Volume Title: Inquiries in the Economics of Aging

Volume Author/Editor: David A. Wise, editor

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-90303-6

Volume URL: http://www.nber.org/books/wise98-2

Publication Date: January 1998

Chapter Title: Cause-Specific Mortality among Medicare Enrollees

Chapter Author: Jayanta Bhattacharya, Alan M. Garber, Thomas E. MaCurdy

Chapter URL: http://www.nber.org/chapters/c7090

Chapter pages in book: (p. 311 - 325)

10 Cause-Specific Mortality among Medicare Enrollees

Jayanta Bhattacharya, Alan M. Garber,
and Thomas E. MaCurdy

10.1 Introduction

Attempts to forecast health expenditures, to determine costs of specific illnesses, and to assess the long-term impact of programs designed to prevent or relieve specific diseases all require accurate estimates of mortality rates. Many such efforts build on information about the cause and timing of death for people who have certain diseases. However, the empirical basis for making accurate projections of cause-specific mortality, particularly for well-defined demographic and clinical subgroups, is often weak.

The standard life table framework offers a simple and powerful method for drawing inferences about the distribution of survival. Yet seldom have the data proved capable of supporting detailed studies of mortality by cause for well-defined populations. Standard U.S. life tables, based on birth records and death certificate data, with cause of death data, are published every several years by the National Center for Health Statistics (1991). Life tables compiled by age, race, and sex are published annually (National Center for Health Statistics 1994). Although these sources offer useful information about mortality trends by demographic group, they provide little information about the survival distribution pertinent to people with specific health conditions and risk profiles. Thus it is difficult to obtain, for example, a life table applicable to 70-year-

This research was supported in part by grants R29-AG07651 and P01-AG05842 from the National Institute on Aging. Jayanta Bhattacharya was supported in part by training grant T32-HS00028 from the Agency for Health Care Policy and Research.

Jayanta Bhattacharya is a graduate student in economics and a medical student at Stanford University. Alan M. Garber is a Senior Health Services Research and Development Senior Research Associate of the Department of Veterans Affairs; associate professor of medicine, economics, and health research and policy at Stanford University; and a research associate and director of the Health Care Program at the National Bureau of Economic Research. Thomas E. MaCurdy is professor of economics at Stanford University and a research associate of the National Bureau of Economic Research.

old men who are discharged from a hospital with a diagnosis of myocardial infarction. Small clinical studies and registries often provide information of this kind, but they usually are limited either by the selection criteria used to define the study population or by small sample sizes. They are not sufficiently comprehensive to cover a wide range of conditions, or to analyze a nationally representative sample.

In this paper, we describe the first steps toward developing such life tables. We lay out an approach to estimating survival patterns among the elderly that is based on longitudinal analysis of data from Medicare eligibility and claims files. These files offer a nationally representative sample of the elderly. Information about the cause of death, derived from hospital discharge files, allows us to link additional information about the terminal hospitalization and gives us the opportunity to obtain confirmatory data that are not routinely available from death certificate information. For our statistical modeling, we develop a flexible functional form to relate annual mortality rates to a set of individual characteristics.

The longitudinal analysis described below, which focuses on cause of death, can be a building block for studies that address a number of additional issues. For example, it can be extended to estimate future Medicare expenditures for the care of individuals who carry specific diagnoses (i.e., the longitudinal costs of incident cases of specific diseases). It can provide information about the expected pattern of expenditures for persons with a given set of characteristics, including not only age and gender, but also race, comorbidities, and prior hospital utilization. Similarly, such analyses can be used to identify populations who should be targeted for either preventive interventions or the identification and treatment of diseases. Finally, it can inform efforts to determine whether otherwise identical patients who receive different treatments have different outcomes.

10.2 Data Source

We obtained from the Health Care Financing Administration (HCFA) a 5 percent random sample of all Medicare enrollees, recorded in the Health Insurance Skeleton Eligibility Write-Off (HISKEW), for the years 1986–90 inclusive. This 5 percent sample consists of 1,518,108 people. The HISKEW file includes a unique identifier for each enrollee, in addition to basic demographic information such as age, sex, and race, and the date of death for each enrollee who died during the period of study. We also obtained the MEDPAR files, which contain information on every hospital admission during the study period, for every patient included in the 5 percent sample. MEDPAR includes dates of admission and discharge, discharge diagnoses, and discharge status, including whether the patient died in a hospital.

Using the unique identifier from the HISKEW file, we link each patient's

demographic information to a complete hospitalization record over the five-year period. This allows us to confirm the mortality information in the demographic file and to ascertain whether people who died during the study period died in a hospital. Furthermore, for those who died in a hospital, we are able to observe their primary discharge diagnosis, which we assign as the main cause of death.

