J Clin Epidemiol Vol. 49, No. 1, pp. 95–103, 1996

0895-4356/96/$15.00
SSDI 0895-4356(95)00037-5

ELSEVIER

Estimating the Costs Attributable to a Disease with Application to Ovarian Cancer

Ruth Etzioni, Nicole Urban, and Mary Baker
FRED HUTCHINSON CANCER RESEARCH CENTER, SEATTLE, WASHINGTON

ABSTRACT. This article is concerned with the methodological issues that arise when estimating the expected costs attributable to a disease. In particular, the article considers methods appropriate for handling incomplete or censored cost and survival data, incorporating discounting, and computing attributable costs. After motivating the need for an estimate of the average, present value of the attributable costs, we present the Kaplan–Meier sample-average (KMSA) estimator, which takes into account the censored nature of the data that are typically available. We investigate the statistical properties of the estimator and compare it to others employed in the literature, showing how certain methods for incorporating discounting can introduce bias. We demonstrate the utility of the estimator by applying it to estimation of the costs attributable to ovarian cancer, using data from a database linking Medicare claims with the Surveillance, Epidemiology, and End Results cancer registry. Our analysis suggests that the average, present value of the 15-year costs attributable to ovarian cancer is $21,285 for local-stage cases and $32,126 for distant-stage cases in 1990 dollars. J CLIN EPIDEMIOL 49;1:95–103, 1996.

1. INTRODUCTION

The health care costs attributable to a disease are of considerable interest to policy analysts. They measure an important component of the burden to society of the disease [1]. Accordingly, they contribute significantly to the cost-effectiveness of strategies to prevent or detect the disease [2,3]. This article is concerned with the estimation of disease-attributable health care costs.

For analytical purposes, the present value of a stream of health care costs accumulating over time is usually needed. The *present value* of a stream of costs refers to the value of the accumulated costs at the start of their accrual. Use of an appropriate discount rate takes into account the fact that costs in the future are valued less than immediate costs. In the *incidence approach* to measuring the costs of a disease, adopted in this article, all costs are discounted to the time at which the disease is incident [1]. Section 2 reviews the economic concepts and terminology used, including the present value.

This article focuses on methods for estimating the average present value of the medical costs attributable to a disease, where the average is over the population of disease cases. Of interest are the costs of medical and institutional care associated with treating the disease, which are incurred after the disease is diagnosed and until the individual dies. Although the average present value of the attributable cost is routinely reported, methods for its estimation are seldom described in the cost-effectiveness literature. Determining an unbiased estimator of this quantity raises several statistical issues; this article enumerates these issues, and proposes some solutions.

Medicare claims are a typical source of data for estimating costs attributable to a disease. Such data have been used to estimate costs attributable to cancer by dividing the period postdiagnosis into initial, maintenance, and terminal phases [4,5]. In the absence of discounting, the present value of the attributable costs can be calculated from cost data and an estimate of the average duration of the maintenance phase. However, knowledge of the expected length of the maintenance phase does not suffice in the presence of discounting. To estimate the present value of the costs in an unbiased fashion, the entire survival distribution is necessary, for reasons detailed in Section 3.

Section 3 presents a method of estimating the present value of the health care costs attributable to cancer. A key feature of the method is that it takes the censored nature of the data into account. The estimator is fairly straightforward to compute, and is designed for potential use by policy analysts, clinicians, and, in general, those involved in cancer prevention and screening research—the intended audience of this article. Variants of the estimator have been used before to estimate the external costs of smoking and excess alcohol use [6] and lack of physical exercise [7], as well as the lifetime costs attributable to smoking [8]. The estimator is similar to the pathwise estimator [9], proposed by Lin *et al.* [10] for estimating the costs of care associated with a disease. It is suitable for practical use, since it can be applied to cost data collected over a short period of calendar time rather than for the entire duration of postdiagnosis followup. The cost data and survival data need not come from the same source. We illustrate the method in Section 4, where we estimate the costs attributable to ovarian cancer among Medicare enrollees.

To summarize, Section 2 reviews economic terminology, and motivates the use of the present value and of the attributable costs. Section 2 also suggests how the estimated attributable costs may be used in evaluating the cost-effectiveness of screening interventions for early disease detection. Section 3 develops a method for estimating the present value of the costs of care attributable to a disease. Section 3 also considers the issue of extrapolation beyond the time period for which cost and survival data are available. Section 4 applies the method to ovarian cancer, describing our data sources in some detail. Finally, Section 5 presents some discussion, comparing the Kaplan–Meier sample-average (KMSA) estimator with others proposed for the cost estimation problem.

2. TERMINOLOGY

We begin by reviewing the economic terminology used in this article. In particular, we motivate and explain the concepts of *present value*

(Received in revised form 16 September 1994).

and *attributable* costs. We consider only *direct costs*, or costs of goods and services actually used in treatment of the disease. This is in contrast to indirect costs, which refer to lost earnings associated with disease morbidity and mortality. References [1–3] and [11] provide detailed discussions of the concepts introduced here and of other relevant issues.

In estimating the costs of postdiagnosis care, we are effectively summing a series, or stream, of costs that begins at diagnosis and continues until death. We take the incidence approach [1] to estimating the costs, which means that we are interested in their value when the disease is incident, that is, at the time of diagnosis. When costs are specified over time, but assessed at one point in time, the comparability of present and future costs is an issue. In general, because of society's preference to have things now, rather than in the future, a dollar amount in the future is worth less now than the same dollar amount in the present. This time preference reflects uncertainty about the future, as well as the fact that something available now can be invested or put to productive use, generating returns that accrue over time. Time preference is not the same as inflation; it would still exist even if there were no inflation. Inflation is taken into account as well, by measuring all costs in the same year's dollars regardless of when they are incurred.

The implementation of society's time preference is through *discounting*. If the year of disease incidence is year 0, then costs in year j are devalued by a factor that is a function of j and the real rate of return on investment, termed the discount rate. Thus, an amount of 100 dollars j years from diagnosis is worth $\$100/(1 + d)^j$ at diagnosis, where d is the annual discount rate. The choice of discount rate is a somewhat controversial topic; we use 5% in this article for consistency with the literature on the cost-effectiveness of prevention (Ref. [3], p. 116).

