Technical Documentation

A Microsimulation Model to Forecast Disease, Disability and Expenditures by the Elderly

CHAPTER 1. INTRODUCTION

This technical report documents the details of a model projecting health expenditures, disease, and disability among future generations of the elderly. At the core of the model development project is the development of a demographic-economic model to project future health care expenditures. The first goal is to answer the question: *If current health status and disability trends continue, what will be the costs to Medicare for treating the elderly, and how will they affect health and functional status?* The second goal is to serve as the simulation vehicle for evaluating "what if" scenarios about the future health care environment.

The model diverges from traditional approaches in that it includes a multi-dimensional characterization of health status. In addition, conventional actuarial approaches employ cell-based models in which each cell represents a subpopulation of interest. While it is theoretically possible to extend cell-based models to support health care projections, practical shortcomings make it difficult to simulate changes of the sort identified in Chapter 2. The desirability of a rich characterization of health status, by sex and age group, implies that the number of cells would need to be very large. Cell sizes would be correspondingly small, and the very large Markovian transition probability matrix difficult to estimate. Microsimulation models offer a conceptually and analytically superior alternative.

THE MECHANICS OF THE MICROSIMULATION

Microsimulation models start out with as large a sample of individuals as possible. The sample needs to contain information on all health status measures that are strong predictors of health expenditures. For expositional purposes, suppose health measures A, B, and C are relevant. In our preliminary specification, these measures reflect ADLs, clinical diagnoses (cancer, diabetes) or perhaps states such as "institutionalized in a nursing home." One may well both suffer from diabetes and be institutionalized, i.e., the states are not mutually exclusive. The measures may or may not be "absorbing," i.e., one may recover from a subset of health statuses. Denote with H the "healthy" state in which the person is free from A, B, and C, and with D the "deceased" state. Individuals may then be H; A; B; C; A+B; A+C; B+C; A+B+C; or D.

At the time the sample was drawn, we know individuals' health status. The goal is to map out individuals' remaining life paths and identify at what point(s) in time they transition into other health statuses and when they are likely to become deceased. This requires that we estimate transition models into all possible health states. In the example, we need at least four models: transition into A; transition into B; transition into C; transition into D (deceased), plus potentially additional recovery models. We don't need to distinguish, say, transitions H→B from A→A+B; the fact that an individual suffers from A may be treated like any other covariate, so that the models are conditional on existing health status.

The first step is to estimate individual health transition models. Several types of models may be chosen, depending on the richness of available (longitudinal) data. For example, a simple logit or probit transition model may be estimated if information is available on health status at two points in time. With more than two health status observations per individual, such simple models may account for health history; with yet more detailed information, continuous-time hazard models may be estimated. Transition models may be estimated using any data source that contains health measures that are identical to those distinguished in the microsimulation sample. It is, of course, preferable to estimate transition models directly off the microsimulation sample, so that the definition of health outcomes is exactly right.

The second step is to project future health transitions. Regardless of the estimated model type, we can compute interval (discrete) transition probabilities conditional on a rich set of demographics,

current health status, and (if available) health status history. These transition probabilities are used to forecast health transitions. If the probabilities only account for current information, a first-order Markovian process is generated; if they account for lagged covariates, such as accumulated health histories, higher-order Markovian processes result. Note that the probabilities depend on potentially many individual-specific characteristics and initial state, unlike the generic transition probabilities in cell-based models which apply to cells consisting of a fairly heterogeneous subpopulation.

By illustration, consider an individual who at baseline suffers from health condition A. The model computes the following four transition probabilities:

1) Probability of recovering (transition into state H) in the next year (say, p_h=.002);
2) Probability of attracting health condition B (transition A→A+B) in the next year (say, p_h=.06);
3) Probability of attracting health condition C (transition A→A+C) in the next year (say, p_h=.05);
4) Probability of dying (transition A→D) in the next year (say, p_h=.08);

We draw a random number between zero and one from a uniform distribution to simulate a health shock. If the transition probability exceeds the corresponding random draw, we project that the transition took place. It may well be that all four random draws are larger than the transition probabilities. In that case, the person remains in state A throughout the year. It may also be that multiple transitions are projected to take place. In the example, transitions into both B and C may be possible, so that the person ends up with multiple health conditions, A+B+C. The transition into death logically dominates all others. If multiple transitions are conceptually implausible or impossible, the transition interval may be shortened (from a year to perhaps just a week or a day), so that multiple transitions are ruled out[1].

Continuing the example, the model projects that the individual will remain in state A throughout the first year. Transition probabilities for the next year change, because the individual is one year older, and perhaps because there are time trends in the transition models. We then draw new random numbers. If he remains in state A for four additional periods, until in the sixth period, he is projected to attract illness B, so his new state is A+B. Then the set of potential next transitions changes. Further, the transition probabilities have changed not just because of age and time, but also because of a change in health condition. For example, his health has now deteriorated severely so that his mortality risk is much higher than before. We compute new transition probabilities and compare them with randomly drawn numbers. The result is a simulated life path in which the person accumulates multiple disease conditions, and then dies.

CHOICE OF THE HOST DATA SET

The microsimulation sample needs to be a large data base with information on many personal characteristics: sex, date of birth, health conditions, income, supplemental health insurance status, and as many other covariates as possible. These requirements point to large-scale survey data. This data base is the "host" survey.

After consultation with the social science expert panel, we chose to use the Medicare Current Beneficiary Survey (MCBS). The MCBS is a nationally representative data set designed to ascertain utilization and expenditures for the Medicare population, especially those expenditures born by the beneficiary or supplemental insurance. The sample frame consists of aged and disabled beneficiaries enrolled in Medicare Part A and/or Part B although we use only the aged. The MCBS attempts to interview each person twelve times over three years, regardless of whether he or she resides in the community, a facility, or transitions between community and facility settings. The disabled (under 65 years of age) and the oldest-old (85 years of age or over) are oversampled. The first round of interviewing was conducted in 1991. Originally, the survey was a longitudinal sample with periodic supplements and indefinite periods of participation. In 1996, the MCBS switched to a rotating panel

design with limited periods of participation. Each fall a new panel is introduced, with a target sample size of 12,000 respondents, and each summer a panel is retired. The MCBS contains detailed self-reported information, including the prevalence of various conditions; measures of physical limitation in performing daily activities (ADLs) and instrumental activities of daily living (IADLs); and height and weight. In addition, the MCBS contains very detailed self-reported data on health service use, as well as Medicare service use records. Institutionalized respondents are interviewed by proxy. To increase sample size, we pool multiple rotation groups. Table 1 shows the sample size for MCBS in each year after dropping observations with missing data (mostly due to missing self-reported conditions).

Table 1. Sample Size For the MCBS Analytic File, 1992 to 1998

Year	N	Percent
1992	10,584	14.6%
1993	10,188	14.1%
1994	10,557	14.6%
1995	9,974	13.8%
1996	9,866	13.6%
1997	10,426	14.4%
1998	10,881	15.0%
Total	72,476	100.0%

For our simulations, we select all individuals age 65 and older in MCBS 1998 dataset. This leaves 10,881 individuals. (For some simulations, we restrict attention to a cohort of 70 year olds whom we follow through time. In these cases, we use data from the pooled sample 1992 to 1998 to ensure adequate sample size). Original MCBS cross-sectional weights indicate the number of persons in the population that every sample member represents. The weights range from 1106 to 12,131 due to stratified survey sampling and non-response rates. We re-scale the weights such that they add up to the 1998 population of individuals aged 65 and older (representing 34,385,239 individuals). A simulation with this host data set of 10,881 individuals would generate unbiased projections. However, the sample size for rare subpopulations (as measured by their multi-dimensional health status) is limited. We therefore replicate observations in the sample. This allows for multiple health status paths per sample member and yields more precise (smoother) estimates of future health status distributions. We replicate in accordance with individuals' relative weight in the sample; the minimum number of replications is two, the maximum 55. The average replication is 10 times, so that the resulting host data consists of 108,810 individuals. Their weights are now more uniform and range from 276 to 355.