Diagnoses are coded using the standard ICD-9-CM coding scheme. We perform two separate analyses of causes of death using broad and more specific diagnostic information. The analysis using the broad diagnostic classification, which employs the standard list of ICD-9 major diagnostic categories, permits a comprehensive picture of the main causes of death. There are 17 mutually exclusive ICD-9 code major categories ranging from code I, "infectious diseases," to code XVII, "injuries and poisonings." Codes that have a very small or zero sample size, reflecting the age composition of Medicare enrollees, are excluded from the analysis. In particular we exclude patients with the following causes of death: code XI, "complications of pregnancy, childbirth, and the puerperium"; code XIV, "congenital abnormalities"; and code XV, "conditions originating in the perinatal period." With these categories excluded, there remain 14 mutually exclusive broad causes of death. There are also two supplementary codes for special purpose categories, as described below.

The analysis using finer level diagnostic information allows us to determine the relative contributions of certain diseases, which are of broad policy interest, to total mortality rates. These categories include "heart attacks" (codes 410.XX and 411.XX, where XX denotes all subcodes), "strokes" (codes 430.XX through 438.XX), "congestive heart disease" (code 428.0), "lung cancer" (codes 162.XX), "breast cancer" (codes 174.XX), and "prostate cancer" (codes 233.4, 222.2, 236.5, 239.5, and 185.XX).

Most of the diagnostic labels in the broad scheme are self-explanatory. "E- and V-codes," however, are special purpose categories that supplement the standard diagnostic classifications. Essentially, V-codes apply to patients who are seeking care for a past diagnosis, such as a patient receiving chemotherapy for an already diagnosed cancer. E-codes allow the classification of environmental conditions which are the main cause of accidents or poisonings. The vast majority of Medicare patients classified in this category are admitted for V-codes, rather than E-codes.

We use information on all Medicare enrollees in the HISKEW file between the ages of 65 and 100 inclusive. We exclude enrollees younger than 65 years of age; they constitute a distinct population who are eligible because of a disability or because they require renal dialysis. For the analysis of cause of death, we use the sample of patients who died between 1986 and 1990 inclusive, whether or not they died in a hospital. For those patients who experienced a hospital stay and who died outside the hospital within one week of their discharge, we attribute the cause of death to the primary discharge diagnosis. For

all other patients who died outside the hospital, we designate the cause of death as unknown. Of the 1,518,108 people in the 1986 HISKEW file, 397,383 people died during the sample period.

10.3 Empirical Approach

To develop a statistical framework describing the incidence and causes of death, we separate the modeling tasks into two steps: the first introduces a distribution characterizing age-specific mortality in the elderly population; the second models health circumstances near the time of death.

10.3.1 Formulating a Model for Mortality Rates

A duration analysis provides a natural framework for characterizing age-specific survival probabilities. We describe here how such an analysis summarizes subsequent survival for an individual drawn from a population at age 65 of a given demographic makeup. A duration distribution describing the likelihood that an individual lives τ years beyond age 65 takes the form

$$f(\tau \mid X) = S(\tau - 1 \mid X)H(\tau, X), \tag{1}$$

$$S(\tau - 1 \mid X) = \prod_{t=1}^{\tau-1}[1 - H(t, X)]. \tag{2}$$

where the covariates X include factors other than duration that influence the lengths of survival times. The hazard rate $H(t, X)$ determines the fraction of the population who, having lived until age $65 + t - 1$, will die at age $65 + t$; the function $f(\tau \mid X)$ specifies the likelihood that an individual with attributes X will die exactly at age τ, and the quantity $S(\tau - 1 \mid X)$, the survivor function, depicts the probability that an individual will live until at least age $65 + \tau - 1$, given survival to age 65. The covariates X in the subsequent analysis include race and sex, the observed demographic characteristics. We break the sample into cells based on these observed characteristics and estimate separate survivor functions for each cell.

We estimate the hazard rate at age t, $H(t, X)$, by calculating the fraction of people in a given cell, alive at age t, who do not survive to $t + 1$. We subsequently calculate the survival distribution using equation (2).

10.3.2 Modeling Causes of Death

A second aspect of our empirical analysis characterizes the health conditions present at the time of death. For those who die, we designate one of 15 diagnoses as the cause of death, with a 16th category termed "other" for no diagnosis assigned at time of death (sometimes this is termed "natural causes"). Define

$$\Pr(\text{alive} \rightarrow i) \equiv \Pr(\text{alive} \rightarrow i \mid \tau, X) \qquad i = 1,\ldots,16, \tag{3}$$

as the probability that an individual who dies at τ and is a member of demographic group X has diagnosis i assigned as the cause of death. Formally, the quantity Pr (alive $\rightarrow i$) represents the probability that an individual dies from disease i conditional on dying at age τ and on the covariates X. There are 16 potential causes of death corresponding to the 16 diagnosis categories.

To offer a flexible specification, we parameterize these quantities using a multinomial logit specification, of the form

$$\Pr(\text{alive} \rightarrow i) = \frac{e^{g_i(\tau, X, \alpha_i)}}{\sum_{j=1,\ldots,16} e^{g_j(\tau, X, \alpha_j)}}, \qquad i = 1,\ldots,16, \tag{4}$$

where the function $g_i(\tau, X, \alpha_i)$ determines how the likelihood of various sources of death changes with age and α_i, $i = 1, \ldots, 16$, is a set of parameters to be estimated that determines the shape of g. We normalize the model by setting $g_{16}(\tau, X, \alpha_{16})$ to zero.