The *present value* of postdiagnosis cost refers to the accumulated, discounted costs at the time of diagnosis of the disease. We focus here on methods for estimating the average present value of the cost, where the average is per-disease case. The *average present value* indicates that the sequence of operations is first to discount the costs corresponding to each period postdiagnosis, and then to average the accumulated, discounted costs over the population of cases. This is in contrast to the *present value of the average* cost, which Section 3.4 examines in more detail.

In this article, we focus on costs *attributable* to the disease, also described as marginal or incremental costs. These are the extra medical costs incurred by disease cases over the costs that would have been incurred had they not been diagnosed with the disease. The attributable costs are particularly important in evaluating the cost-effectiveness of strategies to prevent disease, since cost-effectiveness analysis involves comparing incremental, or attributable benefits, to attributable costs [11]. We discuss methods for computing attributable costs in Section 3, and implement one of these methods in our analysis of ovarian cancer costs in Section 4.

In the case of screening, cost-effectiveness is expressed as a ratio that measures the cost per year of life saved attributable to screening. The numerator of the ratio (net money cost) is the cost of the screening, plus the cost of diagnostic workup for the false positives, less the savings in treatment costs attributable to earlier diagnosis among some proportion of the incident cases. The denominator of the ratio (effectiveness) is the years of life saved attributable to earlier diagnosis among the same proportion of the cases, net of any loss in life years attributable to risks associated with definitively diagnosing cases who screen positive, but are free of the disease. This relationship can be summarized in an equation that can be expressed equivalently for a population in the aggregate, or per participant in screening.

The cost effectiveness of screening is estimated by the following equation:

$$[C/E]_{scr} = (c_{scr} + c_{dx} + c_{tr})/pyls$$

where c_{sr} is the cost of screening, c_{dx} is the expected cost of diagnosis attributable to screening, c_{tr} is the expected cost of treatment attributable to screening, and $pyls$ represents potential years of life saved attributable to screening.

In some cases, c_{tr} may be negative if treatment costs associated with earlier diagnosis are larger than those associated with later diagnosis. The methods presented in this article may be used to calculate c_{tr}. If savings in treatment costs are sufficient to offset the costs of screening and diagnosis, then screening is said to be *cost-saving*. Otherwise, a cost-effectiveness ratio is calculated that measures the cost per additional year of life that can be saved through screening.

3. METHODS

In this section, we derive a method for estimating the average present value of the costs attributable to a disease. We begin by considering how one might, from first principles, estimate the average postdiagnosis costs. First, we assume that followup is complete, so that the date of death is known for each case. We specify an estimator of the average cost, and then adjust the estimator for incomplete followup, or censoring. We then suggest a method for obtaining attributable costs, and show how to incorporate discounting in our estimation procedure, to yield the average, present value of the attributable costs. We also consider the issue of extrapolation beyond the period for which survival and cost data are available.

Our methods make use of information on postdiagnosis costs, as well as on survival post diagnosis. In our presentation of the methods, we assume that cost data are available for the entire period over which survival data are available. In other words, the monthly, quarterly, or annual costs of a patient are available from the time of diagnosis (and, possibly, before diagnosis) through the time of death or loss to followup. Later, we address estimation from less complete cost data.

3.1. *The KMSA Estimator*

We begin by considering the case of complete followup, that is, where each case in the sample has a death date recorded. Let $c(i, j)$ denote the cost incurred by case i in time period j. Once a case survives to the beginning of a given time period, costs are recorded for her in that time period. Thus, a case who survives to the beginning of month 3 postdiagnosis, but who dies during month 3, will have costs recorded for month 3. Let n_j denote the number of cases alive at the beginning of month j; n_1 is the number of cases diagnosed. Then the sample average cost is equal to the total costs divided by the total number of patients:

$$\begin{aligned}\text{avcost} &= (\text{total cost in month 1} + \text{total cost in month 2}\\ &\qquad + \text{total cost in months 3} + \cdots)/(\text{no. of cases}) \qquad (1)\\ &= \left(\sum_{i=1}^{n_1} c(i, 1) + \sum_{i=1}^{n_2} c(i, 2) + \sum_{i=1}^{n_3} c(i, 3) + \cdots\right)\Big/ n_1\end{aligned}$$

Let $\bar{c}(j)$ denote the average cost in time period j. Then Eq. (2) can be rewritten:

$$\text{avcost} = \bar{c}(1) + \frac{n_2\bar{c}(2)}{n_1} + \frac{n_3\bar{c}(3)}{n_1} + \cdots \qquad (2)$$

where n_j/n_1 is the proportion of cases surviving to the beginning of interval j.

If all cases are followed from diagnosis until death, it is straightforward to estimate Eq. (2). However, this is rarely the case. Patients whose followup terminates while they are still alive are said to be censored at their time of last contact—all that is known about their survival time is that it exceeds the time from diagnosis to last contact. A consequence of censoring is that the proportion of cases surviving to the beginning of each time interval is not known; thus, Eq. (2) cannot be computed. However, if the censored cases are not at especially high or low risk of death relative to the noncensored cases, then the *Kaplan–Meier* or *product-limit* [12] estimator provides an unbiased estimator of the quantities n_j/n_1.

The Kaplan–Meier estimator of the probability of surviving to time t can be defined as follows. Suppose that there are n subjects, with k distinct death times, $t_1,t_2, \ldots, t_k$. Let d_j be the number of deaths at t_j, and let n_j be the number of subjects at risk just prior to t_j. For example, if there are 10 subjects, 2 of whom die in month 1, and 2 of whom are lost to followup during month 1, then $t_1 = 1$, $n_1 = 10$, $d_1 = 2$, and $n_2 = 6$. The Kaplan–Meier estimator of the probability of surviving to the beginning of time period t is given by

$$\hat{S}_t = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j} \tag{3}$$

Substitution of the Kaplan–Meier estimate $\hat{S}_j$ for n_j/n_1 in Eq. (2) yields

$$\text{avcost} = \bar{c}(1) + \hat{S}_2\bar{c}(2) + \hat{S}_3\bar{c}(3) + \cdots \tag{4}$$

In other words, the estimate of the average lifetime cost of care is given by

$$\text{avcost} = \sum_j \text{Prob(alive at } j\text{)}. \tag{5}$$
$$\text{mean(cost at } j \text{ given alive at } j\text{)}$$

where the probability of being alive at j is the probability of surviving to the start of time interval j, and is estimated by the Kaplan–Meier survival probability. The mean(cost at j given alive at j) is estimated using cost data from subjects who survive to the start of time interval j. The estimate is based on subject costs during time period j; this includes costs of cases who survive to the start of time period j, but who die or who are censored during time period j. In other words, death costs are included in the average cost estimates for each time interval. The costs for individuals censored during the interval are not prorated to take the time of censoring into account; this may lead to some downward bias in the mean cost estimates if the time intervals are wide. Consequently, we use monthly rather than yearly intervals, when monthly data are available.