DEFINING HEALTH STATES

We define health states based on self-reported health conditions and disability. The MCBS asks about a multiplicity of health conditions. For the preliminary model, we chose to focus our analysis on diseases being investigated by our medical panels. Because of the way these diseases were chosen, these conditions are the ones that are most prevalent in the elderly population and also the most expensive to treat. The conditions we use are shown in Table 2 along with their prevalence in the MCBS. For comparability with other studies, these rates exclude individuals residing in a facility at any point during the year.

Table 2. Prevalence of Select Conditions, MCBS Non-Institutionalized Population

Condition	Prevalence (%):		
	65+	**65-69**	**70+**
Cancer	17.7	14.3	19.0
Breast[1]	6.5	6.8	6.4
Prostate[2]	6.6	4.3	7.5
Uterus[1]	2.9	2.4	3.0
Colon	2.5	1.6	2.9
Bladder	0.9	0.3	1.1
Lung	1.0	0.9	1.1
Kidney	0.3	0.3	0.3
Throat	0.5	0.2	0.7
Head	0.2	0.1	0.3
Brain	0.1	0.1	0.1
Other	3.1	2.7	3.3
Heart Disease	38.2	29.5	41.4
Angina pectoris/CHD	14.4	11.3	15.5
Myocardial infarction	14.7	12.4	15.6
Other	27.6	20.4	30.4
Alzheimer's	2.4	0.7	3.0
Stroke	10.4	7.4	11.5
Diabetes	16.0	15.2	16.3
Hypertension	55.8	49.5	58.1
Lung	14.2	13.7	14.4
Arthritis	57.3	48.5	60.6
BMI[3]	26.0	27.1	25.5
Ever Smoke	60.3	64.4	58.8
Disability			
ADL≥1	25.8	16.1	29.4
ADL≥3	8.4	3.7	10.2

Note: Results from 1998 survey sample and exclude nursing home residents. Responses are weighted using MCBS 1998 cross-sectional weights.
[1] Universe includes women only.
[2] Universe includes men only.
[3] Measured as kg/m^2, not as percentages.

As a consistency check, we compared several of these rates with data from the 1994 and 1995 National Health Interview Surveys. The NHIS serves as the data source for the under 65 population who will age into Medicare in the microsimulation. The result of this comparison is shown in Table 3.

Table 3. Comparison of Condition Prevalence between the MCBS and NHIS

	MCBS Prevalence by Age (%)			NHIS Prevalence by Age (%)		
Condition	**65+**	**65-69**	**70+**	**65+**	**65-69**	**70+**
Cancer	19.3	15.9	20.7			
Breast[1]	6.6	6.2	6.7	2.6	1.5	3.1
Prostate[2]	5.8	4.4	6.4	4.5	2.6	5.5
Uterus[1]	3.1	2.9	3.1	0.2	0.2	0.2
Colon	2.3	1.2	2.8	0.6	0.4	0.7
Lung	0.8	0.8	0.8	0.4	0.1	0.5
Heart Disease	38.3	30.2	41.7	27.5	21.5	30.2
Hypertension	54.4	47.9	57.1	36.4	30.8	38.9
Diabetes	17.2	16.0	17.6	10.1	8.7	10.8
Disability						
ADL≥1	27.2	17.1	39.4	9.6	4.5	11.9
ADL≥3	9.5	5.0	11.5	4.1	2.0	5.1

Notes: NHIS prevalence rates are from the 1994 survey, except for disability, which comes from the 1995 Disability Phase I supplement. Tabulations are based on the recodes provided by NHIS (Diagnostic Recode C). The NHIS asks about stomach, intestine, colon, and rectal cancer in one question, the response to which is reported as "colon cancer" in the table; the list of cancer types asked by the MCBS is shown in Table 2. MCBS data are from 1995.

[1] Universe includes women only.

[2] Universe includes men only.

Clearly there are some large differences between the two sets of prevalence estimates. Some of the difference can be explained by question wording. The MCBS asks about all conditions in the form "Has a doctor ever told you had [condition]?" However, the NHIS varies its wording depending on the condition.[1] For diabetes, and the cancers listed above, the questions are of the form "During the past 12 months, did anyone in the family have [condition]?" For cardiovascular disease and hypertension, the NHIS asks "Has anyone in the family ever had…?", except for tachycardia and heart murmurs which were asked in the form "During the past 12 months, did anyone in the family have…?"

This wording difference means the rates of cancer should be much lower in the NHIS, since cancer survivors are much less likely to report having disease in the NHIS than the MCBS. For example, if a woman had an early stage, non-metastatic tumor removed from her breast 10 years ago, she will not report this cancer in the NHIS but she would in the MCBS. On the other hand, the NHIS has much lower rates of cardiovascular disease[2], hypertension, and diabetes that cannot be explained by differences in question wording.

Disability in the MCBS is defined as having any difficulty with or inability to perform bathing *or showering*, dressing, eating, getting in and out of bed or chairs, *walking*, and using the toilet. In the NHIS supplement, disability is defined as having any difficulty with or inability to perform bathing, dressing,

[1]The NHIS does not ask each respondent all conditions. Instead, the family is randomly assigned to one of six condition lists: skin and musculoskeletal conditions; impairments; selected digestive conditions; selected conditions of the genitourinary, nervous, endocrine, metabolic, and blood forming systems; selected circulatory conditions; or selected respiratory conditions. Since the list of cancers crosses condition lists, we cannot calculate an overall prevalence rate for any cancer.

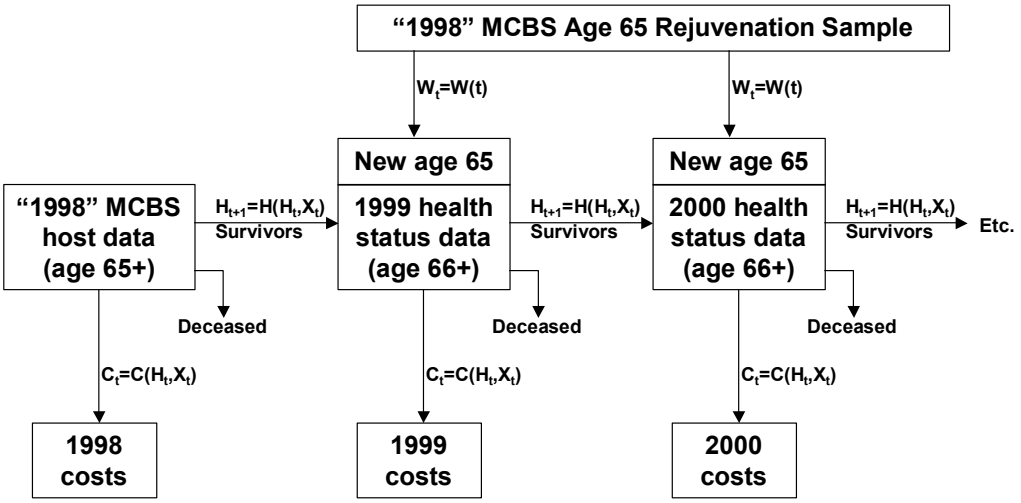
[2]Cardiovascular disease includes the following recodes from the NHIS: rheumatic fever with or without heart disease (501); ischemic heart disease (502); heart rhythm disorders including tachycardia or rapid heart (503), heart murmurs (504), other and unspecified heart rhythm disorders (505); congenital heart disease (506); other selected diseases of heart (excludes hypertension) (507); and hardening of the arteries (510).

eating, getting in and out of bed or chairs, *getting around inside the home*, and using the toilet. Because the MCBS asks about walking—which results in higher rates of difficulty than getting around inside the home—and showering, it should have higher rates of disability. These wording differences explain much—but not all—of the difference in disability rates, since the differences persist even when looking at similarly-worded ADLs (data not shown).

