1960).

A Model for a Tumor-registry-based Cohort Study

Consider the following idealized model of a tumor-registry-based cohort study. A large, well-defined cohort, such as the LSS, is to be used as a basis for studying cancer incidence. Incident cancers occurring in the cohort are recorded in a regional, population-based tumor registry, whose catchment area includes the area of residency of all members of the cohort at the time follow-up was started. It can be assumed that all cancers of interest occurring in cohort members when they are residents of the tumor-registry area will be recorded in the tumor registry, and that their times will be recorded without error. However, as the cohort is followed, some individuals in the cohort may, either permanently or for various periods of time, migrate from the tumor-registry area. Cancers occurring in these individuals while they are not residents will not be recorded in the tumor registry. Cancers that are identified on the basis of death certificates only are assumed to have occurred among nonresidents and are not included, because, in our idealized model, other information corroborating the death certificate would have been available for individuals who were residents. We will discuss separately the issue of including death-certificate-only cases as incident cases. To simplify the mathematical treatment of nonresidency, we are intentionally ignoring a number of complications and problems that occur in actual tumor registries and leaving the discussion of these until later.

We assume that we know without error whether an individual is still alive. We will consider both the case when we know absolutely whether an individual is a resident of the tumor-registry area at a given time and the case when we know only the probability that he or she is a resident. The data that we can accrue on an individual include an indicator of whether a cancer was recorded in the tumor registry and the time associated with this, that is, either the date of cancer diagnosis or the date of last contact. If nobody ever left the tumor-registry area, we could do a proper analysis of cancer incidence by considering individuals at risk until the minimum of the time to first cancer, the time to last follow-up, or the time to death. If there is migration from the tumor-registry area, but we have precise residency data (ie, the dates when individuals were and were not resi-

dents of the tumor-registry area), we could do a proper analysis of incidence by incorporating this residency information into a maximum likelihood analysis, although as we will show below this is computationally difficult for large cohorts. Our goal is to produce reasonable but computationally feasible analyses using known residency histories or known probabilities of residency.

The Incidence Function for Observed Cancers in the Presence of Nonresidency

Let $\lambda(t)$ be the cause-specific hazard function (Prentice et al, 1978) for incident cancer (ie, the incidence function), which depends on time t, and let $\gamma(t)$ be the cause-specific hazard for death without cancer. $F(t)$, the corresponding survival function of the minimum of time to cancer or time to death without cancer (the probability of being alive at time t without having had cancer), is therefore

$$F(t) = \exp\left\{-\int_0^t [\lambda(u) + \gamma(u)]du\right\} \quad , \tag{1}$$

where time 0 is the beginning of follow-up. In addition, let $G(s;t)$ be the probability of survival until time $t + s$ after being diagnosed with cancer at time t, let $g(s;t)$ be the density of survival time s, and let $\pi(t)$ be the probability that an individual is a resident of the tumor-registry area at time t. Although these functions will depend on a number of individual characteristics, such as sex and age at the beginning of follow-up, we will assume this implicitly and not include it in the notation. We also assume that censoring due to the end of follow-up is independent of outcome.

The term *observed incidence function* will refer to the hazard function for cancers recorded in the tumor registry among individuals who are alive but who have not had a cancer recorded previously in the tumor registry. This is the hazard rate that would be estimated in an analysis of the tumor-registry data that ignores the nonresidency problem. In such an analysis, incident cancers are only those detected by the tumor registry, but the population *assumed* to be at risk in reality comprises individuals who are alive and have never had cancer regardless of whether they are residents, plus those alive who had a cancer that was not detected by the tumor registry because the individual was not a resident at the time of diagnosis. In terms of the quantities above, the observed incidence function, $\lambda^*(t)$, is

$$\lambda^*(t) = \frac{\pi(t)\lambda(t)F(t)}{F(t) + \int_0^t [1 - \pi(u)]\lambda(u)F(u)G(t-u;u)du} \quad . \tag{2}$$

The numerator is the density of cancer incidence times detected in the tumor registry at time t. The first term in the denominator is the probability of being alive at time t and never having had cancer. The second term in the denominator is the probability of being alive at time t but having a cancer before time t that was not detected.

Likelihood Function for Observed Cancers

The observable data from the tumor registry on each individual comprise a 0/1 indicator δ of whether a cancer was observed at t, where t is either the time to death or time to the end of follow-up if $\delta = 0$. Assume for the moment that a complete residency history is also available in the form of the indicator function $R(t)$, which takes on the value 1 at any time t that the individual is a resident of the catchment area and 0 otherwise.