We refer to the estimator [Eq. (5)] of average cost as the KMSA (Kaplan–Meier sample-average) estimator. It is an intuitive quantity, which is simple to compute. It is also a *consistent* estimator, in the sense that it is close to the true value in large samples, if the following two conditions are satisfied:

1. The time interval is fine enough, so that censoring of costs does not bias the estimates $\bar{c}_j$ downward. For practical purposes, monthly time intervals should suffice, but annual time intervals may be too broad. Cases censored during the year may significantly undercontribute to the cost sample mean for that year.
2. The censoring mechanism is independent of survival and cost in the sense that individuals are not censored because they appear to be at especially high or low risk of death, and, in addition, they are not censored because of accrual of especially high or low costs relative to the uncensored cases.

Lin and Etzioni [13] present a proof of the consistency of the KMSA estimator. The independence of censoring and survival is satisfied under type I censoring, that is, when followup terminates at a certain time, at which point all surviving individuals are censored. S. Taplin (personal communication) has pointed out that the independence of the censoring mechanism may be questionable when the cost source is insurance data, and censoring is due to individuals switching from one insurance provider to another. Under such circumstances, individuals in poor health and with higher ongoing costs will be reluctant to switch providers, and therefore will tend to be censored less frequently than those in good health.

3.2. *Attributable Costs*

There are several alternative approaches to estimating the costs attributable to a disease. One approach is to count the health services associated with the disease, multiply each service by its unit cost, and aggregate the results to yield an estimate of attributable costs. Unless an automated system tracks utilization, this method requires chart or claims review. In addition, judgment is necessary as to whether each service is related to the disease of interest rather than to coexisting conditions.

An alternative method, used by Baker *et al.* [4], employs a suitable set of controls, or individuals free of the disease, which may be matched to cases on the basis of age, sex, and/or general health indicators. Average costs per time period are estimated for the controls as for the cases, and subtracted from the costs of the cases to yield attributable costs. The KMSA estimate of attributable costs is given by Eq. (4) with $\bar{c}(j)$ replaced by $\bar{c}_j - \bar{c}_j^0$, the difference between the average case cost and the average control cost in time interval j. We employ this method of calculating attributable costs in Section 4.

3.3. *Discounting*

Each of the possible methods described for generating attributable costs yields a stream of costs by time period postdiagnosis. To obtain their present value, these costs must be discounted, as explained in Section 1, to the time of diagnosis. The incorporation of discounting is straightforward for the KMSA estimator. With a discount rate d, the KMSA estimator becomes

$$\bar{c}(1) + \frac{\hat{S}_2\bar{c}(2)}{(1 + d)} + \frac{\hat{S}_3\bar{c}(3)}{(1 + d)^2} + \cdots \tag{6}$$

Note that the costs in each time period are discounted first, then entered into the KMSA expression. Thus, the KMSA estimator computes the average, present value of the cost. This is in contrast to the present value of the average cost, a concept we focus on in the following section.

3.4. *Extrapolation*

The KMSA estimator is based on estimates of the entire survival distribution and the corresponding costs. Often, the tail of the survival distribution is not identified, due to termination of followup. This is especially true in patients with early-stage diagnoses. In general, costs beyond a certain point postdiagnosis are also typically unavailable. In this section, we consider some problems that may arise when extrapolating beyond the period of followup, and propose an approach that avoids such problems.

A common approach to extrapolation is to make assumptions about survival and costs beyond the followup period. For example, suppose cost data are collected for 10 years postdiagnosis. In such a case, it might be assumed that average continuing per-annum costs remain at the 10-year level. If life expectancy or expected survival is known, then, in the absence of discounting, the average cost beyond 10 years can be calculated by counting costs up to life expectancy, without further knowledge about the shape of the survival distribution. However, information about the tail of the survival distribution is still necessary to estimate discounted costs. The remainder of this section illustrates the kinds of biases that may arise when discounting costs that are counted only up to life expectancy.

Let N denote the postdiagnosis survival of a patient, with expected survival from diagnosis $M = E(N)$. Suppose, for simplicity, that the annual postdiagnosis cost is constant; denote this annual cost by c. Denote the annual discount rate by d. Then, counting costs only up to life expectancy yields the following estimate of the average present value of the cost:

$$\sum_{i=1}^{M} \frac{c}{(1+d)^i} = \frac{c}{d}\left[1 - \left(\frac{1}{1+d}\right)^M\right] \tag{7}$$

This expression discounts the stream of costs for the average person. It is therefore more accurately described as the present value of the average cost, rather than as the average present value of the cost, which would be given by

$$\text{mean}\left[\sum_{i=1}^{N} \frac{c}{1+d)^i}\right] = \frac{c}{d}\text{mean}\left[1 - \left(\frac{1}{1+d}\right)^N\right] \tag{8}$$

Note that computation of the average present value of the cost [Eq. (8)] relies on knowledge of the distribution of N, while the present value of the average cost [Eq. (7)] only requires knowledge of the mean of N. Baker *et al.* [4] use Eq. (7) to estimate the average cost of care in the maintenance period, a period of relatively constant cost following the 3 months postdiagnosis, and preceding the terminal phase of the disease.

Although Eqs. (8) and (7) are similar, they are equal only if the undiscounted cost in year i can be expressed as

$$\text{cost}(i) = (a + bi)(1 + d)^i \tag{9}$$

where a and b are constants. We can go further than noting that Eq. (8) is not equal to Eq. (7) by observing that Eq. (8) is a *concave* function of N, that is, it has a negative second derivative. In this case, we can invoke Jensen's inequality [14], which states

> If g is a concave function of X, then the mean of $g(X)$ is less than or equal to $g[\text{mean}(X)]$.