MODEL OVERVIEW

Figure 1 depicts how the cost models, transition models, and rejuvenation models are integrated into our microsimulation model. The model is designed to yield predictions in constant dollars and—at baseline—using 1990's 'technology.' We start with MCBS data for 1998 as the host cohort. The characteristics of these individuals are used to predict per capita 1998 medical expenditures. The weights of the host data are adjusted such that they add up to the 1998 population of individuals age 65 and older. The product of per capita expenditures and population size yields aggregate 1998 medical expenditures. The host data include some individuals who, at the time of their last interview of their first year in the MCBS had become deceased. These are dropped from the sample. We then project individuals' health status in 1999. By then, the sample has aged to 66 years of age and older. We rejuvenate the sample using a rejuvenation sample which consists of age-65 MCBS respondents. The weights of newly entering individuals are adjusted, first, in accordance with 1999 prevalence rates of health conditions among 65-year olds, and second, such that the sum of weights newly entering 65 years old equals the 1999 population of individuals age 65. The resulting sample is representative of the 1999 age 65+ population. We use the health status and demographic characteristics of this sample to predict per capita 1999 medical expenditures and derive the 1999 aggregate expenditures. We then drop individuals who have become deceased, project the health status of survivors in 2000, rejuvenate the sample, and compute 2000 costs; et cetera, through the year 2030. More detail is provided in this and subsequent chapters.

Figure 1. Overview of the FEM



Sample Rejuvenation

As our initial host sample ages, it is no longer representative of the age 65+ population. We therefore rejuvenate the sample annually with a newly entering cohort of 65-year olds. These individuals consist of 65-year olds in the 1992-1998 MCBS; each individual enters only once, with his or her characteristics measured as of the first year of the MCBS in which he or she was interviewed.

There are 2,863 respondents age 65 in the 1992-1998 MCBS. We conducted a separate analysis of the "diversity" of these 2,863 individuals, distinguishing all possible combinations of cancer, heart disease, neurological disorder, hypertension, diabetes, and disability (0 vs 1+ vs 3+ ADLs). The number of theoretically possible combinations is 2*2*2*2*2*3 = 96. The 10881 first-year MCBS respondents of all ages represent 95 health status combinations; the number of combinations among 2,863 respondents age 65 is 89. In other words, there are 7 health status combinations missing among the 65-year olds. Naturally, as individuals age, they may attract more health conditions and move into new health condition combinations.

Components of the Model

Subsequent chapters describe the three models that form the components of our microsimulation model: health care costs, health status transitions; and characteristics of future newly entering Medicare enrollees. Chapter 2 describes the cost estimation using data from the MCBS. We consider two outcomes: total Medicare payments and from any source. The explanatory covariates include self-reported health status, interactions of health status with disability measures (to capture severity of the condition), residency in a (nursing home) facility, and demographic characteristics. The product of these cost models are functional relationships that predict medical expenditures; we denote these relationships by $C_t=C(H_t,X_t)$.[3] In so doing, we make several assumptions.

1. We assume that future individuals with a given set of health conditions receive the same medical care as individuals in the MCBS. This is tantamount to saying that our baseline case corresponds to 1990's "technology."
2. We assume that 1998 unit prices continue throughout our forecast period. This (obviously unrealistic) assumption implies that our results are in 1998 dollars. The applicable price index is the price index for medical services, not the standard consumer price index.
3. Cost regressions are based on non-HMO Medicare enrollees, so our per capita projections apply to the non-HMO population only.
4. We assume that the elderly do not migrate across Census region borders (North-East, Midwest, West, South, other) as they age. We also assume that elderly that live in urban areas continue to do so, and that those in rural areas do not move to an urban area.
5. We assume that there are no changes in the age patterns of omitted and potentially time-varying covariates, such as marital status and private retiree health insurance coverage.

We also convert per capita medical expenditures into population aggregates using elderly population estimates from the Census Bureau. This requires several more assumptions:

[3] For flow variables, such as in annual costs, C_t, subscript t denotes a calendar year; for stock variables, such as health status, H_t, it denotes the year of interview (typically administered in the fall).

6. Medical costs of HMO enrollees and the non-HMO elderly are the same.
7. We assume that all elderly are covered by Medicare Parts A and B. This implies a slight overestimate of projected aggregate HI costs and an overestimate of roughly 3 percent of projected aggregate SMI costs.
8. The population forecasts do not distinguish race or Hispanic ancestry, so we assume that the fractions African Americans and Hispanics remain constant.

Chapter 3 develops models of health transitions. It currently only uses data from the MCBS. We project transitions of self-reported cancer, heart disease, neurological disorder, hypertension, diabetes, and disability. Mortality is calibrated to national figures using Vital Statistics, thereby allowing a global time trend in life expectancy. Finally, we project transitions into facilities, such as nursing homes. We assume that residency in a facility is an absorbing state. The explanatory covariates include health status and demographic characteristics as measured in the previous year. The product of these transitions models are functional relationships that predict health status one year into the future; we denote these relationships by $H_{t+1}=H(H_t,X_t)$. Because these states are measured by questions as "Did a doctor ever tell you...." we treat them as absorbing. We also project future disability status (number of ADLs), which may improve or deteriorate with age. Finally, we project entry into facilities such as nursing homes. We assume that residence in a facility is an absorbing state.

Chapter 4 describes how we estimate prevalence in future years—i.e., how we forecast the health status of new entrants into Medicare at age 65. It uses data from several years of the NHIS, and exploits prevalence and incidence rates of individuals as young as 30 years. It projects joint prevalence rates of cancer, heart disease, neurological disorder, hypertension, diabetes, and disability status among 65-year olds through the year 2030. In addition, it takes account of co-morbidity patterns of newly entering Medicare enrollees in the MCBS and forces MCBS prevalence correlations to continue in its forecasts. It then rescales projected joint prevalence rates into weight adjustment factors, which are used for annual rejuvenation of the sample with newly entering Medicare enrollees. The product of these trend models are relative weights for each health condition combination for 65-year olds in 1995 through 2030; we denote these relative weights by $W_t=W(t)$. Before rejuvenating the simulation sample with newly entering 65-year olds, we adjust their weights in accordance with projected joint prevalence levels. We then apply a second adjustment to the weights of newly entering individuals to ensure that the total population of individuals age 65 and older matches projections from the Census Bureau. Finally, to boost sample size, we replicate newly entering individuals and adjust their weights accordingly.

CHAPTER 2. ESTIMATING COSTS

A major determinant of health care expenditures among elderly Americans is the prevalence of chronic disease and disability. While not all of these conditions lead to persistently high medical costs, the presence of a stroke, cancer, and many other conditions can have a lasting impact on health status, disability and the demand for medical services.

We use longitudinal data from the Medicare Current Beneficiary Survey (MCBS) Cost and Use files, as described in Chapter 2. Reimbursements in the MCBS are categorized into nine different service groups, such as inpatient care, ambulatory services, outpatient prescription drugs, home health, and institutional care. This level of cost detail allows us to explore how new therapies and technologies affect treatment and outcomes and how the mix of services change over time and across patient subgroups.