The likelihood contribution for an individual with observed data (δ, t) conditional on the residence history $[R(u): 0 < u < \infty]$ is

$$L = [R(t)\lambda(t)F(t)]^{\delta}\left\{F(t) + \int_0^t [1 - R(u)]\lambda(u)F(u)G(t-u;u)du\right\}^{1-\delta} \quad . \tag{3}$$

Since $\delta = 1$ implies $R(t) = 1$, the left term in the product is the same as the likelihood for observed cancer when there is no residency problem. The right term is the sum of the chance of surviving cancer free until time t and the chance of surviving to time t after experiencing a cancer that was missed by the tumor registry because of nonresidency at some time before t. When there is no nonresidency $[R(t) \equiv 1]$, Equation (3) reduced to the familiar form

$$L = \lambda(t)^{\delta}F(t) \quad . \tag{4}$$

Since Equation (3) is linear in $R(t)$ for given δ, the likelihood of δ and t unconditional on $[R(u): 0 < u < \infty]$ is obtained by replacing $R(\cdot)$ everywhere in Equation (3) by its expected value, $\pi(\cdot)$.

Approximate Observed Incidence Function in the Case of Rare Disease

Since cancer has, mathematically speaking, a relatively low incidence, it is interesting to consider what happens to Equation (2) when $\lambda(t)$ is small. If one lets $\lambda(t) = \eta\varepsilon(t)$ for positive constant η, the linear term of a Taylor expansion of Equation (2) around $\eta = 0$ gives an approximation to $\lambda^*(t)$ for small $\lambda(t)$,

$$\lambda^*(t) = \pi(t)\eta\varepsilon(t) = \pi(t)\lambda(t) \quad . \tag{5}$$

This suggests that for sufficiently small $\lambda(t)$, dividing the observed hazard rate $\lambda^*(t)$ by the residency probability $\pi(t)$ will yield approximately the desired true hazard rate $\lambda(t)$.

The Observed Hazard for Cancer Subtypes

The tumor-registry records the occurrence of all types of cancer, but typical analyses are based on cancer subtypes rather than on all cancers. In analyses of cancer subtypes, follow-up is nevertheless censored at the time of occurrence of any cancer. If we let $\theta_j(t)$ be the proportion of cancers occurring at time t that are of type j, then the true incidence function and observed incidence function for cancer of type j are

$$\lambda_j(t) = \theta_j(t)\lambda(t) \tag{6}$$

and

$$\lambda_j^*(t) = \theta_j(t)\lambda^*(t) \quad , \tag{7}$$

respectively.

In the case of estimating cause-specific hazards, a small value of $\theta_j(t)$ alone is not a sufficient condition for Equation (5) to lead to a good approximation to the true incidence function. Total cancer incidence $\lambda(t)$ must be sufficiently small.

An Approximate Correction for Residency in Poisson Regression Analyses

The analysis of survival data for large cohorts is simplified greatly by modelling the incidence function $\lambda(t)$ using a piecewise-constant function. In this method one assumes that the incidence function depends on a time-varying vector of covariates $Z(t)$ and models this incidence function using a piecewise-constant failure function taking values $\tilde{\lambda}(z_k)$, where k indexes one of K cells of a partition of the space of $Z(t)$, and the covariate vector $Z(t)$ is discretized into one of K values z_k, $k = 1, \ldots, K$; $Z(t)$ might, for example, comprise sex, age at exposure, and time since exposure, with age divided into K_A categories, time into K_T categories, and sex into two categories. In this case there would be $K = 2K_AK_T$ categories in the partition of $Z(t)$. The vector z_k would constitute an indicator of sex and the midpoint age and time in each of the K categories. The product of likelihood terms such as Equation (4) over all individuals in the cohort reduces in this case to

$$\prod_{k=1}^{K} \tilde{\lambda}(z_k)^{D_k} \exp[-P_k\tilde{\lambda}(z_k)] \tag{8}$$

(Clayton, 1988), where P_k is the total number of person-years spent by all individuals whose values of $Z(t)$ fall into the kth cell of the partition, and D_k is the number of incident cases observed among these individuals. The collection of person-years and incident case totals for the K-cell partition of $Z(t)$ is referred to as the "person-years table." Although these data are not really Poisson in nature, the likelihood resembles the likelihood of K independent Poisson random variables D_k with mean $P_k\tilde{\lambda}(z_k)$, so that estimates of the parameters of $\tilde{\lambda}(z_k)$ can be obtained using a computer program for Poisson regression. Complex nonlinear models of $\tilde{\lambda}(z_k)$ can be fit easily using the computer program AMFIT (Preston et al, 1993), which is designed for this purpose.

One can see from Equation (3), however, that when there is nonresidency, whether the residency history $[R(u); 0 < u < \infty]$ is known for all persons or simply the probabilities of residency $[\pi(u); 0 < u < \infty]$, a simplification similar to Equation (8) is not possible. Note also that the likelihood [Equation (3)] involves the probability of dying from cancer $G(s;t)$, which is ancillary to the problem of interest. Approximate methods of analysis based on Poisson regression techniques are possible, but they do not lead to an exact solution of the maximum likelihood problem.