This implies that Eq. (8) is always less than or equal to Eq. (7). In other words, the present value of the average cost is an overestimate of the average present value of the cost when the cost per time period is constant. The extent of the overestimation depends on several factors, primarily the distribution of N. For example, suppose N has an exponential distribution with mean 9.25 years and c = $5316 (the average maintenance period and annual cost for breast cancer from Baker *et al.* [4]. Then, assuming a 5% discount rate, the average present value of the maintenance cost is approximately equal to $36,635,[1] while the present value of the average maintenance cost is equal to $42,103, a 15% overestimate.

We conclude that coping with extrapolation by counting costs only up to life expectancy can lead to biased estimates of the present value. To avoid extrapolating beyond time periods for which information is available, it may be more reasonable to consider the average costs of care for a fixed period postdiagnosis, such as 10 or 15 years. This would appear to be especially reasonable if attributable costs are of interest, and these costs can be assumed to be negligible beyond 10 or 15 years postdiagnosis for patients surviving this long; this may be checked empirically, if appropriate control data are available. In Section 4, we compute the average 15-year costs of care for ovarian cancer.

4. APPLICATION TO OVARIAN CANCER

This section describes our application of the KMSA estimator to obtain the average cost attributable to epithelial ovarian cancer among Medicare enrollees. Our cost data source consists of a database linking Medicare claims data with the Surveillance, Epidemiology, and End Results (SEER) cancer registry database. Potosky *et al.* [15] describe in detail the components and creation of this linked database. Payments as well as charges are recorded by Medicare; payments, henceforth referred to as costs, generally amount to roughly two-thirds of the charges. The cost data cover the years 1984 through 1990, and contain cost information on all Medicare patients alive at some point during this period, who could be reliably matched to patients recorded by SEER. Consequently, different patients have cost data covering different periods in the disease process, and comprise what might be termed a *synthetic cohort*. For example, a patient diagnosed in 1986 may have costs recorded for the 2 years preceding diagnosis, and up to 4 years postdiagnosis, while a patient diagnosed in 1980 and who survives to 1990 may have costs recorded from years 4 to 10 postdiagnosis.

Although it is possible to estimate survival probabilities from such data, we note that the data exclude patients dying before 1983, and censor cases in 1990. We use the SEER database instead to estimate survival, since it includes all cases diagnosed from 1973 through 1990, and contains more complete followup than does the cost database. The survival estimates from SEER and the cost estimates from the SEER–Medicare linked database are combined to produce the KMSA estimator of attributable costs.

We begin by describing our data sources. We then move on to detail our computation of control costs and calculation of attributable costs. We present the KMSA estimates of the average present value of the attributable costs, for all cases, and separately by clinical stage at diagnosis. For comparison purposes, we also present estimates of the average present value of the total postdiagnosis costs, including comorbidity as well as disease-attributable costs, and the average attributable and total costs, without discounting. Finally, we note some limitations of our analysis and point out some possible extensions and refinements.

4.1. *Data Sources*

4.1.1. COST DATA. The data set consisted of 5012 Medicare beneficiaries above the age of 65 years, diagnosed with ovarian cancer between 1973 and 1989 inclusive. Since our analysis concerned screening using the CA125 radioimmunoassay, which identifies only epithelial malignancies, we selected only cases with epithelial ovarian cancer. Patients whose clinical stage could not be determined, and patients diagnosed below age 65 years, were excluded. Only cases over age 65 years were used, since patients in the database diagnosed at earlier ages were either entitled to Medicare below age 65 years or became entitled at age 65 years. This latter group of patients tended to be diagnosed

[1] Based on Monte Carlo simulation, with 10,000 replications.

relatively earlier in time, with cost data missing immediately postdiagnosis. We felt that these individuals constituted a special group and hence did not include them in the analysis.

Twenty-two percent of cases were classified as local stage at diagnosis, 20% were classified as regional, and 54% as distant. Four percent were unstaged and were excluded. Cases had cost data recorded for at least 1 month and as many as 84 months. Figures were recorded up to 18 years postdiagnosis for some patients. Cases had costs recorded only for months during which costs were incurred. For each case, we assigned zero costs to months during which no costs were recorded but patients were entitled to Medicare.

Information was available on inpatient and skilled nursing facility costs, home health and hospice services, physician services, and outpatient services. See Potosky *et al.* [15] for a detailed description of the major Medicare files encoding these costs.

Costs were assigned to month according to the date of service recorded by Medicare. Inpatient costs were apportioned uniformly over the months comprising each hospital stay. Since only the year of office visits was identified, the office visit costs were spread uniformly over the months within the relevant calendar year. Riley *et al.* [5] used a similar scheme, noting that office visit charges accounted for only 5% of total payments for their sample.

Costs were adjusted to 1990 dollars using the Medical Care Price Component of the Consumer Price Index, to avoid distortion from inflation between 1984 and 1990. All of the monthly costs were accumulated and recorded for each patient by month postdiagnosis. The starting time for the analysis was taken to be 1 month prior to diagnosis, to take into account the costs associated with making the definitive diagnosis. Some patients have costs recorded even before this time, so it is therefore possible, in principle, to take costs before this time into account.

Figure 1 plots the monthly cost in 1990 dollars by month from diagnosis for a local stage patient who died in month 35 postdiagnosis. Notice the U-shaped pattern of the postdiagnosis costs. Costs are highest at diagnosis and death and lower in between. Figure 2 shows boxplots of the costs for distant-stage cases by month for the first year postdiagnosis. The boxplot for a given month shows the cost among patients surviving to the beginning of that month, thus it includes the costs for the patients who died in that month.