The cost analyses exclude enrollees under age 65, persons enrolled in HMOs, and those without Part B Supplemental Medicare Insurance due to incomplete ascertainment of utilization. Because of these exclusions, the sample sizes for these analyses will be smaller than those shown in the previous chapter. The average yearly sample consists of approximately 8,400 beneficiaries.

The annual number of enrollees and average Medicare reimbursements over the 7-year period are reported in Table 2.1. Average Medicare expenditures increased nearly 11.5 percent in real terms between 1992 and 1998, reflecting possibly increased per capita utilization. The number of enrollees in our sample declined over time, primarily due to increased HMO enrollment and greater numbers of younger beneficiaries who were excluded from the analyses.

Table 2.1. Sample Size and Medicare Reimbursement, by Year

		Medicare Reimbursement	
MCBS Year	N	Mean	Std Dev.
1992	9,406	$4,441	$11,303
1993	8,966	4,501	11,790
1994	9,212	5,021	13,208
1995	8,469	5,160	13,322
1996	8,073	5,315	13,432
1997	8,200	5,416	13,339
1998	8,325	4,953	11,747
Total	60,651	4,960	12,614

Source: 1992-1998 MCBS.

Because we are interested in forecasting future Medicare outlays, the primary cost measures used in the analyses are total Medicare reimbursements and their major components. CMS calculates and projects allowed charges or costs for Medicare covered services and subtracts the deductibles and coinsurance owed by the beneficiary. Part A reimbursements cover inpatient hospital services, up to 100 days of post-hospital skilled nursing facility (SNF) care, home health services and hospice care. Part B provides coverage for physician services, outpatient hospital services, durable medical equipment, and other medical and ancillary services. Secondary analyses examine out-of-pocket expenses, Medicaid reimbursements, and medical spending by other third-party payers.

DISABILITY, HEALTH STATUS, AND DISEASE

We first examined how alternative measures of health and disability affect expenditures, both independently and interactively.

Disability. Past efforts to model the effects of medical interventions on utilization and costs typically include various measures of physical health such as functional limitations, disability, or the presence of chronic diseases. Two measures of physical functioning common in survey data are *functional limitations* and *activities*. Functional limitations generally reflect an inability to carry out physical tasks such as bending or lifting without help or aids. Alternatively, activities of daily living (ADLs) are more closely tied to social roles, particularly those deemed necessary to meet an individual's personal needs, e.g. eating, bathing, and dressing. A related concept, instrumental activities of daily living (IADLs), are more complex activities, such as managing money and shopping for groceries.

The MCBS asks respondents if they have any difficulty performing each of six daily activities because of health or physical problems. The fraction of the sample reporting difficulty with each activity is reported in Table 2.2. Nearly one in five older beneficiaries reports difficulty bathing or getting out of bed or a chair; 6 percent have troubling eating; and almost a third report difficulty walking.

Table 2.2. Frequency of Activity Limitations

Condition	Percent of Sample Reporting Difficulty
Bathing	16.9
Dressing	11.9
Eating	5.2
Getting Out of Bed/Chair	16.9
Using the Toilet	9.4
Walking	27.2

Notes: Analyses come from the 1992-1998 MCBS. All calculations are weighted using normalized annual cross-sectional weights—i.e., the weights for each year sum to one. All costs are reported in 1998 dollars and are inflated using the medical CPI.

In aggregate, over 40 percent of older beneficiaries report one or more ADLs, which are highly correlated with Medicare reimbursements. Beneficiaries age 65 and older who experience difficulties walking, dressing, or getting out of bed have substantially higher medical expenditures than those without limitations (Table 2.3). For example, persons reporting five or more ADLs incur nearly $17,000 in annual Medicare expenses compared to under $2900 for seniors without limitations.

Table 2.3. Average Medicare Reimbursement by ADL Counts

ADL Counts	N (Unweighted)	% of Sample	Mean $	Median $
0	36,469	60.1	$2,875	$451
1	7,242	11.9	$5,685	$1,071
2	3,751	6.2	$6,510	$1,361
3	2,098	3.5	$9,215	$2,514
4	1,665	2.7	$10,865	$3,271
5	1,634	2.7	$14,629	$6,649
6	985	1.6	$20,675	$10,355
Nursing Home	6,807	11.2	$11,303	$3,369

See notes for table 2.2

ADL's are widely used in empirical studies because they are highly predictive of medical care utilization and costs and easily interpretable. However, ADLs are inconsistently defined across

surveys. Disability rates in the MCBS tend to be higher than other surveys of the same population, particularly the fraction reporting difficulty walking. Further, some researchers argue that ADL measures are biased by cultural norms and societal roles of how older men and women function (Freedman & Martin, 1999).[4]

Self-reported health status. Another common measure of physical well-being is self-reported health status. The MCBS asks respondents to rate their general health using a 5-category Likert scale (excellent, very good, good, fair, poor). The Likert scale is widely used in national surveys and highly predictive of medical expenditures (Table 2.4). Our data indicate that nearly 70 percent of older beneficiaries report being in good to excellent health, despite the fact that over 40 percent report 1 or more ADLs. In addition, the Likert scale of general health status is highly correlated with Medicare expenditures. Older beneficiaries reporting to be in "poor" general health have a nearly 3-times the costs of those in "good" health and more than a 7-fold increase in Medicare expenses relative to those in "excellent" health.

Table 2.4. Medicare Reimbursement by Self-Reported Health Status

S.R. General Health	N (Unweighted)	% of Sample	Mean $	Median $
Excellent	8,854	14.6	$1,919	$233
Very Good	15,012	24.8	$2,639	$422
Good	18,523	30.5	$4,351	$794
Fair	12,771	21.1	$7,580	$1,728
Poor	5,339	8.8	$14,640	$5,567
Missing	152	0.3	$11,149	$2,929

See notes for table 2.2.

The principal limitation of the Likert scale is the difficulty translating advances in medical technologies and treatments to changes in self-reported health states. In other words, how we map input from the Medical TEPs on emerging technologies and treatment breakthroughs into discrete changes in health states is unclear. For this reason, the Social Science Expert Panel cautioned against using self-reported health in a forecasting model, preferring more medically-based definitions of health status and disease states.

Chronic disease. In addition to measures of physical functioning and self-reported health states, many studies characterize morbidity by the presence of chronic disease and related symptoms. The MCBS contains both self-reported and claims-based measures of specific conditions.[5] Self-reported measures are based on participant responses to ever being told by a physician they had a specific condition. Thus, they reflect lifetime prevalence of a condition. Claims-based measures are derived from diagnostic codes recorded on administrative data. As a result, they are more likely to pick up incident cases or recurrences that require acute treatment.

How disease incidence and prevalence are defined has a considerable effect on both the number of observed cases and average medical costs. Frequencies of self-reported conditions exceed claims-based measures for every condition (Table 2.5). Moreover, average Medicare reimbursements associated with self-reported diseases are substantially lower than claims-based definitions. Both of these findings underscore the distinction between prevalence and incidence. Prevalence reflects the number of existing cases in a population at a given time, or during a given period. Prevalence rates

[4] Freedman VA, Martin LG. "The role of education in explaining and forecasting trends in functional limitations among older Americans." *Demography,* Nov;36(4):461-73, 1999.