In the case of cancer incidence analyses based on a tumor registry, when nonresidency is not taken into account, D_k in the Poisson regression analysis described above represents the number of cases occurring among residents of the tumor-registry catchment area. However, the person-year totals, P_k, that are available do not take into account residency status, so that the quantity $P_k\tilde{\lambda}(z_k)$ in Equation (8) is not appropriate. If overall cancer incidence is low enough for the approximation in Equation (5) to be reasonable, the incidence function appropriate for observed cancer cases is approximately $\tilde{\pi}(z_k)\tilde{\lambda}(z_k)$, where $\tilde{\pi}(z_k)$ is a piecewise-constant function that describes the probability of residency for individuals in the kth cell of the partition. One can therefore obtain estimates for the parameters of $\tilde{\lambda}(z_k)$ by using $\tilde{\pi}(z_k)P_k$ rather than P_k in Equation (8). If the probabilities $\tilde{\pi}(z_k)$ are unknown but residency histories are known, then a natural estimate of $\tilde{\pi}(z_k)$ is r_k/P_k, the proportion of total person-years at risk that were spent as residents in the tumor-registry area, where r_k is the sum of person-years spent as residents of the tumor-registry area. That is, one of the two following approximate likelihood functions is maximized via Poisson regression:

$$L' = \prod_{k=1}^{K} \tilde{\lambda}(z_k)^{D_k} \exp[-\tilde{\pi}(z_k)P_k\tilde{\lambda}(z_k)] \tag{9}$$

or

$$L'' = \prod_{k=1}^{K} \tilde{\lambda}(z_k)^{D_k} \exp[-r_k\tilde{\lambda}(z_k)] \quad . \tag{10}$$

Adequacy of the Simple Correction for Nonresidency in the Case of Cancer

To investigate the adequacy of the approximate Poisson regression analysis described above, one set of simulated cancer and residency data from a large cohort was generated, and five analyses of the data were performed: (1) Poisson regression analysis using the cancer cases and person-years observable if all individuals were residents at all times. (2) Poisson regression analysis using only cancer cases that occurred during periods of residency, but with person-years accumulated as if individuals were residents at all time—the analysis that ignores the nonresidency problem. (3) Poisson regression analysis using the same cancers cases as in analysis 2 but with person-years accumulated only for times when individuals were residents. This assumes knowledge of the residency history of each individual in the cohort, and corresponds to maximizing the approximate likelihood, Equation (10). (4) Poisson regression analysis using the cases and person-years used in analysis 2, with the person-years corrected using probabilities of residency, as described above. This assumes that residency probabilities are known and corresponds to maximizing the approximate likelihood, Equation (9). (5) A maximum likelihood analysis based on the likelihood in Equation (3), using individual data on follow-up and residency. This is the preferred analysis if complete residency data are available but is computationally difficult for large cohorts.

In all analyses, and in the generation of the data, cancer incidence was assumed to follow a piecewise-constant hazard function that was described by the fourth-order polynomial:

$$\lambda(t) = \alpha_0 + \alpha_1[\tau(t) - 40] + \alpha_2[\tau(t) - 40]^2 + \alpha_3[\tau(t) - 40]^3 + \alpha_4[\tau(t) - 40]^4 \quad . \quad (11)$$

Time t represents attained age, and $\tau(t)$ is the midpoint of the time interval containing t. Thirteen time intervals were used, with interval cutpoints 10 to 70 by 5, and 80. The population parameter values were obtained by fitting this piecewise-constant hazard model to the annual incidence of all cancers reported in the Danish Tumor Registry (Muir et al, 1987), both sexes combined, using Poisson regression methods. The population parameter values are shown in the last column of Table 1.

The 79,972 persons in the RERF tumor-registry cohort were used as a basis for the simulated data set. These persons constitute all LSS members with known radiation dose estimates who were alive in 1958. The cohort was followed until the end of 1985, although about 35% of those in the cohort died before then. To simplify the simulation, 1357 individuals who were over 65 yr old ATB were excluded. In addition, follow-up of the cohort was assumed to have started 10 yr after the bombings, in August 1955. The data used from the cohort were age in August 1945 (the time of the bombings) and age at last follow-up, regardless of the reason for terminating follow-up. This cohort was used only as a basis for a realistic pattern of follow-up and reasonable sex and age distribution in a large cohort. No actual cancer-incidence or residency data from this cohort was used.

All persons in the cohort were assumed to be alive and at risk at the beginning of 1955, and the cancer-incidence function in Equation (11) was then used to generate cancer cases. A random cancer time was generated for each individual in the cohort. Cancers occurring before the end of follow-up for the individual were recorded, as well as the time to cancer. Cancers occurring after the end of follow-up were not recorded.

In addition, for each individual, a residency history was also generated. Individuals were considered either resident or not resident throughout each time interval. A residency indicator for each interval was generated using the probability of residency estimated from RERF data for males living in Nagasaki (see the Appendix), resulting in a person-year-weighted average residency rate of 81% in the cohort. The residency indicator for each interval was generated to be independent of the indicator for other intervals. Although in reality one would expect residency status at one time to be dependent on status in previous times, since the analyses performed were either conditional on the residency history or dependent only on the expected value of the residency indicators, it is of no importance that the residency histories used arose from a mechanism that ignored inter-interval dependence.