4.1.2. SURVIVAL DATA. We obtained survival data from the SEER database. This program is sponsored by the National Cancer Institute, and collects data from nine population-based tumor registries, covering about 10% of the U.S. population. Data recorded include month and year of diagnosis, age at diagnosis, type of cancer, clinical stage at diagnosis, histology, and month and year of death. For the analysis, data consisted of 9700 epithelial ovarian cancer cases, who were diag-

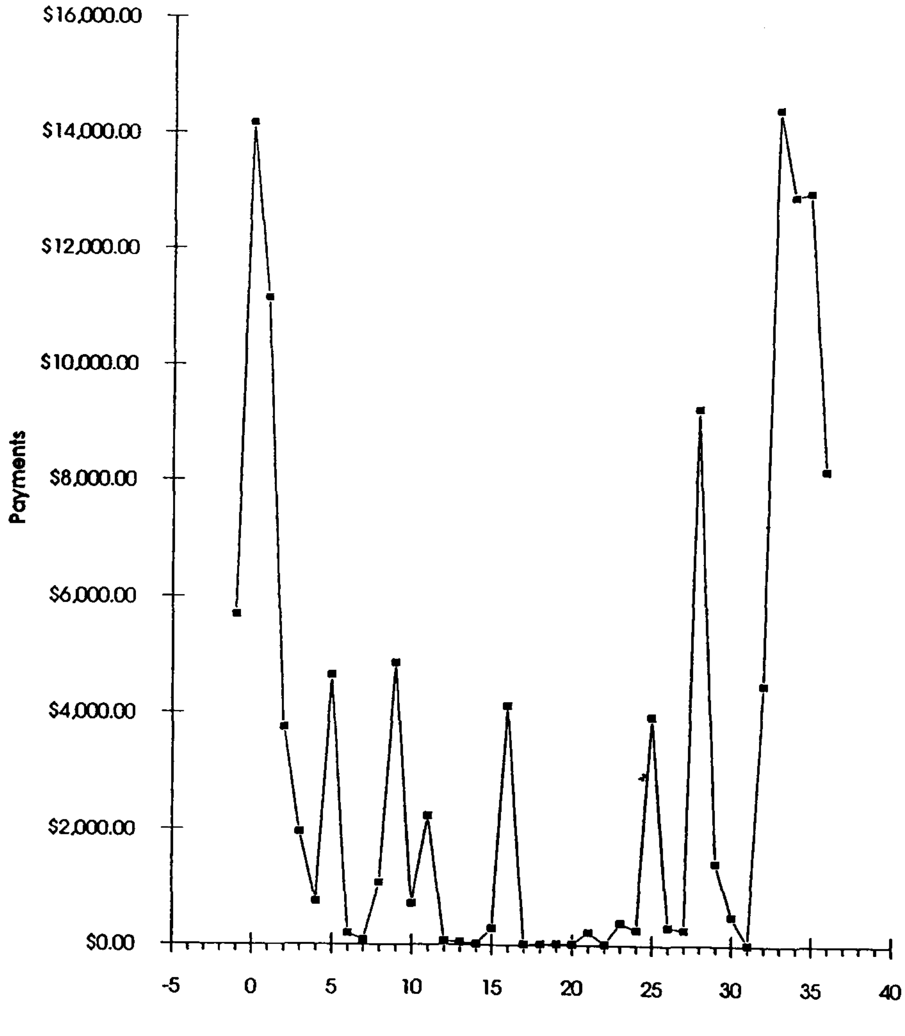

FIGURE 1. Monthly Medicare cost from 1 month prior to diagnosis for an early-stage ovarian cancer case with 3 years survival postdiagnosis. Month of diagnosis is month 0. (Source: Linked SEER–Medicare database.)

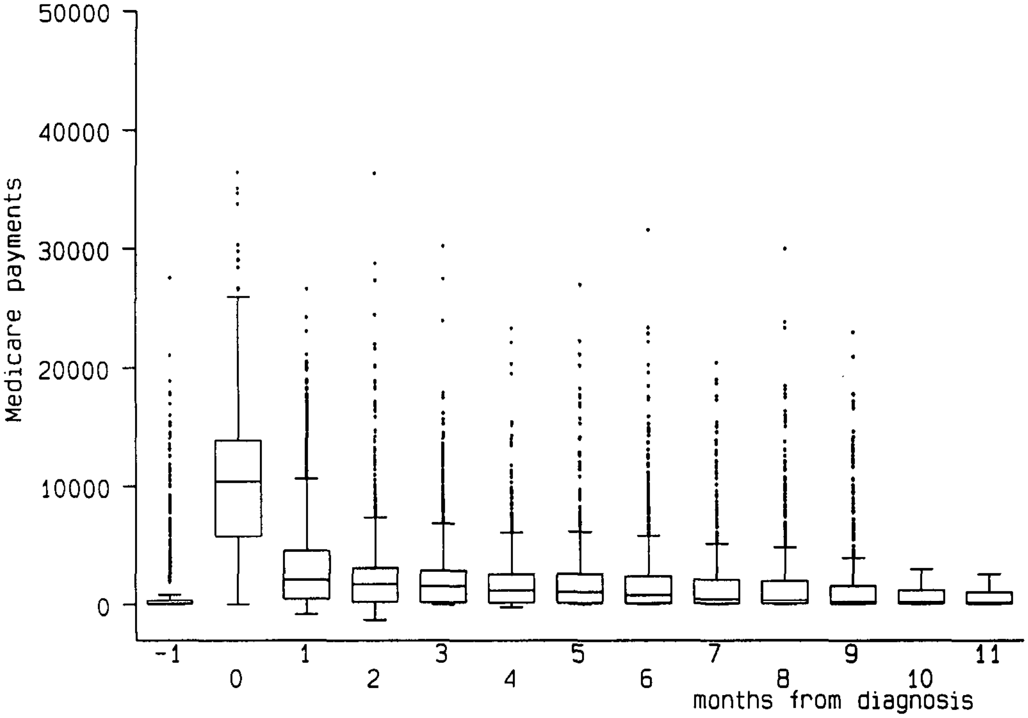

FIGURE 2. Boxplots of Medicare payments by month from diagnosis for distant-stage cases in the first year postdiagnosis. Month of diagnosis is month 0. The boxplot for a given month shows the cost among patients surviving to the beginning of that month, thus it includes the costs for the patients who died in that month. (Source: Linked SEER–Medicare database.)

nosed between 1973 and 1989 inclusive and who were age 65 years or above at diagnosis. Cases with missing stage information, and cases diagnosed at autopsy, were excluded. Seventeen percent of the cases were diagnosed as local stage, 14% as regional, and 64% as distant. Five percent were not staged and were excluded from analysis.

Figure 3 plots the Kaplan–Meier survival by month postdiagnosis, for the different clinical stages at diagnosis.