[5] We are currently working to get more historical claims data for all MCBS respondents; these additional data will change the results shown here.

are often used in health-care planning and management because they reflect the need or demand for health services between disease onset and recovery or death. Incidence reflects occurrence of new cases in a well-defined population during a given period, typically a year. It is commonly used for studying disease etiology, yet also provides more accurate information on the costs of treating an initial episode of care.

The number of self-reported and claims-based cases of specific conditions differ substantially for illnesses with high survival rates, low recurrence, or slow disease progression, such as skin cancer or Alzheimer's. Conversely, incidence and prevalence begin to converge for conditions with high mortality rates such as lung cancer.

Similarly, average Medicare expenditures within the same condition vary substantially between self-reports and claims-based measures (Table 2.5). For example, average Medicare expenses are nearly three times higher for beneficiaries with an ICD-9 code for colon cancer or arteriosclerosis than for persons with self-reported measures of those conditions. Alternatively, we observe only modest differences in expenditures for both breast and prostate cancers for self-report and claims-based measures. In part, this reflects the distinction between defining health using concurrent claims data—which by definition require some episode of care for that illness—and a self-reported measure of "ever having the disease."

Table 2.5. Medicare Reimbursement by Self-Reported Conditions

	N (Unweighted)	Mean $
Cancer	11,510	6,775
Breast	2,589	5,823
Prostate	1,786	7,937
Uterine	1,194	5,142
Colon	1,816	7,389
Bladder	548	10,070
Lung	549	12,266
Kidney	253	7,729
Throat	246	10,321
Head	207	6,406
Brain	140	12,764
Other	2,632	7,238
Heart disease	25,124	7,268
CHD	10,272	8,153
Myocardial infarction	9,742	8,853
Other	18,964	7,563
Alzheimer's	4,125	8,363
Stroke	8,335	9,228
Diabetes	10,201	8,079
Hypertension	32,812	5,764
Lung	8,633	7,533
Arthritis	34,205	5,160

See notes for table 2.2.

Neither measure is ideal. Prevalence estimates derived from patient self-reports may be sensitive to the type of data being used. Claims-based definitions are more objective, but may be biased if providers overreport high reimbursement conditions. In the present context, we are mainly concerned with predictive power and the ability to map health status into our "what-if" scenarios. In the present context, we re mainly concerned with predictive power and the ability to map health status

into our "what-if" scenarios. Claims-based definitions are superior for the later, and we explore the former later in the chapter.

Multiple conditions. The presence of a chronic illness or a functional limitation increases the likelihood that additional impairments will develop. For instance, difficulty walking can lead to lack of exercise, which decreases cardiopulmonary function and further reduces mobility. Reduced mobility, in turn, can lead to a bed disability such as bed sores, that can foster a new source of pathology and disablement process. We find that nearly one-third of older beneficiaries suffer from multiple conditions among the limited set of illnesses under study. Further, medical expenditures increase monotonically with each additional condition (Table 2.6).

Table 2.6. Total Medical Care Costs by Number of Conditions (Claims-Based)

Base Condition	No. of Additional Conditions**	Mean $	Median $	N
Cancer (11) †	0	$16,861	$10,293	2,128
	1	24,514	21,375	131
	2	33,031	26,442	10
Cardiovascular (4) †	0	$20,420	$13,109	4,012
	1	25,626	19,006	1,064
	2	37,757	30,197	234
	3	51,676	37,907	36
Neurology (3)	0	$25,213	$19,379	2,926
	1	32,609	24,304	81
	2	49,347	38,024	3
Any Condition*	0	$15,111	$7,871	8,495
	1	22,576	15,292	2,945
	2	29,147	22,899	771
	3	35,509	29,812	216
	4	49,102	32,610	41
	5	75,372	54,606	9

Note: **Cardiovascular disease-** Angina pectoris/CHD, Arteriosclerosis, Myocardial infarction and other; **Neurological disease-** Alzheimer's, Stroke and Parkinson's.

* Based on ICD-9 codes for any of the 18 conditions.

† Excludes skin cancer and hypertension.

** Within disease class only; except for "any condition."

Interaction of ADLs and chronic disease. While functional limitations and chronic diseases are correlated with medical care spending, neither measure necessarily explains costs or predicts future health states. For instance, an incident case of cancer may predict higher than average expenditures next year, as the patient receives follow-up therapy. But if the cancer goes into remission or is cured, the patient's expenditures may not be much higher than average in subsequent years (Garber et al, 1997). Similarly, an early diagnosis of prostate or breast cancer may indicate high future expenditures or concern for preventive care and health-conscious behavior that results in low medical costs in the long-run. Interacting chronic disease and functional limitations provides a more accurate assessment of underlying health and medical spending.

Table 2.7 presents average Medicare reimbursements by disease and ADL categories. We categorized ADLs into 3 groups (0, 1-2, 3+) and defined diseases separately based on patient self-reports and administrative claims. Medicare expenses rise substantially with increases in physical

limitations, particularly among persons reporting three or more ADLs. This pattern occurs consistently across conditions, for both self-reported and claims-based disease measures.

Table 2.7. Medicare Costs by Self-reported Conditions and ADL Counts

	Self-Reported			
Condition	0	1-2	3+	Nursing Home
Cancer	$4,491	$7,284	$14,025	$13,800
Breast	3,808	5,376	11,232	14,788
Prostate	5,866	8,099	17,586	14,102
Uterus	3,144	4,965	11,250	13,004
Colon	5,386	7,791	12,968	12,003
Bladder	7,734	10,637	17,170	23,652
Lung	8,458	10,602	25,446	12,761
Kidney	4,806	10,332	14,526	14,829
Throat	5,326	12,570	31,247	13,043
Head	3,349	9,482	17,527	4,995
Brain	4,397	4,816	24,737	13,001
Other	4,868	7,828	14,618	14,281
Heart	4,670	7,501	14,055	12,355
Angina pectoris/CHD	5,340	8,339	15,621	11,857
Myocardial infarction	5,928	8,783	16,952	14,087
Other	4,769	7,794	14,124	12,288
Alzheimer's	4,111	5,905	11,681	8,765
Stroke	4,776	7,830	15,434	11,942
Diabetes	4,290	8,143	15,992	16,430
Hypertension	3,457	6,256	13,200	12,773
Lung	4,247	8,079	15,033	15,343
Arthritis	3,143	5,726	11,899	11,429

See notes for table 2.2

Aggregate measures of disease. The number of disease states is potentially quite large, especially using claims-based measures. Our preliminary model takes a conservative approach to this issue by aggregating specific diseases among our clinical domains of primary interest. These are then integrated with ADL counts to create disease-disability states, as shown in Table 2.8. ADLs and medical expenditures remain positively correlated, however the rise in expenditures associated with three or more ADLs is less pronounced than in Table 8 with disaggregated disease measures. While aggregating diseases simplifies the model, it does limit interpretability somewhat by combining conditions with different pathologies and treatment protocols.

Table 2.8. Mean Medicare Costs by Self-reported Aggregate Conditions & ADL Counts

	Cancer		**Heart disease**	
ADL Count	N (Unweighted)	Cost	N (Unweighted)	Cost
0	6,542	$4,491	12,584	$4,670
1	1,519	7,090	3,519	7,293
2	820	7,662	1,915	7,901
3	517	9,828	1,147	11,249
4	386	13,351	985	12,482
5	361	15,891	920	12,246
6	221	23,061	546	21,923
Nursing home	1144	13,780	3,508	12,355

See notes for table 2.2

Predictive power. To compare the predictive power of alternative measures of health and physical functioning, we computed partial R-squared derived from a series of cost regressions. A partial R-square reflects the fraction of variation in the dependent variable explained by a specific independent variable or variables, after controlling for the effects of other regressors in the model.