In equation (4), the function $g_i(\tau, X, \alpha_i)$ not only captures how the diagnoses and rate of death vary with age, but the presence of X in g_i also allows these relationships to differ across demographic groups. Spline models are an attractive approach for modeling duration effects, since they fit the data with a flexible and smooth function of duration. Implicit in conventional spline models, which fit polynomial functions to a series of intervals over duration, is a tradeoff between smoothness and goodness of fit. Fit can be improved by increasing the number of polynomial functions, but nondifferentiability at the boundaries requires a sacrifice in smoothness. Limiting the number of intervals or the order of the polynomial functions yields a smoother curve but diminishes the capabilities of detecting complicated forms of duration dependence.

To develop a flexible empirical specification for $g_i(t, X, \alpha_i)$, we apply a parameterization introduced in Garber and MaCurdy (1993) called overlap polynomials:

$$g_i(t, X, \alpha_i) = \sum_{j=1}^{J} \left[\Phi_{ij}(t) - \Phi_{i,j-1}(t)\right]\left[p_x(t, \alpha_{ij})\right]. \tag{5}$$

The quantity $\Phi_{ij}(t)$ denotes the cumulative distribution function (cdf) of a normal random variable possessing mean μ_{ij} and variance σ_{ij}^2, while $p_X(t, \alpha_{ij})$ is a polynomial in t, parameterized by α_{ij}. We estimate equation (4) separately for every race-sex cell and thus allow g_i to vary flexibly with demographic covariates.

The presence of the cdf's in equation (5) permits us to incorporate spline features in g_i so that the polynomial $p_X(t, \alpha_{ij})$ represents g_i over only a specified range of t. For example, suppose we wish to set $g_i = p_X(t, \alpha_{i1})$ for values of t between zero and t^* and to set $g_i = p_x(t, \alpha_{i2})$ for values of t between t^* and some upper bound $\bar{t}$. To create a specification of g that satisfies the property, assign $J = 2$ in equation (5), fix the three means determining the cdf's as $\mu_{i0} =$

0, $\mu_{i1} = t^*$, and $\mu_{i2} = \bar{t}$, and pick very small values for the three standard deviations σ_{i0}, σ_{i1}, and σ_{i2}. These choices for the μs and the σs imply that the quantity $\Phi_{i1}(t) - \Phi_{i0}(t)$ equals one over the range $(0, t^*)$ and is zero elsewhere, and the quantity $\Phi_{i2}(t) - \Phi_{i1}(t)$ equals one over the range $(t^*, \bar{t})$ and is zero elsewhere. Since the differences in the cdf's serve as weights for the polynomials, g_i possesses the desired property. Further, $g_i(t, X_2, \alpha_i)$ is nth-order differentiable in t for any value of n without imposing any continuity restrictions at the knot t^*, as one would have to specify if one were to use standard splines. With the values of the μ_{ij} and the σ_{ij} set in advance of estimation, $g_i(t, X_2, \alpha_i)$ is strictly linear in the parameters α and in known functions of t and X_2. We control where each spline or polynomial begins and ends by adjusting the values of the μs. We also control how quickly each spline cuts in and out by adjusting the values of the σs, with higher values providing for a more gradual and smoother transition from one polynomial to the next.

10.4 Empirical Results

10.4.1 Estimation Results for Survival Rates

As noted above, we estimate distinct survival models for each age-race cell in the sample. In particular, we estimate survivor functions for the entire population of Medicare-eligible elderly and for four demographic groups: white males, black males, white females, and black females. Our formulation for X allows for distinct hazard rates within each cell.

Table 10.1 presents survival estimates (percentage still alive at given ages) for individuals who have survived until age 65. These figures reflect well-known racial and gender differences in age-specific mortality rates; for example, blacks have higher mortality rates than whites, except at far-advanced ages, and mortality rates for men exceed those for women. The qualitative similarity with findings from other sources of demographic information help to validate the use of the long-term survivor functions that are estimated by taking advantage of the longitudinal aspect of this administrative data set that spans only five years.

In fact, the estimates that we obtain are quantitatively similar to those found

Table 10.1 Survival Rates for Elderly

	Percentage of 65-Year-Olds Living until at Least Age					
Demographic Group	70	75	80	85	90	95
Entire population	89.75	76.08	59.28	40.77	23.00	9.86
White men	86.05	68.90	49.15	30.47	15.40	6.54
White women	91.93	80.89	66.17	48.26	28.87	13.80
Black men	80.59	60.42	41.06	24.16	13.14	6.71
Black women	88.39	74.22	57.36	41.58	26.14	14.09

in standard life tables for all ages and demographic groups. For example, for white males, the standard 1987 life table, compiled by the National Center for Health Statistics (1988), reports that the probability of mortality within five years given for a person who reaches age 70 is 20.4 percent. Using the HCFA administrative database, we estimate the corresponding probability to be 20.0 percent. Similarly, the 1987 life table predicts that an unspecified individual 65 years old will have a five-year mortality of 10.6 percent while our estimate is 10.3 percent. For the demographic categories generally, our results closely match the life tables derived from the U.S. vital statistics system.