The generated data were unrealistic in one other respect. Cancer was assumed to be nonfatal, which is to assume that survival is independent of cancer status [ie, $G(s;t)$ is independent of t]. One can see from Equation (2) that the adequacy of the approximation depends on the degree of fatality of the cancer; the more likely that cancer is to be fatal, the better the approximation. Hence, this simulation is an example from the unrealistic worst case in that cancer is completely nonfatal.

Table 1. Comparison of parameter estimates and estimates of annual incidence, from five analyses of simulated data set

	Analysis 1. Poisson regression, no nonresidency	Analysis 2. Poisson regression, no residency information	Analysis 3. Poisson regression, complete residency information	Analysis 4. Poisson regression, probability of residency	Analysis 5. Maximum likelihood	Population
Parameter[a]						
α_0	184.7 (4.95)[b]	135.9 (4.25)	184.1 (5.78)	182.4 (5.73)	184.2 (5.67)	180.6
α_1	17.6 (0.457)	15.0 (0.415)	17.2 (0.501)	17.2 (0.497)	17.2 (0.497)	17.3
α_2	0.641 (0.0352)	0.655 (0.0319)	0.623 (0.0396)	0.627 (0.0393)	0.626 (0.0395)	0.625
α_3	0.00817 (0.000657)	0.00883 (0.000601)	0.00870 (0.000728)	0.00840 (0.000720)	0.00884 (0.000726)	0.00919
α_4	-9.05×10^{-6} (3.83×10^{-5})	-4.74×10^{-5} (3.51×10^{-5})	1.83×10^{-5} (4.27×10^{-5})	3.32×10^{-6} (4.23×10^{-5})	2.16×10^{-5} (4.29×10^{-5})	4.46×10^{-5}
Age (yr)	Annual incidence per 100,000					
12.5	10.4	9.3	11.7	11.5	11.6	11.6
17.5	17.8	17.3	17.8	18.1	17.5	14.2
22.5	28.4	24.1	28.2	28.3	27.9	23.9
27.5	48.7	38.0	48.8	48.4	48.4	44.9
32.5	85.3	67.2	85.5	84.8	85.2	82.1
37.5	144.6	119.1	144.6	143.7	144.4	141.0
42.5	232.8	200.5	232.1	231.3	232.1	228.0
47.5	356.2	318.0	354.4	353.7	354.7	349.7
52.5	520.6	477.2	517.8	517.0	518.6	513.9
57.5	732.0	683.4	728.8	727.4	730.5	728.7
62.5	996.0	941.4	994.0	990.8	997.1	1003.1
67.5	1318.3	1255.3	1320.0	1313.1	1325.1	1346.6

[a]Parameter values are scaled to reflect annual incidence per 100,000.

[b]Standard errors are shown in parentheses.

The top half of Table 1 shows the parameter estimates and standard errors from each of the five analyses described above, along with the population parameter values. (Parameter values have been scaled to represent annual incidence per 100,000). The bottom half of the table shows corresponding annual incidence per 100,000 from the model. All analyses give essentially the same estimates with the exception of analysis 2, that is the analysis that simply ignores the problem of nonresidency. Note that the standard errors from analyses 3, 4, and 5, are all about the same, roughly 10% to 15% larger than those from analysis 1. One would expect more-precise estimates from analysis 1, since it is not subject to the additional uncertainty introduced by nonresidency. Analyses 3 and 4 agree closely with analysis 5, both in estimates and in standard errors, indicating that making the simple correction based on individuals known residency histories or on known residency probabilities adequately corrects for nonresidency and for the additional uncertainty in estimation due to nonresidency. A comparison of analyses 2–4 with the maximum-likelihood estimate (analysis 5) is shown in Figure 1, that gives the percentage difference in annual incidence per 100,000 for each age interval. Analysis 2 substantially underestimates the incidence rate for all ages, whereas the other analyses agree quite closely. Note that few cancers occur at ages less than about 25 yr, so that incidence is poorly estimated for these ages.