The dependent variable is the log of Medicare reimbursements. The independent variables include patient demographics, year dummies, nursing home status, and alternative measures of physical health. The latter include ADL counts (with and without walking), ADL categories (0, 1-2, 3+), Likert scale of general health, self-reported and claims-based measures of disease, aggregate diseases, and various combinations of these measures.

Given that MCBS respondents answer health status questions late in the calendar year, it is unclear whether these measures should be used to predict expenditures in the current or future year. We report partial R-squares for both contemporaneous and lagged measures of the independent variables. The models are highly predictive of Medicare expenditures, with a maximum R-square of .173 in the lagged model and .343 with contemporaneous health measures.

Selected results are reported in Table 2.9. As expected, multiple measures of physical health status are more predictive of costs than single constructs. For example, the Likert scale of general health explains about 8 percent of the variation in Medicare reimbursements, after controlling for the effects of patient demographics and other independent variables. Adding self-reported diseases to this model increases the partial R-square from .081 to .129 in a model with contemporaneous regressors. Difficulty walking is an important predictor of expenditures, despite concerns the MCBS measure overstates national prevalence. Disaggregated measures of diseases (specific cancers, cardiovascular, and neurologic conditions) are only modestly more predictive of costs than aggregate measures. Finally, a Likert scale of general health status is an independent predictor of costs, even in models that include binary indicators of chronic diseases and ADLs.

Table 2.9. Partial R-squares of alternative measures of health status.

Model	Partial R^2	
	Lagged measures	Contemporaneous measures
Disability only		
5 ADLs†	0.023	0.037
ADL counts††	0.031	0.047
6 ADLs	0.032	0.049
Without Disability		
Claims (3 Categories)*	0.046	0.179
Likert	0.047	0.081
Likert & Claims (20 Conditions)**	0.047	0.223
Claims (20)**	0.050	0.180
Self-Reported Conditions (SRC)	0.055	0.086
SRC & Claims (20)	0.055	0.212
Likert & SRC	0.080	0.129
Disability and Health Status		
Likert & 6 ADLs	0.058	0.095
Likert, Claims (20), & 6 ADLs	0.058	0.233
Claims (3 Categories) & 6 ADLs	0.070	0.208
Claims (20) & 6 ADLs	0.073	0.208
SRC & 6 ADLs	0.074	0.114
SRC, Claims (20), & 6 ADLs	0.074	0.232
Likert, SRC, & 6 ADLs	0.088	0.140

†Excludes "walking".
†† ADL categories are defined as 0, 1-2, 3+
*Aggregate claims (3) categorized as cancer, cardiovascular and neurology.
**Claims-based conditions include 12 for cancer, 5 for cardiovascular and 3 neurologic.

Despite the limited duration of the MCBS panel, we examined trends in average Medicare expenditures following an incident condition for a subsample of patients with three or more years of data. Incident cases were defined by an ICD-9 diagnostic code for a specific condition in year (*t*) and the absence of the condition in year (*t-1*). We compared average expenditures in the incident year and up to three successive years based on the respondents reporting status. We examined the pattern of expenditures separately for three categories of respondents: those who died in the current year; persons who were alive and interviewed in the subsequent year; and those who attrited.

Among the 1,922 respondents with an ICD-9 code for hypertension in year (*t*) but not year (*t-1*), 125 died in year (*t*), 1,214 were alive in year (*t+1)*, and 583 attrited in (*t+1*). Among those who were alive, mean Medicare expenditures were $9,927 in the incident year (t) and modestly lower in subsequent years, although the comparison is limited by declining sample sizes. Among those who died, average yearly Medicare expenses declined by nearly $6,000 between those who died in the incident year and persons dying in the subsequent year. The reliability and statistical significance of this analysis was limited by the relatively short panel in which we observe trends in treatment costs.

COST REGRESSIONS

We impute costs in the microsimulation by computing fitted values from cost regressions. The primary dependent variables used in the cost regressions are Medicare reimbursements and their components (Part A and Part B reimbursements), and total medical experts.[6] The set of independent variables include demographics such as age, gender, ethnicity, education, and geography (region and urban residence), nursing home residence, death, and time dummies. Measures of physical health include self-reported health, ADL counts and categories, self-reported and claims-based disease indicators, and interactions of these measures. We have used both lagged and contemporaneous measures of health status.

The final regressions are based on weighted least squares rather than alternative approaches such as the two-part model or modified versions of it. Least squares is robust to asymmetric and highly-skewed errors, although there is a loss of efficiency compared to more complex estimators. The dependent variable in the model presented is total Medicare reimbursements. The contemporaneous set of independent variables are described above, with health status measures consisting of a ADL categories (0, 1-2, 3+), self-reported disease categories (binary measures of any cancer, cardiovascular, hypertension, neurology, and diabetes), and interactions of ADLs and disease.

Admission to a nursing home, ever having smoked, residing in the northeast, mortality, and physical health status have considerable effects on expenditures. Individuals who die during the year have substantially higher medical expenses than survivors, which is consistent with the literature. Medical expenditures increase with age, until about age 85. Lower expenditures among the oldest old may reflect biological differences among those who have survived to that age, as well as less aggressive medical treatment. We also find that costs increase substantially with ADLs, particularly 3 or more. The interactions of ADLs and disease vary in magnitude and significance, both in this model and other specifications. Once we have additional years of data from Medicare claims files, we can distinguish between incident and non-incident cases of treatment using claims-based measures and more fully explore interactions of disease and disability status. The final models are shown in Table 2.10.

[6] A panel of social science experts recommended not distinguishing the components of costs—e.g., inpatient, outpatient, and home health—because trends during the 1990's were so extreme, and this is the period spanned by our data.

Table 2.10. OLS estimates from MCBS cost regressions

	Total Expenditures		Medicare expenditures	
Characteristic	Estimate	Std. Error	Estimate	Std. Error
Age 70 to 74	1,218	187	630	151
Age 75 to 79	1,165	200	605	166
Age 80 to 84	1,133	222	586	178
Age 85+	-146	267	-822	212
Male	605	150	370	118
Black	817	261	984	220
Hispanic	833	354	945	263
Death	6,101	569	9,870	470
Less than high school	-233	158	75	130
Some college	251	205	110	166
College or above	154	193	-149	138
Northeast	2,308	194	1,105	151
Midwest	-16	145	-165	117
West	883	208	526	177
Other (except South)	-2,603	398	-2,345	299
1-2 ADLs (no nursing home residency)	2,968	384	1,943	304
3+ ADLs (no nursing home residency)	10,819	1,037	7,776	874
Nursing home residency	31,929	1,063	6,985	627
Diabetes	1,559	194	1,052	167
Cancer	2,278	163	1,478	133
Heart disease	2,784	133	1,988	112
Stroke	1,287	288	932	251
Alzheimer's disease	570	577	548	500
Hypertension	981	110	651	92
Arthritis	555	111	303	93
Lung disease	1,453	211	898	181
Cancer and 1-2 ADLs	-736	413	-392	340
Cancer and 3+ ADLs	-183	857	-238	728
Cancer and Nursing home residency	-110	1,322	528	1,022
Heart disease and 1-2 ADLs	53	329	57	263
Heart disease and 3+ ADLs	272	765	46	672
Heart disease and nursing home residency	-1,114	901	-1,185	699
Stroke and 1-2 ADLs	621	638	237	483
Stroke and 3+ ADLs	1,964	965	1,650	839
Stroke and nursing home residency	190	968	-914	733
Arthritis and 1-2 ADLs	-581	956	-893	740
Arthritis and 3+ ADLs	-1,553	1,236	-2,192	1,042
Arthritis and nursing home residency	-141	963	-4,306	713
Hypertension and 1-2 ADLs	-440	315	-227	249
Hypertension and 3+ ADLs	-646	712	100	607
Hypertension and nursing home residency	-860	866	939	621
Diabetes and 1-2 ADLs	1,242	457	1,166	379
Diabetes and 3+ ADLs	3,482	873	2,677	747
Diabetes and nursing home residency	5,679	1,452	4,216	1,166
Lung and 1-2 ADLs	1,455	488	1,161	408
Lung and 3+ ADLs	1,559	1,046	1,112	921
Lung and nursing home residency	259	1,647	1,795	1,294
Alzheimer's and 1-2 ADLs	-776	353	-516	286
Alzheimer's and 3+ ADLs	-3,410	899	-2,158	759
Alzheimer's and nursing home residency	-2,806	887	-418	680
Ever smoked	756	134	773	106
Spline for BMI<20	-358	156	-337	136
Spline for 20<BMI<25	-56	45	-71	36
Spline for BMI>25	-101	25	-90	21
Medicare Part A only	-2,671	341	-2,774	182
Medicare Part B only	-3,267	741	-3,112	250
Constant	7,506	3,057	6,797	2,659