Such agreement is not surprising given that the date of death reported in the HCFA statistical files is likely to correlate well with death certificate data, from which life tables are constructed. While others have noted this correlation (Kestenbaum 1992), this is the first report of life tables calculated from these data.

10.4.2 Implications of the Findings for Survival

Accurate life tables are integral tools for health care policymakers interested in predicting the consequences of changing survival patterns. Construction of these tables, however, can be cumbersome. Death certificate information must be compiled, coded, and analyzed. The preceding results suggest that mortality rates, and the components of the life tables, can be estimated from the HCFA statistical database without requiring the use of death certificate data.

One limitation of life tables is that they are usually constructed only for a limited range of demographic subgroups, such as age categories by race and sex. This level of detail may be sufficient for many situations, but particularly when interest centers on the prognosis associated with certain diseases and treatments, more narrowly defined subgroups are needed. For example, one may be interested in the prognosis associated with the presence of a diagnosis of coronary heart disease. With our approach, it is easy to calculate the survivor function associated with such patients at a specific age.

In other words, the estimation of survivor functions for clinically important groups of people can serve as an effective tool in developing better information about prognosis. Even a well-conducted prospective observational study or a randomized clinical trial, the usual source of disease-specific prognostic information, may not provide comparable information. Randomized trials, for example, often lack generalizability: it is difficult to extrapolate from the results of the trial to infer results in classes of patients who were not included. By using a nationally representative sample, with no subgroup exclusions, our method avoids this pitfall.

10.4.3 Diagnosis Probabilities: Estimation Results and Implications

To estimate the probabilities Pr(alive $\rightarrow i$) defined by equation (4), we apply standard maximum likelihood procedures in a multinomial logit framework to compute values for the parameters α appearing in equation (4). We use the sample of 397,383 patients who died during the study period.

In this analysis we use a specification of $g(t,X_2,\alpha)$ that sets $J = 4$ in equation (5), with $\mu_0 = 0$, $\sigma_0 = 4$, $\mu_i = 70$, $\sigma_1 = 4$, $\mu_2 = 80$, $\sigma_2 = 4$, $\mu_3 = 90$, $\sigma_3 = 4$, $\mu_4 = 150$, and $\sigma_4 = 4$. Thus, the polynomial $p_X(t,\alpha_{i1})$ determines g_i from age 65 to age 70. After age 70, g_i switches to the polynomial $p_X(t,\alpha_{i2})$, which determines its value until approximately age 80, and so on. For every interval, we specify p_X to be quadratic in age. Given this specification, we estimate α by maximum likelihood.

We first perform the analysis using the 15 broad diagnostic categories mentioned in section 10.2, with the 16th category consisting of patients who were not assigned a cause of death because they died outside of the hospital more than a week after final discharge. Table 10.2 presents the smoothed probability of dying from condition i, given that death occurred in the age interval, for the entire population at various ages. For presentation purposes, we integrate the smoothed probabilities over age ranges in the table. Tables 10.3 and 10.4 present these same probabilities for white females, black females, white males, and black males. As table 10.2 shows, the most common causes of death are circulatory diseases (including myocardial infarction, congestive heart failure, stroke, and many other conditions), lung disease, and cancer. About three-

Table 10.2 **Diagnosis Associated with Deaths: Entire Population**

	Percentage of Deceased Who Died with Diagnosis in Age Group						
Diagnosis	65–69	70–74	75–79	80–84	85–89	90–94	95+
Infectious diseases	2.02	2.31	2.53	2.83	2.97	2.90	2.49
Neoplasms (cancer)	15.90	14.52	11.51	8.46	5.84	3.97	2.36
Immune and metabolic disease[a]	3.27	3.60	3.53	4.00	4.26	4.11	4.08
Blood diseases[b]	0.74	0.73	0.70	0.64	0.73	0.59	0.61
Mental disorders	0.71	0.71	0.78	0.84	0.73	0.54	0.45
Nervous system disease[c]	0.98	0.92	0.95	0.79	0.66	0.46	0.41
Circulatory diseases	20.98	23.13	24.05	24.15	22.73	19.53	15.28
Respiratory diseases	9.98	11.05	11.80	12.20	12.70	12.78	12.67
Digestive diseases	5.43	5.67	5.72	5.94	6.18	5.90	5.73
Genitourinary diseases	2.06	2.71	3.33	3.81	4.14	4.03	3.51
Skin diseases	0.58	0.70	0.88	1.04	1.25	1.32	1.35
Musculoskeletal disease[d]	0.86	0.87	0.76	0.75	0.66	0.53	0.55
Ill-defined conditions[e]	2.41	2.51	2.54	2.50	2.49	2.37	0.23
Injury and poisoning[f]	2.45	2.86	3.48	4.25	5.35	5.86	5.96
E & V codes	1.55	1.14	0.78	0.53	0.30	0.25	0.15
Other	30.09	26.57	26.64	27.26	29.00	34.86	42.12

[a]Category includes endocrine, nutritional, immune system, and metabolic disease.
[b]Category includes diseases of blood and blood-forming organs.
[c]Category includes diseases of the nervous system and sense organs.
[d]Category includes diseases of the skin and subcutaneous tissue.
[e]Category includes diseases of the musculoskeletal system and connective tissue.
[f]Category includes symptoms, signs, and ill-defined conditions.