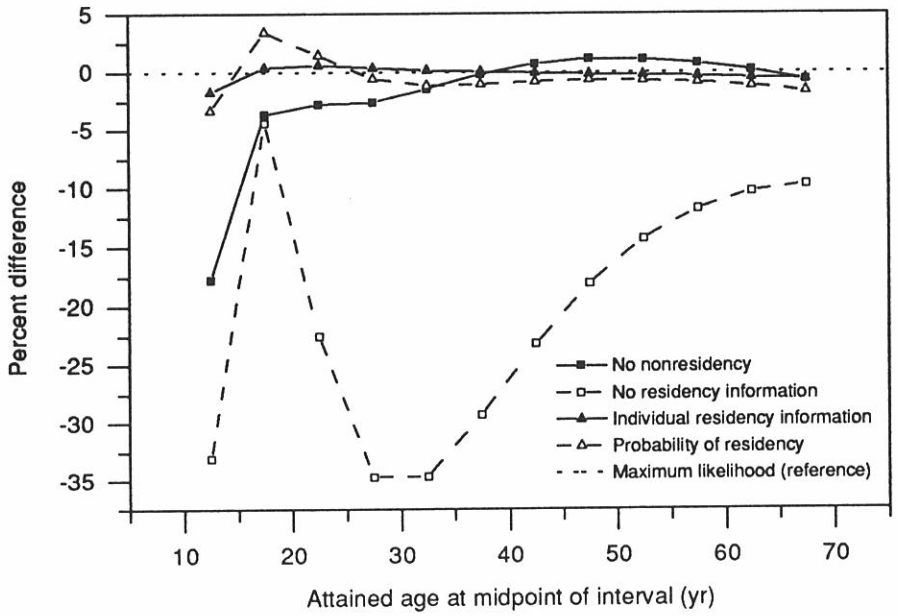


Figure 1. Comparison of Poisson regression estimates of annual incidence rates to maximum-likelihood estimate for simulated data set, percentage difference.

Comparative Analysis of Stomach-cancer Incidence with and without Correction for Nonresidency

To demonstrate further the effect of the correction for nonresidency, the RERF tumor-registry stomach-cancer incidence data were analyzed both with and without the correction for nonresidency. An excess-relative-risk model of the form

$$\lambda(city, sex, age\ ATB, age, dose) = \exp[\gamma_0 + \gamma_1 city + \gamma_2 sex + \gamma_3(age ATB - 25) + \gamma_{4s}\ln(\frac{age}{50}) + \gamma_{5s}\ln^2(\frac{age}{50})](1 + \beta dose) \quad (12)$$

was used, where *city* is an indicator with value 0 for Hiroshima and 1 for Nagasaki, *sex* is an indicator with value 0 for males and 1 for females, *age ATB* is age at the time of the bombings, *age* is attained age, and dose is intestinal radiation dose in sieverts (Sv). The parameters γ_{4s} and γ_{5s} depend on sex. This is the same model used in Thompson et al (1993). The parameter β is the excess relative risk at 1 Sv. Residency probabilities were estimated using patient-contact data from the RERF AHS, as described in the Appendix. The person-year-weighted average residency probability was 86%.

The parameters estimates from the two analyses are shown in Table 2. The estimate (standard error) of the excess relative risk β was 0.342 (0.090) without the correction for nonresidency, and 0.341 (0.090) with the correction, so that the excess relative risk was unaffected by the correction. The estimated background

Table 2. Estimates of parameters in the model for stomach-cancer incidence in the RERF Life Span Study without and with correction for nonresidency

	Estimate (no correction for nonresidency)		Estimate (corrected for nonresidency)	
γ_0	4.93	(0.0492)[a]	5.05	(0.0492)
γ_1	−0.190	(0.0454)	−0.133	(0.0453)
γ_2	−0.676	(0.0654)	−0.709	(0.0655)
γ_3	0.0116	(0.00233)	0.00888	(0.00234)
γ_{4s} (males)	4.47	(0.249)	4.26	(0.243)
γ_{4s} (females)	3.14	(0.198)	3.06	(0.193)
γ_{5s} (males)	−2.94	(0.480)	−2.54	(0.468)
γ_{5s} (females)	−0.650	(0.388)	−0.415	(0.376)
β	0.342	(0.0904)	0.341	(0.0903)
Excess cases (per 100,000 PYSv)	45.1		48.5	

Note: PYSv = person-year-sievert.

[a]The standard errors are shown in parentheses.

rates changed as a result of the correction, however, as demonstrated in Figure 2, that shows the estimate of background rate per 100,000 person-years by city and attained age for males exposed at age 20 ATB. For 40-yr olds the estimated incidence rate increased by 21% in Hiroshima and 29% in Nagasaki as a result of the correction, and for 65-yr olds by 11% and 17%, respectively. The effect of the correction is also reflected in estimates of the number of excess cases per 100,000 person-year-sievert in the LSS cohort that change from 45.1 cases to 48.5 cases as a result of the correction, a 7% increase.

Using Death-certificate-only Cases

It is sometimes suggested that cases identified only from death-certificate (DC) information (death-certificate-only cases) be included in analyses of cancer incidence. To study the consequences of including death-certificate-only cases, we will extend Equation (2) to account for persons who were not residents of the tumor registry at the time of cancer, time t, but who died before the end of follow-up, time T, and whose cancer was detected on the death certificate at time T. This leads to

$$\lambda^{**}(t) = \frac{\pi(t)\lambda(t)F(t) + [1-\pi(t)]\lambda(t)F(t)\int_t^T \omega(u-t;t)g(u-t;t)du}{F(t) + \int_0^t [1-\pi(u)]\lambda(u)F(u)\{G(T-u;u) + \int_t^T g(v-u;u)[1-w(v-u;u)]dv\}du}, \quad (13)$$