CHAPTER 3. PREDICTING HEALTH STATUS

As noted previously, the microsimulation model consists of three main component models. First, parameter estimates from a health status transition model form the basis of individuals' health status forecasts from the moment at which they enter the simulation host data until they become deceased. Second, every year we rejuvenate the host data with age-65 individuals to ensure that the data remain representative of the entire population age 65 and older. We estimate a model to forecast trends in various measures of health status and adjust the relative weights of the rejuvenation sample in accordance with those trends. Third, we apply a model of health care expenditures as a function of demographic characteristics and health status to project Medicare and total health care expenditures. Chapter 2 explained the cost model; the current chapter describes the health status transition model; and Chapter 4 describes the trend model for future Medicare entrants.

Our model of health status transition probabilities is based on historical experiences of the respondents to the 1992-1998 MCBS. These data also form the basis of the microsimulation host data, so that there is no comparability issue. We pool multiple MCBS waves and use 21,495 individuals for the transitions model. Other health surveys, such as the NHIS, may have larger samples, but would lack the comparability and provide only subsets of information on subsets of respondents. The MCBS sample is very heterogeneous with respect to health status: Distinguishing six health conditions with potentially 96 combinations (cells), the 21,495 MCBS respondents span almost the entire spectrum of conditions.

The sample selection criteria are as follows. Individuals must be at least 65 years old. This yields 28,371 respondents with a total of 72,774 interview years.[7] Our outcomes are annual transitions, so we keep only individuals who participated in two or more contiguous interview years. This leaves 21,534 individuals and 65,937 interview years. Finally, we drop all interviews of individuals with any missing value for any health measure of interest or for nursing home residency. This affects 39 individuals and the final estimation sample consists of 21,495 individuals and 65,575 interview years. Each outcome (transition) requires two contiguous interview years; the 65,575 interview years translate into 44,160 interview-pairs.

The health status measures include cancer (excluding skin cancer), heart disease, neurological disorder, hypertension, diabetes, number of ADLs, and general health status. Table 2.1 presents prevalence and incidence rates in the MCBS estimation sample, including facility-based respondents but excluding respondents who were only interviewed once or had missing information, as of respondents' year of entry into the MCBS. (Tables in Chapter 1 presented prevalence rates by broad age categories in the community-based MCBS population, for comparison with NHIS prevalence rates.) Table 2.1 also includes the percent of respondents that was interviewed in a facility and the distribution.

[7] Health status information is only collected in the fall interview round, so for our purposes, there is only one interview per year.

Table 3.1. Prevalence and Incidence of Select Conditions, MCBS Estimation Sample

	Prevalence			Incidence		
Condition	**65+**	**65-69**	**70+**	**65+**	**65-69**	**70+**
Mortality				3.3	1.2	4.1
Cancer	18.6	15.2	19.9	1.8	1.5	1.9
Breast (women only)	6.5	6.4	6.5			
Prostate (men only)	5.2	3.4	6.0			
Uterus (women only)	3.0	3.1	3.0			
Colon	2.7	1.6	3.1			
Bladder	0.9	0.5	1.0			
Lung	0.8	0.8	0.8			
Kidney	0.4	0.4	0.4			
Throat	0.4	0.4	0.4			
Head	0.3	0.2	0.4			
Brain	0.2	0.2	0.2			
Other	4.3	3.4	4.7			
Heart disease	38.7	29.6	42.2	3.2	2.3	3.5
Angina pectoris/CHD	15.8	11.8	17.4			
Myocardial infarction	15.0	12.2	16.1			
Other	29.0	21.2	32.0			
Alzheimer's	4.9	1.1	6.4	1.2	0.3	1.5
Stroke	11.8	7.6	13.5	1.4	0.8	1.7
Diabetes	16.6	15.2	17.1	1.3	1.1	1.3
Hypertension	54.1	48.3	56.3	3.0	2.6	3.2
Lung[1]	14.1	13.0	14.6	1.4	0.9	1.5
Arthritis	56.3	47.6	59.7	4.4	4.4	4.4
Disability						
ADL>=1	30.8	21.2	34.5			
ADL>=3	10.3	5.6	12.1			
Nursing home	6.8	2.3	8.6	1.5	0.2	2.0

[1] Refers to lung disease which excludes lung cancer

Note that incidence rates increase sharply with age for, in particular, cardiovascular disease, neurological disorder, and entry into a (nursing home) facility. The next set of tables present the distributions of age, sex, race, Hispanic ancestry, education, smoking (by sex), and marital status. All tabulations are based on the first interview year.

Table 3.2. Age Distribution, MCBS Estimation Sample

Age	Freq.	Percent
65-69	5,551	25.82
70-74	3,969	18.46
75-79	4,115	19.14
80-84	4,155	19.33
85-89	2,385	11.10
90-94	1,016	4.73
95-99	264	1.23
100+	40	0.19
Total	21,495	100.00

Table 3.3. Distribution of Sex, MCBS Estimation Sample

	Freq.	Percent
Female	12,914	60.08
Male	8,581	39.92
Total	21,495	100.00

Table 3.4. Distribution of Race, MCBS Estimation Sample

	Freq.	Percent
Native American	145	0.67
Asian, Pacific Islander	255	1.19
African American	1,985	9.23
White	19,110	88.0
Total	21,495	100.00

Table 3.5. Distribution of Hispanic ancestry, MCBS Estimation Sample

	Freq.	Percent
Non-Hispanic	20,325	94.56
Hispanic	1,170	5.44
Total	21,945	100.00

Table 3.6. Distribution of Educational Attainment, MCBS Estimation Sample

	Freq.	Percent
High school drop-out	9,248	43.02
High school graduate	6,575	30.59
Some college	2,892	13.45
College graduate	2,780	12.93
Total	21,495	100.00

Table 3.7. Distribution of Ever Smoked, by Sex, MCBS Estimation Sample

	Women		Men	
Ever Smoked?	Freq.	Percent	Freq.	Percent
No	7,876	60.99	1,789	20.85
Yes	5,038	39.01	6,792	79.15
Total	12,914	100.00	8,581	100.00

Table 3.8. Distribution of Currently Smoking, by Sex, MCBS Estimation Sample

	Women		Men	
Smoke Now?	Freq.	Percent	Freq.	Percent
No	11,552	90.17	7,171	84.20
Yes	1,259	9.83	1,346	15.80
Total	12,881	100.00	8,517	100.00

Table 3.9. Distribution of Marital Status, MCBS Estimation Sample

	Freq.	Percent
Single	10,730	49.92
Married	10,765	50.08
Total	16,839	100.00

MISSING DATA

As stated above, respondents with missing information on health conditions or facility residence were dropped from the estimation sample. For demographic characteristics, we attempted to fill in missing data from other waves and from CMS's program records on sex, date of birth, and race/ethnicity. Small numbers of missing variables remained. We imputed these variables randomly in accordance with their MCBS sample distributions. For smoking, we imputed separately for men and women. All imputed variables were flagged with indicator variables. At first, we included these indicator variables in all transition models. However, very few turned out to be significant, indicating that variables were missing at random with respect to health transitions. We therefore omitted indicators for missing variables from our final model specifications.