Table 10.3 Diagnosis Associated with Deaths: Women by Race

Diagnosis	Percentage of Deceased Who Died with Diagnosis in Age Group						
	65–69	70–74	75–79	80–84	85–89	90–94	95+
			White Women				
Infectious diseases	2.30	2.53	2.58	2.74	2.86	2.68	2.10
Neoplasms (cancer)	17.90	15.37	11.21	7.43	4.90	3.46	2.05
Immune and metabolic diseases	3.94	3.76	3.73	3.83	4.03	3.89	3.60
Blood diseases	0.87	0.83	0.74	0.59	0.67	0.54	0.53
Mental disorders	0.83	0.70	0.75	0.79	0.65	0.48	0.35
Nervous system diseases	1.17	1.10	0.95	0.78	0.63	0.39	0.42
Circulatory diseases	21.41	23.46	25.05	25.03	23.38	19.64	15.10
Respiratory diseases	10.37	10.40	10.38	10.67	11.17	11.46	11.25
Digestive diseases	5.86	6.15	6.22	6.52	6.52	6.15	5.98
Genitourinary diseases	1.96	2.40	2.87	3.38	3.54	3.62	2.82
Skin diseases	0.64	0.75	0.97	1.06	1.23	1.25	1.27
Musculoskeletal diseases	1.13	1.14	0.88	0.88	0.70	0.55	0.45
Ill-defined conditions	2.56	2.40	2.62	2.39	2.30	2.29	1.94
Injury and poisoning	2.88	3.33	4.13	4.95	6.07	6.40	6.29
E & V codes	1.71	1.27	0.84	0.56	0.32	0.27	0.14
Other	24.48	24.41	26.07	28.40	31.05	36.92	45.72
			Black Women				
Infectious diseases	3.20	3.41	4.17	4.47	5.64	6.43	6.32
Neoplasms (cancer)	14.97	13.29	10.26	7.63	5.71	4.04	3.42
Immune and metabolic diseases	5.65	6.08	6.02	7.13	6.95	8.38	11.89
Blood diseases	0.76	0.71	0.74	0.86	0.94	0.73	0.43
Mental disorders	0.58	0.75	0.82	0.68	0.46	0.39	0.44
Nervous system diseases	0.90	0.99	0.97	0.71	1.09	1.24	0.57
Circulatory diseases	23.07	23.31	23.15	23.58	23.09	20.26	18.44
Respiratory diseases	6.30	7.40	7.61	7.61	10.23	10.94	11.99
Digestive diseases	4.75	4.95	5.31	5.79	6.41	5.96	4.87
Genitourinary diseases	2.92	3.86	4.41	5.69	6.87	6.42	4.37
Skin diseases	0.92	1.65	1.81	2.23	2.80	4.09	2.63
Musculoskeletal diseases	0.80	0.68	0.58	0.36	0.51	0.15	0.67
Ill-defined conditions	2.70	2.45	2.42	3.26	3.22	2.65	3.95
Injury and poisoning	2.42	2.09	2.39	2.44	3.12	2.78	3.69
E & V codes	1.57	1.12	0.81	0.33	0.15	0.21	4.68
Other	28.48	27.25	28.53	27.22	22.82	25.34	21.63

fourths of all deaths in this population occur during or soon after hospitalization. Note that black men are somewhat less likely to die of circulatory diseases than white men and, below the age of 85, are less likely to die in the hospital (table 10.4). Black women, on the other hand, are somewhat *more* likely to die in the hospital than white women, above the age of 79 (table 10.3).