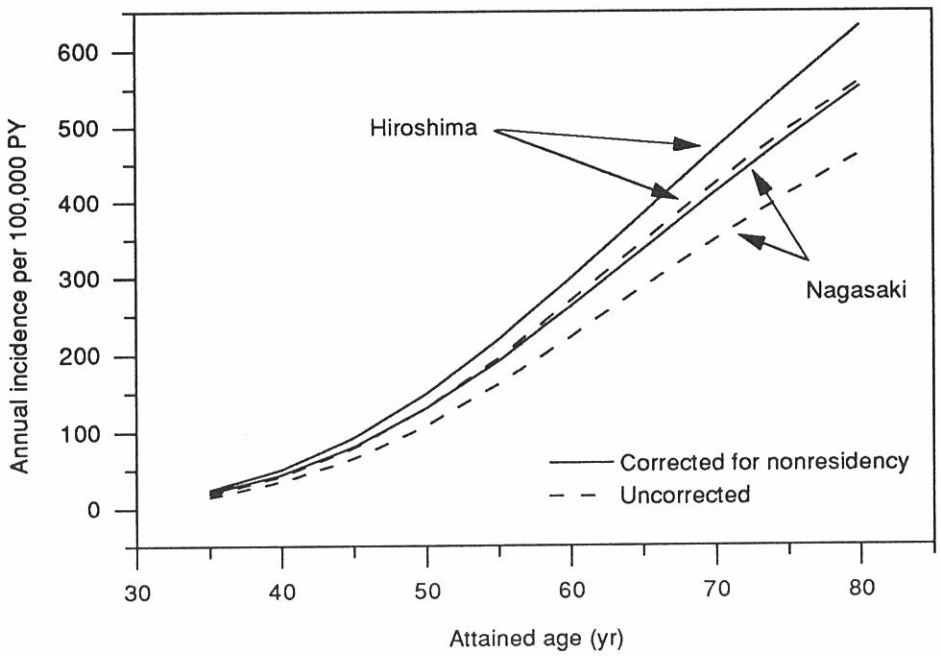


Figure 2. Estimated background cases of stomach cancer per 100,000 person-years (PY) for persons who were 20 yr old at the time of the bombings, with (solid line) and without (dashed line) correction for nonresidency.

where $\omega(s;t)$ is the proportion of cancers diagnosed at time t that are noted on the death certificate for individuals dying at time $t + s$. In the case of type-specific cancer incidence, $\lambda(t)$ is replaced by $\lambda_j(t)$ [Equation (6)], and $\omega(\cdot,\cdot)$ by a type-specific function $\omega_j(\cdot,\cdot)$, where j is an index of cancer type. Note that if cancer were instantaneously fatal and all cancers were reported on the DC, Equation (13) would reduce to the incidence function for cancer, $\lambda(t)$, and if cancer were never recorded on the DC, then this reduces to Equation (2).

When $\lambda(t)$ is sufficiently small, Equation (13) becomes

$$\lambda^{**}(t) = \lambda(t)\{\pi(t) + [1 - \pi(t)]\int_t^T \omega(u - t;t)g(u - t;t)du\} \quad . \tag{14}$$

If DC cases were included in the tumor-registry analyses, one would need estimates of the functions $g(t,s)$ and $\omega(t,s)$ [or $\omega_j(t,s)$ in the case of type-specific analysis] to make the correction similar to that based on Equation (5).

Discussion

We have discussed the effect of nonresidency on estimates of cancer incidence in a tumor-registry-based cohort study and presented a justification for a simple method of correcting for nonresidency in Poisson regression analysis. Ignoring the problem of nonresidency results in underestimation of cancer incidence rates. In a simulated data set based on total cancer incidence in Denmark, the simple correction to the Poisson regression analysis based on known residency probabilities gave essentially the same results as a complete-data maximum-likelihood solution to the problem, as did an approximate analysis in which known individual residency data were used. Hence, in the case of cancer incidence and the levels of nonresidency used here, there seems to be little advantage to using the complete-data maximum-likelihood approach if either residency histories or residency probabilities are available. Clearly, known residency histories would be preferable to estimates of residency probabilities because one can never be certain of the applicability of estimated residency probabilities.

In our example using stomach-cancer incidence in the RERF tumor registry, ignoring nonresidency did not result in noticeable bias in estimates of the radiation effect in a relative risk model. However, estimates of absolute incidence were biased, as were estimates of excess cases. Bias of this sort would similarly also be reflected in estimates of additive risk. Even relative-risk estimates would be biased if residency rates depended on the exposure of interest. Although the degree to which absolute risk is underestimated in the RERF tumor registry by ignoring nonresidency is arguably small (only 7%), as this cohort ages and individuals who have higher nonresidency rates reach the age of high cancer risk, the bias in absolute-risk estimates will be larger.