4.2. *Control Costs*

We used the control costs approach rather than the related service use approach to generate attributable costs. As described below, the control costs were constructed to increase with age postdiagnosis. These costs were based on 1988 data from Lubitz and Riley [16], who presented Medicare payments per person-year, separately for beneficiaries in the last year of life (decedents) and all others (survivors).

Single controls surviving to the start of month *j* include individuals in their last year of life as well as survivors, Medicare payments for both these groups were used in the computation of control costs. Denoting the average annual payment for survivors by C_s, the average annual payment for decedents by C_d and the probability of dying in the forthcoming 12 months by P, we computed the monthly control cost per individual as

$$C_0 = (1 - P)\frac{C_s}{12} + \frac{P}{12}C_d \qquad (10)$$

This expression assumes that the rate of death during a single year is constant, and that the annual cost for survivors may be apportioned equally over the 12 months of the year.

We obtained estimates of P from U.S. life tables [17], and calculated Eq. (10) separately for the age groups shown in Table 1. We then fit a least-squares regression line to the plot of C_0 against the midpoint of each age group; the least-squares fit had $R_2 = 0.92$, and suggested that the monthly costs increase with age by approximately $0.72 per month over the age range considered.

We used this estimate of the increase in costs with time to generate a stream of control costs corresponding to the stream of costs for each case. First, we extracted the control cost for the month of diagnosis from Table 1, to match the age group of the case at diagnosis. We then incremented the control cost by $0.72 each month postdiagnosis, to take into account the fact that the case was aging with time. As outlined in Section 3.2, we subtracted control costs from recorded postdiagnosis costs, to yield a stream of attributable postdiagnosis costs for each individual.

4.3. *Results*

We estimated the average 15-year postdiagnosis costs using the KMSA estimator, presented in Section 3. An annual discount rate of 5%, which implies a monthly discount rate of 0.4074%, was used in calculating the present value of costs. Results are presented in Table 2, which gives overall costs and the costs attributable to the disease, with and without discounting. The computation of the standard errors corresponding to the estimates is somewhat complex; Lin *et al.* [13] present a derivation of the asymptotic variance of the KMSA estimator that takes into account variations within intervals as well as the covariances among the intervals. We note that since 15-year survival of ovarian cancer patients is low, especially for regional and distant stage cases, the average 15-year postdiagnosis costs as presented in Table 2 closely approximate the average lifetime costs for these cases.

The estimated total costs with and without discounting are higher for local diagnoses than for distant diagnoses; this is a consequence of the lower survival rate for late-stage patients, which is apparently not completely offset by significantly higher treatment costs for late-stage

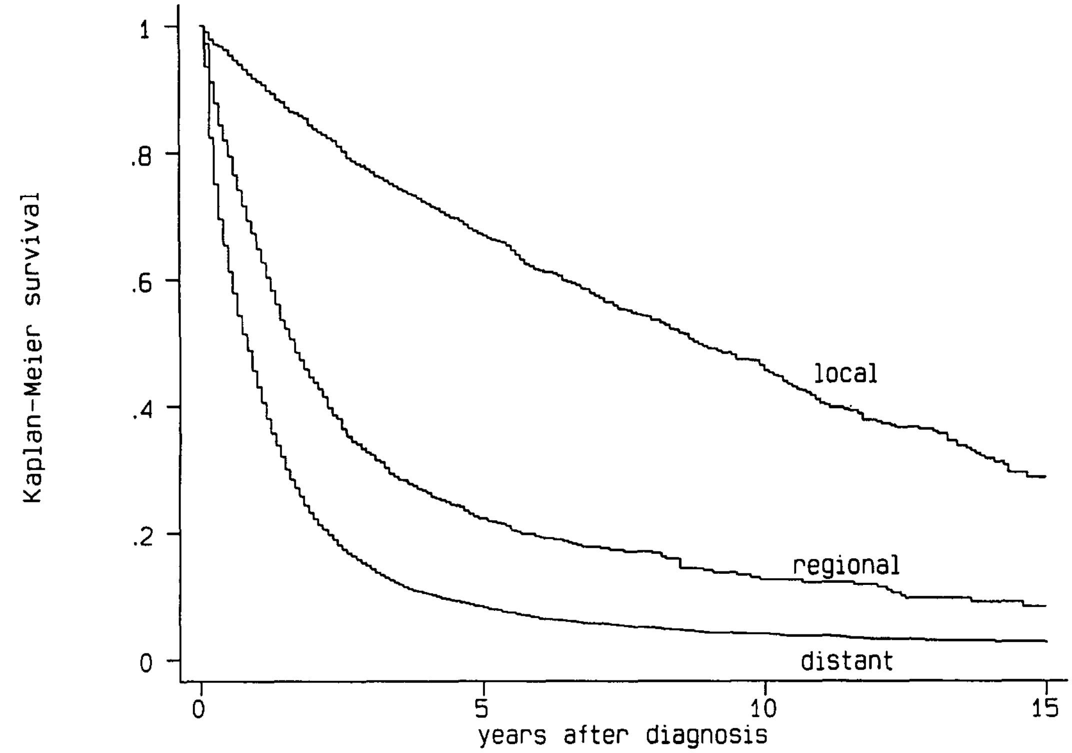

FIGURE 3. Kaplan–Meier survival by months from diagnosis and different clinical stage at diagnosis in women, age 65 years and above, with epithelial ovarian cancer. Month of diagnosis is month 0. (Source: SEER.)

ovarian cancer. A similar phenomenon has been noted by Riley *et al.* [5] in connection with several other cancers, including breast, colon, and prostate cancer. The computations of Riley *et al.* [5] yield estimates corresponding to the second test column of Table 2.

The present value of the attributable costs shows a somewhat different pattern, with the 15-year costs for distant-stage cases significantly exceeding those for local-stage cases. This can be explained as follows. First, even though survival is better among local than among distant cases, the attributable costs decrease with time from diagnosis, and therefore the effect of costs several years from diagnosis on the 15-year attributable costs is limited. In addition, as time since diagnosis increases, the amount by which costs are discounted increases, which further reduces the effect of the later costs on the 15-year costs.