By estimating the same model on the same set of patients with the finer

Table 10.4 Diagnosis Associated with Deaths: Men by Race

Diagnosis	Percentage of Deceased Who Died with Diagnosis in Age Group						
	65–69	70–74	75–79	80–84	85–89	90–94	95+
			White Men				
Infectious diseases	1.70	2.01	2.23	2.64	2.70	2.77	2.42
Neoplasms (cancer)	14.98	14.19	11.96	9.74	7.33	5.05	3.11
Immune and metabolic diseases	2.44	3.09	2.94	3.61	3.93	3.65	3.53
Blood diseases	0.64	0.67	0.69	0.68	0.86	0.75	0.84
Mental disorders	0.65	0.70	0.81	0.93	0.95	0.69	0.78
Nervous system diseases	0.87	0.80	0.93	0.85	0.69	0.52	0.41
Circulatory diseases	21.12	23.71	24.16	24.08	22.11	19.44	15.89
Respiratory diseases	10.11	11.86	13.60	14.58	15.77	16.37	16.81
Digestive diseases	5.32	5.46	5.49	5.43	5.60	5.29	5.22
Genitourinary diseases	1.94	2.74	3.61	4.21	4.80	4.40	5.35
Skin diseases	0.49	0.53	0.65	0.76	0.96	0.98	1.21
Musculoskeletal diseases	0.73	0.68	0.66	0.63	0.63	0.52	0.79
Ill-defined conditions	2.27	2.56	2.48	2.55	2.68	2.47	2.81
Injury and poisoning	2.24	2.64	3.08	3.78	4.59	5.34	5.88
E & V codes	1.50	1.14	0.77	0.56	0.33	0.25	0.20
Other	32.99	27.22	25.96	24.98	26.08	31.50	34.78
			Black Men				
Infectious diseases	2.60	2.94	3.79	4.64	4.38	3.96	5.60
Neoplasms (cancer)	13.90	13.46	11.97	10.68	9.23	6.37	3.92
Immune and metabolic diseases	4.23	4.89	5.13	6.97	7.41	7.17	7.02
Blood diseases	0.88	0.75	0.67	0.84	0.56	0.40	0.45
Mental disorders	0.74	0.93	0.64	0.87	0.80	1.29	0.01
Nervous system diseases	0.81	0.88	1.37	1.11	0.72	0.55	1.01
Circulatory diseases	17.44	19.22	18.91	18.88	19.25	16.47	13.88
Respiratory diseases	9.42	10.64	10.89	12.87	13.17	13.48	17.17
Digestive diseases	3.94	4.91	3.99	4.72	5.92	6.38	5.01
Genitourinary diseases	2.80	3.94	4.73	5.06	6.55	7.80	6.60
Skin diseases	0.52	0.95	1.11	1.85	2.11	2.00	3.85
Musculoskeletal diseases	0.34	0.51	0.38	0.50	0.35	0.35	0.62
Ill-defined conditions	2.55	2.79	2.65	2.63	3.22	2.71	3.35
Injury and poisoning	2.11	2.46	2.62	2.50	2.93	2.85	3.78
E & V codes	1.15	0.72	0.65	0.60	0.20	13.81	3.15
Other	36.57	29.99	30.49	25.28	23.19	14.42	24.60

diagnostic classification scheme, as described in section 10.2, we obtain the results shown in tables 10.5–10.7. In this scheme, the cause of death is classified by the principal diagnosis (hence these figures exclude individuals who had one of these conditions if the condition was only considered a contributory cause of death or an incidental diagnosis); the "other" category includes patients who were not assigned a diagnosis and patients who do not fall into any of the other diagnostic categories. Table 10.5 presents the smoothed probability

Table 10.5 **Selected Diagnosis Associated with Deaths: Entire Population**

	Percentage of Deceased Who Died with Diagnosis in Age Group						
Diagnosis	65–69	70–74	75–79	80–84	85–89	90–94	95+
Heart attack	5.75	6.36	6.25	4.65	3.98	2.62	1.37
Strokes	4.30	5.48	6.08	6.88	6.69	5.63	4.87
Congestive heart failure	4.51	5.50	6.35	7.26	6.82	5.91	5.53
Lung cancer	3.44	3.16	2.07	1.00	0.58	0.28	0.17
Breast cancer	0.23	0.20	0.24	0.11	0.14	0.19	0.10
Prostate cancer	0.43	0.57	0.58	0.51	0.40	0.42	0.04
Other	81.34	78.73	78.42	79.59	81.39	84.95	87.92

Table 10.6 **Selected Diagnosis Associated with Deaths: Women by Race**

	Percentage of Deceased Who Died with Diagnosis in Age Group						
Diagnosis	65–69	70–74	75–79	80–84	85–89	90–94	95+
			White Women				
Heart attack	5.77	6.22	6.21	5.46	4.22	2.88	1.64
Strokes	4.53	5.67	6.75	7.31	7.36	6.17	4.92
Congestive heart failure	4.51	5.31	6.28	6.98	7.02	6.84	5.74
Lung cancer	3.08	2.59	1.48	0.65	0.30	0.12	0.07
Breast cancer	0.69	0.50	0.37	0.26	0.18	0.18	0.10
Other	81.42	79.71	78.90	79.35	80.91	83.82	87.53
			Black Women				
Heart attack	3.99	3.85	3.32	2.90	2.86	1.73	1.86
Strokes	6.58	6.94	7.71	8.41	7.96	7.34	6.26
Congestive heart failure	5.06	5.69	5.32	6.14	6.11	6.09	4.33
Lung cancer	1.94	1.71	0.93	0.50	0.38	0.35	0.14
Breast cancer	0.59	0.48	0.57	0.17	0.16	0.12	0.33
Other	81.83	81.32	82.15	81.89	82.53	84.36	87.08