Our arguments were based on an idealized model of how cancers are obtained in the tumor registry that ignores a number of problems that should be considered in practice. The most important of these is that we have assumed that the tumor registry is 100% effective in detecting cancers that occur among residents of the tumor-registry area and 100% certain of the incidence time. This is clearly not the case, since some cancers go unreported, are missed, or are misdiagnosed,

with the probability of these events depending on the cancer type. We have also assumed that the tumor-registry catchment area provides essentially 100% coverage of the area of residency of members of the cohort, which also may not be strictly true. Some cancers that are initially diagnosed outside the tumor-registry area are detected nonetheless because of subsequent visits to a physician within the tumor-registry area. Finally, in our analyses we have ignored the uncertainty in estimates of the residency probabilities $\pi(t)$ and simply assumed that these are known functions. With the exception of the last, all of these problems are prevalent in all tumor-registry-based studies even without a non-residency problem. Although the methods described above could be extended to include these aspects, the simple correction for nonresidency proposed here is a reasonable way to address one of several potential problems with these types of studies. The impact of ignoring some of these problems is small. Formally including all of them in the mathematical treatment would have obscured the most important aspects of problems of nonresidency.

The simple correction proposed here is reasonable for cancer, which has relatively low incidence compared with other diseases as a group, but is not exact and may not be adequate for higher-incidence diseases. The exact maximum-likelihood method would be computationally impractical in the RERF analyses. The EM algorithm-based method proposed by Espeland et al (1989) for the similar problem of diagnostic misclassification in longitudinal studies would likewise be difficult to apply in a large cohort.

The brief discussion of the cases of incident cancers being discovered on the DC shows that the proper correction when death-certificate-only cases are included is much more complicated than when they are not included. In particular, estimates of the survival density for individuals with cancer as well as time-dependent estimates of the probability that the cancer is recorded on the death certificate are required. Different estimates would be required for each type of cancer considered. These estimates are not readily available and would be difficult to compute. Further study is required to determine whether the additional quantities are important, or whether the advantage of including death-certificate-only cases outweighs the additional uncertainties and difficulties introduced by the requirement to estimate these quantities.

Acknowledgments

We would like to thank Dr. M. Yamada of the Department of Clinical Studies and Dr. F. L. Wong of the Department of Statistics for their work in compiling the AHS patient contact data into analyzable form and Dr. K. Kodama, chief of the Department of Clinical Studies, for making these data available.

References

Beebe GW, Ishida M: Joint JNIH–ABCC study of life span of atomic bomb survivors. Research plan. ABCC TR 4-59

Clayton D: The analysis of event history data: A review of progress and outstanding problems. Stat Med 7: 819–41, 1988

Espeland MA, Platt OS, Gallagher D: Joint estimation of incidence and diagnostic error rates from irregular longitudinal data. J Am Stat Assoc 84: 972–9, 1989

Hollingsworth JW, Beebe GW: ABCC–JNIH Adult Health Study. Provisional research plan. ABCC TR 9-60

Muir C, Waterhouse J, Mack T, Powell J, Whelan S: Cancer Incidence in Five Continents (Vol 5). Lyon, France, IARC Scientific Publications, #88, pp 453–7, 1987

Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE: The analysis of failure times in the presence of competing risks. Biometrics 34, 541–54, 1978

Preston DL, Lubin JH, Pierce DA: Epicure: Users Guide. Seattle, Wash, HiroSoft International, 1993

Thompson D, Mabuchi K, Ron E, Soda M, Tokunaga M, Ochikubo S, Sugimoto S, Ikeda T, Terasaki M, Izumi S, Preston D: Cancer incidence in atomic bomb survivors, Part II: Solid tumors, 1958–87. Radiat Res (in press)

Appendix

Estimates of Probability of Residency

The RERF Adult Health Study (AHS) is a subcohort of the approximately 20,000 members of the LSS who have been invited since 1958 to participate in biennial physical examinations (Hollingsworth and Beebe, 1960). The AHS patient contact data were used to compute the probability that an individual was in the tumor-registry area for the analysis of stomach-cancer incidence presented above. These data consist of 2-yr-cycle-specific information on whether a member of the AHS was examined, was contacted in the area but refused, had moved out of the area, or had died and the corresponding dates of these events. From these data, dates of immigration and emigration transition events were recorded, along with information on covariates that could affect the rates of such events. We assumed that the experience of the AHS cohort was similar to that of the full LSS cohort from which it is drawn.

The analysis involved estimating hazard functions for both emigration from the tumor-registry area and immigration to the tumor-registry area. The form of the model for the hazard function, which was the same for both in and out transitions, was the log-linear function