It is interesting to observe that the attributable costs for regional cases exceed those for both local and distant cases. Although regional cases have poorer survival than local cases, their treatment costs are sufficiently high to offset this difference in survival. However, the treatment costs for distant cases are not sufficiently high relative to those for regional cases to offset the large difference in survival between these two groups. An explanation for this phenomenon has been offered by a referee, who notes that there is wide variation in the burden of disease within stages. Given that the general standard of care among gynecological oncologists is aggressive surgical debulking plus chemotherapy, it is possible that a patient with regional disease might receive more aggressive care than one with distant disease. This might explain the finding of higher costs for regional cases.

TABLE 1. Computation of control costs[a]

Age group (yr)	P	C_s (in U.S. dollars)	C_d (in U.S. dollars)	C_0 (in U.S. dollars)
65–69	0.02002	1,713	18,168	170.16
70–74	0.02002	2,172	18,571	208.32
75–79	0.04942	2,561	17,540	275.12
80–84	0.04942	2,828	15,110	286.28
85+	0.13728	2,901	12,014	346.03

[a] P is the annual death rate, from 1990 U.S. life tables (female only), given in 10-year age groups; C_s is the annual cost for survivors and is given by the 1988 annual survivors' cost from Lubitz and Riley; C_d is the 1988 annual cost for decedents. Costs from Lubitz and Riley are adjusted to 1990 dollars. C_0 is the monthly control cost.

4.4. *Limitations*

The analysis described in this section is subject to several limitations. We note these here, and, where appropriate, provide some suggestions for overcoming them. We hope that discussion of these issues will be useful to researchers dealing with similar or even with the same databases.

The first issue concerns the comparability of the survival and cost data and the interpretation of the estimate that combines data from the two sources. Naturally, we would like our estimate to be applicable to the U.S. population. Potosky *et al.* [15] note some slight differences between the population of SEER cases and the cases included in the linked SEER–Medicare database. In addition, there is the question of whether the SEER data are in fact representative of the U.S. population [18]. However, the SEER database and the linked SEER–Medicare database represent the most extensive and accurate sources of survival and cost data available to us today. We feel that combining data from these two sources should yield a more reliable and generaliz-

TABLE 2. KMSA estimates of the 15-year average total and atributable costs of care for epithelial ovarian cancer among women aged 65 years and above, by stage at diagnosis, with and without discounting at 5%/year

SEER stage	Total costs, no discounting (in U.S. dollars)	Attributable costs, no discounting (in U.S. dollars)	Total costs, present value (in U.S. dollars)	Attributable costs, present value (in U.S. dollars)
Local	51,405	23,212	42,164	21,285
Regional	46,191	34,942	41,668	32,958
Distant	38,858	33,507	36,655	32,126

able estimate of expected costs than that based on data from any single source.

In terms of actually using the linked database, a word about missing versus zero costs is in order. Strictly speaking, cases should only be included in the computation of monthly average costs for the months during which Medicare represents their primary health coverage. Costs are generally missing, i.e., not recorded, if other sources (spouse insurance, employer-provided insurance, VA) provide primary coverage. However, costs are also not recorded if no services are used in a given time period. Ideally, we should be able to identify whether costs are missing because no services were used, or because other sources provided primary coverage. The only other entitlement information besides Medicare available on the linked files is an indicator of Health Maintenance Organization (HMO) entitlement. Costs that are zero in patient-months covered by an HMO should be excluded, rather than set to zero in the estimation of monthly average costs. The analysis described in this section does not use the HMO entitlement information, and thus may slightly underestimate average monthly costs. The use of HMO entitlement information constitutes a refinement to be implemented in future analyses of these data, and is not expected to change the substantive conclusions.

Finally, we note that our analysis yields estimated attributable costs only for the age group 65 years and over. We do not feel comfortable extrapolating these estimates to patients below the age of 65 years since younger, healthier patients may be treated more aggressively with more extensive surgical procedures or more chemotherapy regimens. Estimates extrapolated from those based on Medicare data may therefore underestimate the costs for the younger age group. Attributable costs for younger patients may be estimated from alternate data sources, for instance, health maintenance organizations. However, these sources have their own limitations, such as representativeness issues, selectivity bias, and differing health care policies.

5. DISCUSSION

Accurate estimation of the costs attributable to a disease is a more challenging problem than is generally recognized. With the construction of a linked SEER–Medicare database [15], and the availability of other cost data sources such as health maintenance organization records, accurate estimation of attributable costs has become a real possibility. However, the very nature of the data raises statistical issues that must be dealt with during the estimation process.

The statistical issues essentially arise from the incompleteness of the available data; the need to incorporate discounting simply adds to the complexity of the problem. First, the survival data are incomplete—right censoring occurs during followup due to general loss to followup and to cases entering the case cohort at different times. Typically, the tail of the survival distribution is not identified due to long survival times and limited followup, particularly for early-stage cases. Second, cost data are generally collected over a fixed calendar period, so that different patients have cost data covering different periods in the disease process, forming what we have termed a synthetic cohort. The cost data are therefore subject to left and right censoring. Third, it is difficult to determine whether deaths are due to the disease or not. This can be important, for instance, in the method of Baker *et al.* [4], which requires that terminal care costs be estimated separately for cancer and noncancer deaths. Although SEER does record a cause of death variable, it should be interpreted with caution, because it is not always clear whether deaths were indirectly due to the disease or unrelated to the disease.

This article has proposed an estimator of attributable costs that takes the incompleteness present in the data into account, and can be adapted to yield a consistent estimate of the present value of the attributable costs. The KMSA estimator may be applied to censored survival data and synthetic cohort cost data, possibly from different sources. Costs from both censored and uncensored cases are used. Since the sample average component of the KMSA estimator is based on the average cost among patients surviving to the start of each time interval, it may be biased downwards if intervals are wide, due to censoring of costs during intervals. Under such circumstances, modifications to the sample average component may be considered. For instance, costs for individuals censored during an interval may be prorated to take the time of censoring into account. Alternatively, the sample average for each interval could be based only on individuals dying during the interval or surviving to the end of the interval. A sample average calculated in this way should be unbiased if the independent censoring conditions, enumerated in Section 2, hold.