Table 10.7 **Selected Diagnosis Associated with Deaths: Men by Race**

	Percentage of Deceased Who Died with Diagnosis in Age Group						
Diagnosis	65–69	70–74	75–79	80–84	85–89	90–94	95+
			White Men				
Heart attack	5.82	6.08	6.06	5.41	4.57	3.23	1.91
Strokes	3.60	4.42	5.21	5.90	6.00	5.53	4.23
Congestive heart failure	4.72	5.68	6.23	6.71	6.72	6.36	6.58
Lung cancer	4.17	3.67	2.64	1.71	1.01	0.52	0.34
Prostate cancer	0.70	0.88	1.05	1.07	1.09	0.79	0.42
Other	80.99	79.27	78.81	79.19	80.60	83.56	86.53
			Black Men				
Heart attack	2.61	2.49	2.83	2.20	3.27	1.93	0.34
Strokes	5.35	6.27	5.64	6.76	5.97	5.52	5.53
Congestive heart failure	4.10	4.63	4.69	4.77	4.71	4.67	5.18
Lung cancer	3.62	3.02	2.99	1.86	1.05	0.67	0.02
Prostate cancer	1.26	1.68	1.67	1.62	1.90	2.21	1.10
Other	83.05	81.91	82.18	82.78	83.11	85.00	87.83

of dying from the conditions included in the scheme: acute myocardial infarction (heart attack), stroke, congestive heart failure, lung cancer, breast cancer, and prostate cancer. The first three diagnoses usually reflect diseases of the blood vessels; congestive heart failure is often a consequence of myocardial infarction, which is usually due to obstruction of the coronary arteries, and stroke is usually a consequence of obstruction in the arteries supplying blood to the brain or of blood clots that form in other arteries. These six diagnoses account for about 20 percent of all deaths among the elderly, according to these results. None shows a clear age trend except lung cancer, which accounts for a declining fraction of all deaths at greater ages.

Information on what diseases are most likely to be causes of death is clearly an important intermediate product in locating sources of cost growth in medical care. Since the overlap polynomial method allows flexible identification of those diseases that have the largest impact on mortality, one can cull detailed data—stratified by age, sex, and other clinical information—on the most important causes of death. Combining this information with cost data allows one to think about such questions as: Is a disproportionately large share of medical resources devoted to the oldest old, who may benefit little from the types of care they receive? Is there a cutoff age below which health care interventions are cost-effective? And what is the most appropriate method by which to determine priorities for the allocation of resources to research on the prevention and treatment of various diseases?

10.5 Conclusions

The Medicare claims files offer an important set of building blocks for studies that focus on mortality, health care utilization, expenditures, and health outcomes among the elderly. The preliminary work presented here demonstrates how the eligibility and hospital insurance claims files can be used to estimate survival curves by demographic group and by other characteristics, such as a diagnosis of one or more chronic diseases. The claims files further offer a basis for analyses of cause-specific death rates; although claims files are not considered as accurate as detailed audits of cause of death that are often performed as part of clinical research studies, they may well be more accurate than the death certificate data that usually serve as the major source of cause-of-death information in population-based studies. Furthermore, the longitudinal features of the claims files make it possible to explore supporting information for cause-of-death codes, by searching prior hospitalizations for discharge diagnoses of the cause of death and of related conditions (e.g., revealing a prior hospitalization for congestive heart failure or myocardial infarction in a person with a death diagnosis of cardiac arrhythmia, which is often associated with one of the other diagnoses).

Insofar as national data sources can be found to compare to the estimates of our models, the results are comparable. For example, our survival figures are

comparable to the life table figures supplied as part of the series of vital statistics of the United States. Our data on cause of death are largely consistent with results of studies that look at both causes of death and morbidity in the elderly (see, e.g., Johnson, Mullooly, and Greenlick 1990), although the national vital statistics system may be more successful in assigning a specific cause to each death, when compared with our procedures (Sutherland, Persky, and Brody 1990).

The data used here will soon be expanded in three ways: new data files will provide us with a larger percentage of enrollees, we will soon have more years of the files, and we will merge Part B (outpatient) data. The expanded data capabilities will enable us to estimate the survival models for a larger number of years per observation, and to pursue more finely detailed diagnostic categories (for analysis of both antecedent conditions and of cause of death).

With these data, we plan to attach Medicare expenditure and cost data for both inpatient and outpatient care, adapting our methods to estimate lifetime profiles of Medicare expenditures for individuals falling into various demographic and clinical categories. The framework can also be extended to analyze the mortality and utilization associated with use of specific procedures. We can compare, for example, the profile of expenditures and mortality for individuals with admissions for coronary heart disease who either do or do not undergo surgical treatment; the claims files offer a great deal of information about clinical characteristics of the patients, which when combined with geographic information has been used for instrumental variables analysis of the effects of alternative treatment strategies on health outcomes and costs.

Finally, this work can be extended to model the effects of preventive interventions on subsequent utilization and expenditures. For example, interventions that prevent or delay the development of prostate cancer will change the pattern of expenditures in ways that the longitudinal approach developed here can help predict.