$$\psi_{ijk} = \exp(\nu_0 + \nu_{Ci} + \nu_{Aj} + \nu_{Tk}) \quad , \tag{15}$$

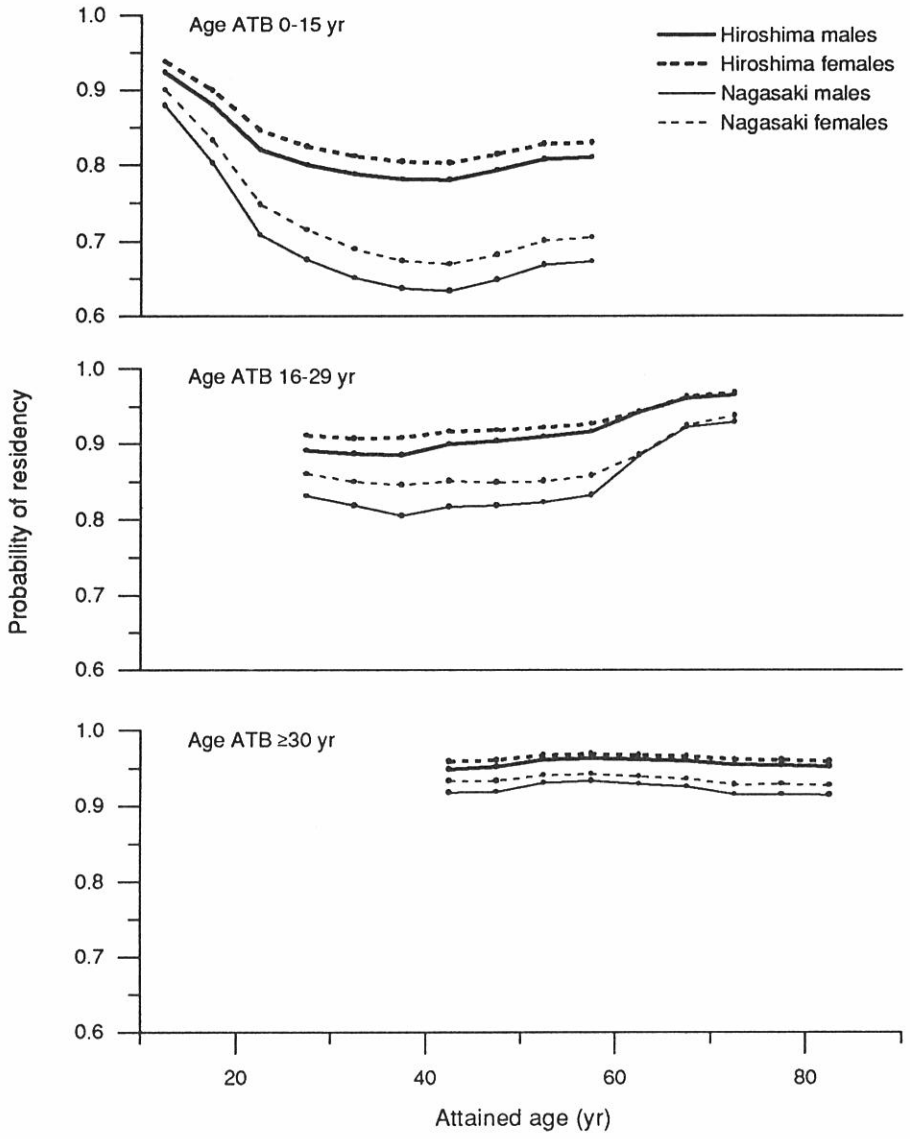
where υ_{Ci} represents the city effect (i = 1, 2: Hiroshima or Nagasaki), υ_{Aj} is the effect of age ATB (j = 1, 13: 5-year categories), and υ_{Tk} is the effect of calendar time (k = 1, 7: with the first interval from the beginning of 1958 to the end of 1960, followed by 5-yr intervals through the end of 1985, and a final interval from the end of 1985 to the end of 1987.) Separate models were fit for males and females.

Let T_k be the right endpoint of the time intervals, where $T_0 = 0$. If it is assumed that migration follows a simple Markov process with piecewise-constant hazards of transition $\psi_{ijk}{}^{\mathrm{I}}$ and $\psi_{ijk}{}^{\mathrm{O}}$ in period k for in (I) and out (O) transitions, respectively; then it can be shown that the probability of being in the area at the midpoint of the kth interval is

$$\begin{aligned}\pi_{ij}(M_k) = {} & \pi_{ij}(T_{k-1})\exp[-0.5\,(\psi^{\mathrm{I}}_{ijk} + \psi^{\mathrm{O}}_{ijk})(T_k - T_{k-1})] \\ & + \frac{\psi^{\mathrm{I}}_{ijk}}{\psi^{\mathrm{I}}_{ijk} + \psi^{\mathrm{O}}_{ijk}}\{1 - \exp[-0.5\,(\psi^{\mathrm{I}}_{ijk} + \psi^{\mathrm{O}}_{ijk})(T_k - T_{k-1})]\} \quad ,\end{aligned} \tag{16}$$

where $\pi_{ij}(t)$ is the probability of being a resident in the area at time t for an individual in city i with age ATB in category j, M_k is the midpoint of the kth interval, and $\pi_{ij}(0)$ is estimated as the proportion of persons who were resident in 1958. [Note that in the body of the report the quantities $\pi_{ij}(M_k)$ were expressed in different notation as $\tilde{\pi}_k$, where k indexed one cell in the person-years table.]

The Appendix Figure shows residency probabilities estimated from the AHS patient contact data, by attained age, city, and sex, averaged within three categories of age ATB.



Appendix Figure. Average estimated probability of residency among persons in the tumor registry for three ranges of age at the time of the bombings (ATB), by sex and city.