Other estimators have been proposed in the literature. The estimator of Lin *et al.* [10] is similar to the KMSA estimator. In fact, it can be shown that the estimator of Lin *et al.* [10] and the KMSA estimator are identical in the absence of censoring. However, the estimator of Lin *et al.* [10] uses cost data only from cases with a death date recorded, and thus may not be efficient. The rationale for excluding censored cases is that this eliminates any bias that may arise from the association between the cost at censoring, and survival among censored cases. A detailed comparison of the two estimators is beyond the scope of this article, but their bias and efficiency are studied through simulation in [10] and [13].

A second method for estimating attributable costs is that of Baker *et al.* [4]. This method is quite different from either the pathwise or the KMSA estimators. The method divides the period from diagnosis until death into three phases: the initial phase, the maintenance phase, and the terminal phase. While this is a conceptually appealing formulation of the problem, motivated by the pattern of postdiagnosis costs, and used by many researcher and policy makers, care must be exercised when discounting costs with this method. We have shown that discounting the average maintenance cost leads to an overestimate of the average present value of the maintenance cost. Interestingly, it can be shown that the Baker *et al.* discounted death costs are an underestimate, so that the bias in discounting the death costs is in the opposite direction to that in discounting the maintenance costs. However, it is not clear that these biases will cancel each other out.

In terms of extrapolation beyond the period for which cost data and survival data have been recorded, our approach has been to avoid extrapolating, especially when discounting. A finite-period version of the KMSA estimator may readily be defined. If attributable costs are being considered, then the 10-year version of the KMSA estimator estimates the average attributable costs, assuming that these costs are negligible beyond 10 years postdiagnosis. Followup and cost data spanning a maximum period of 10 years postdiagnosis are necessary to estimate this quantity. In the ovarian cancer example, cost data were collected over seven calendar years, but this was sufficient to cover at least 10 years postdiagnosis, owing to the synthetic cohort nature of the data.

The KMSA estimator differs somewhat from previous approaches in the incorporation of terminal care costs. For instance, the method of Baker *et al.* [4] estimates terminal care costs separately from continuing care costs, distinguishing between deaths due to the disease and deaths unrelated to the disease. Terminal care costs are not an issue in the KMSA estimator, because the costs for terminal patients are included along with the costs for nonterminal patients in the average monthly cost among patients surviving to the start of the month. If a control group is available, the KMSA estimate of attributable costs implicitly distinguishes between cancer and noncancer deaths so long as terminal/death costs are included in the costs for the controls. In the analysis of the ovarian cancer data, a control group was not available for comparison, so the control cost, based on survivor Medicare costs, had to be adjusted to take deaths in this population into account.

Many questions, both statistical and nonstatistical, remain to be explored. On the statistical front, variance estimates and variance properties of the KMSA estimator are of interest. The relative bias and efficiency of the KMSA and pathwise estimators are key topics for further investigation [10], [13]. Evaluating the cost-effectiveness of preventive strategies is an important application of this work, and accurate estimation of the other components of cost-effectiveness is an important topic for further investigation. We hope that this contribution to the methodology for estimating attributable costs will be a useful step toward reliable cost-effectiveness estimates, and soundly based public health policy.

The authors thank Addy Tseng, Fran Maloy, and Ann Fowler for assistance with programming and data management; Arnie Potosky for helpful communication and, with Gerald Riley, for making the combined SEER-Medicare database available to us; Bill Barlow and Stephen Taplin for helpful discussions, and Eric Feuer, for sharing his work with Yochanan Wax with us which helped to shape this article. The authors are especially grateful to Larry Kessler for his in-depth reading of, and comments on, an earlier draft of this manuscript. This work was supported by Contract N01CN-05230 from the Division of Cancer Prevention and Control of the National Cancer Institute.

References

1. Hartunian NS, Smart CN, Thompson M. The incidence and economic costs of cancer, motor vehicle injuries, coronary heart disease and stroke: A comparative analysis. **Am J Public Health** 1980; 70: 1249–1260.
2. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. **New Engl J Med** 1977; 296: 716–721.
3. Russell LB. Is Prevention Better than Cure. The Brookings Institution, Washington D.C., 1986.
4. Baker MS *et al.* Estimating the treatment costs of breast and lung cancer. **Med Care** 1991; 29: 40–49.
5. Riley GF *et al.* Medicare payments from diagnosis to death for elderly cancer patients by stage at diagnosis. **Med Care** 1995; 33: 828–841.
6. Manning WG *et al.* The taxes of sin. Do smokers and drinkers pay their way? **J Am Med Assoc** 1989; 261: 1604–1609.
7. Keeler EB *et al.* The external costs of a sedentary lifestyle. **Am J Public Health** 1989; 79: 975–981.
8. Hodgson TA. Cigarette smoking and lifetime medical expenditures. **Milbank Q** 1992; 70: 81–125.
9. Langberg NA, Shaked M. On the identifiability of multivariate life distribution functions. **Ann Probability** 1982; 10:773–779.
10. Lin DY, Feuer EJ, Etzioni R, Wax Y. Estimating Medical Costs from Incomplete Follow-up Data: II. When the cost histories are not recorded. 1995. Technical Report #139, Department of Biostatistics, University of Washington.
11. Urban N. Cost effectiveness. In: Prevention of Coronary Heart Disease (Ockene IS *et al.*, eds.). Little, Brown and Company, Boston, Massachusetts, 1992.
12. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. **J Am Statist Assoc** 1958; 53:457–481.
13. Lin DY, Etzioni R. Estimating medical costs from incomplete follow-up data: I. When the cost histories are recorded. 1995. Technical Report #138, Department of Biostatistics, University of Washington.
14. Chung KL. A Course in Probability Theory. Academic Press, Orlando, Florida, 1974.
15. Potosky AL *et al.* Potential for cancer related health services research using a linked Medicare-Tumor registry data base. **Med Care** 1993; 31: 732–747.
16. Lubitz JD, Riley GF. Trends in Medicare payments in the last year of life. **New Engl J Med** 1993; 328: 1092–1096.
17. U.S. National Center for Health Statistics. Vital Statistics of the United States. 1992. US Department of Health and Human Services, Hyattsville, Maryland. DHHS publication No. (PHS) 93-1104.
18. Frey CM *et al.* Representativeness of the Surveillance, Epidemiology, and End Results Program data: Recent trends in cancer mortality rates. **J Natl Cancer Inst** 1992; 84: 872–877.