References

Garber, Alan M., and Thomas E. MaCurdy. 1993. Nursing home discharges and exhaustion of Medicare benefits. *Journal of the American Statistical Association* 88:727–36.

Johnson, Richard E., John P. Mullooly, and Merwyn R. Greenlick. 1990. Morbidity and medical care utilization of old and very old persons. *Health Services Research* 25:639–65.

Kestenbaum, B. 1992. A description of the extreme aged population based on improved Medicare enrollment data. *Demography* 29:565–80.

National Center for Health Statistics. 1991. *Vital statistics of the United States, 1988.* Vol. 2, *Mortality, part A.* Washington, D.C.: Public Health Service.

———. 1994. *Vital statistics of the United States, 1990.* Vol. 2, Sec. 6, *Life tables.* Washington, D.C.: Public Health Service.

Sutherland, John E., Victoria W. Persky, and Jacob A. Brody. 1990. Proportionate mortality trends: 1950 through 1986. *Journal of the American Medical Association* 264:3178–84.

Comment Angus Deaton

This paper is the first report on an interesting and important research project. Using Medicare eligibility and claims files, it is possible to replicate life tables for the elderly, and more important, to extend them by conditioning on a richer set of covariates than just sex and race. As the authors emphasize in their introduction, such information could be useful for a wide range of economic and health purposes, from assessing the likely costs of demographic change and alternative delivery systems, to helping target medical innovations toward groups that can benefit the most. The current paper is a very preliminary one; it establishes that standard life tables can be more or less replicated—though it would have been useful to see the correspondence more fully and formally demonstrated—and it extends the life tables by breaking up death by various causes. This is certainly a useful first cut, though for the reasons I shall discuss, cause of death is perhaps not the most interesting or useful of the covariates that will be examined in the subsequent research.

The data consist of 397,383 Medicare enrollees aged 65 or over who died in calendar years 1986 through 1990. These deaths came from 1,518,000 people at risk, themselves a 5 percent sample of the population. The first set of calculations in the paper are of hazard rates by age, race, and sex, calculations that are obtained essentially by cross-tabulation, and that can be compared against standard mortality tables by age. The authors then disaggregate these mortality rates by cause of death. The decomposition by sex and race is retained, but some of the cause-specific age cells are now too small to support accurate estimation of the hazards, so the authors smooth by age using (nonstandard) splines. As the project advances into more complex calculations, the spline technology is likely to be more useful than in this paper, where a more straightforward alternative would have been to increase the sample size from the 5 percent used here. I am also a little skeptical that this particular spline technology is the most appropriate for the task. Since the estimation is done by generalized logit, a simpler—and presumably more efficient—solution would have been to use locally weighted logits, so that the calculation, if not fully disaggregated by age, would use mortality information from neighboring ages, with information weighted more heavily the more relevant it is.

Turning to the substance, there is surely cause for concern in pooling the

Angus Deaton is the William Church Osborn Professor of Public Affairs and professor of economics and international affairs at Princeton University and a research associate of the National Bureau of Economic Research.

information for the five years and all ages, with no allowance made for possible cohort effects. Although the pooling is perhaps natural over a five-year period, there is no reason to suppose that mortality or cause-specific mortality for a 70-year-old black male should be the same in 1990 as it was (say) in 1970. Patterns and standards of living change, as does the health delivery system, and there is no reason to suppose that there will not be time or cohort effects in mortality, as well as the age effects that are modeled here. The separate identification of age, year, and cohort effects is not straightforward, but it is surely an issue that will have to be faced in this work before it can be used with any confidence to forecast health expenditure patterns or mortality rates.

The other substantive issue that is insufficiently acknowledged in the current paper is measurement error. Even when recording is perfect, there are difficult conceptual issues in classifying causes of death. A great deal more attention is paid to cause of death for young patients than for older ones, the immediate cause of death is often not the same as the fundamental cause of death, and some diagnoses are wide enough to be little more than a confirmation that the patient is dead. It is also unfortunate that so many of the deaths in the analysis are diagnosed as "other" or "natural causes," especially when this is one of the few categories where there are systematic patterns with age. While I think that cause of death poses the greatest difficulty for this paper, it is not the only variable for which there are problems. Age reporting is notoriously inaccurate for the very old, and it would have been useful if the paper had been clearer about the source of the age information used in the analysis; for example, self-reported age at hospital admission may be different from that in the social security records.

Of course, measurement error is no reason to abandon the data or the exercise. But the specific difficulties over cause of death seem to require a good deal of further investigation. In particular, I would welcome some demonstration that these reported diagnoses are at all useful for any of the original aims of the paper. For example, is there enough information in the data to allow any link with costs? Do age/sex/race decompositions of cause of death tell us anything about which groups to target in future health care reforms? Is it reasonable to suppose that cause-of-death diagnosis will be invariant under changes in the health delivery system? The paper would have been strengthened by the discussion of any of these important issues.