RESULTS OF ESTIMATION

The health conditions that we use in our analysis are all self-reported. One may expect health measures based on claims data to be more predictive of costs. In addition, medical costs vary by duration since the onset of a condition and tend to be particularly high in the final year of life. In order to account for these duration effects, it is required to know the year of onset of each condition. We are currently working with CMS to obtain historical claims records. The preliminary results in this report, however, are based on self-reported health conditions without information on the year of onset.

Mortality is an absorbing state. For cancer, cardiovascular disease, neurological disorder, diabetes, and hypertension, the MCBS questions were worded as "Did a doctor ever tell you that ..." In other words, the question wordings define these conditions as absorbing states. Accordingly, we only model transitions into these states, without allowing for recovery. Similarly, we assume that residence in a facility is an absorbing state. We model transitions into mortality, cancer, cardiovascular disease, neurological disorder, diabetes, hypertension, and facility residence as proportional hazard models:

$$\ln h_j(t) = \gamma' Age(t) + \beta X_j,$$

where $\ln h_j$ is the log-hazard of onset of the *j*-th condition (including mortality and entry into a facility); $Age(t)$ is a piecewise-linear spline transformation of age at time *t* (see below); and X_j are demographic characteristics and co-morbidities that affect the onset of condition *j*.

The baseline duration dependency is the dependency on respondent age, $\gamma' Age(t)$. The hazards of various conditions' onset are assumed to be linear in age, with potentially different slopes before and after age 77, i.e., the baseline log-hazard is piecewise-linear (also known as piecewise Gompertz or generalized Gompertz).[8]

The unit of observation is an interview-pair. All explanatory covariates are measured with a one-year lag. Only individuals who, at the time of the first interview, did not suffer from a specific condition contribute to the model estimation. The sample sizes for various health status transition models vary therefore. For example, consider an individual who entered the MCBS in 1993 without

[8] Formally, γ is a vector of two age slopes and $Age(t)$ is a spline transformation, $Age(t) = \begin{pmatrix} \min(A, 77) \\ \max(0, A - 77) \end{pmatrix}$, where *A* is (scalar) age at time *t*.

cancer but with a heart condition. In 1994, his conditions are unchanged; in 1995, he is diagnosed with cancer; in 1996, his conditions are unchanged. This person starts out with a heart condition, so he does not at all contribute to the heart disease transition model. In 1993 and 1994, he is free of cancer, so he contributes two observations to the cancer transition model. The outcome in his first contribution (1993 to 1994) is zero, because he remained free of cancer; the outcome in his second contribution (1994 to 1995) is one, because he was diagnosed with cancer. He is out of the sample for subsequent years. We ignore the clustering that arises from the fact that the same individual may contribute more than once to a model.

Table 3.10 presents the results of estimation for hazard models of onset of cancer, heart disease, stroke, Alzheimer's, hypertension, diabetes, lung disease, arthritis, disability, and entry into a facility. The coefficients on age indicate the baseline slopes on age. They are generally positive, i.e., the risks of onset of various conditions tends to increase with age. It may surprise that the age coefficients tend to be smaller after age 77 than before, i.e., that there is a deceleration in the risk pattern. Note, however, that this age pattern applies only to individuals without any co-morbidity. As individuals get older, they are more likely to suffer from various conditions, which have positive effects on the onset of other conditions. The net result is typically an acceleration of the log-hazard with age. We return to this issue below, in the discussion of mortality.

Table 3.10. Results of Health Transition Estimation
(Log-hazard parameters)

	Cancer	Heart	Stroke	Alzheimer's	Hypertension	Diabetes	Lung	Arthritis	ADL1+	ADL3+	Nursing home
Cancer			-0.1121						0.1234[2]	0.1763[2]	-0.0961
			(0.0906)						(0.0488)	(0.0670)	(0.0987)
Heart disease			0.2661[3]						0.1472[3]	0.2273[3]	-0.0597
			(0.0819)						(0.0400)	(0.0569)	(0.0817)
Stroke									0.2579[3]	0.5653[3]	0.4320[3]
									(0.0600)	(0.0712)	(0.0891)
Alzheimer's									-0.9946[3]	-0.4609[3]	1.1078[3]
									(0.1220)	(0.1275)	(0.1062)
Hypertension		0.4723[3]	0.3768[3]						0.2314[3]	0.2317[3]	-0.0946
		(0.0569)	(0.0858)						(0.0404)	(0.0603)	(0.0824)
Diabetes		0.2598[3]	0.2646[2]		0.2399[3]				0.2121[3]	0.4148[3]	0.3063[3]
		(0.0726)	(0.1049)		(0.0832)				(0.0511)	(0.6713)	(0.0973)
Lung									0.4215[3]	0.2760[3]	0.0279
									(0.0519)	(0.0734)	(0.1122)
Arthritis									0.4987[3]	0.5052[3]	-0.1952[2]
									(0.0404)	(0.0613)	(0.0834)
ADL>=1											0.9173[3]
											(0.1027)
ADL>=3											0.4708[3]
											(0.0932)
Age<77 (spline)	0.0588[3]	0.0721[3]	0.0653[3]	0.1739[3]	0.0441[3]	0.0581[3]	0.0472[3]	0.0461[3]	0.0845[3]	0.0919[3]	0.1913[3]
	(0.0119)	(0.0096)	(0.0141)	(0.0197)	(0.0091)	(0.0142)	(0.0142)	(0.0072)	(0.0065)	(0.0102)	(0.0218)
Age>77 (spline)	-0.0102	0.0223[3]	0.0297[3]	0.0904[3]	0.0058	-0.0520[3]	0.0031	0.0059	0.0224[3]	0.0504[3]	0.0804[3]
	(0.0097)	(0.0067)	(0.0094)	(0.0083)	(0.0069)	(0.0135)	(0.0103)	(0.0061)	(0.0052)	(0.0061)	(0.0071)
Ever smoked	0.1498[1]	0.0394	0.2168[2]				0.7279[3]		0.2355[3]	0.1113[1]	0.0331
	(0.0842)	(0.0609)	(0.0934)				(0.0999)		(0.0436)	(0.0637)	(0.0869)
Under Weight		0.0963	0.3118[3]		-0.2386[3]	-0.2291[2]		-0.3483[3]	-0.0942[2]	0.1200[1]	0.4810[3]
		(0.0636)	(0.0889)		(0.0639)	(0.1100)		(0.0552)	(0.0462)	(0.0643)	(0.0813)
Obese		0.2431[3]	-0.1093		0.2911[3]	0.7130[3]		0.2642[3]	0.3953[3]	0.3575[3]	-0.2620[1]
		(0.0741)	(0.1274)		(0.0819)	(0.1038)		(0.0657)	(0.0524)	(0.0767)	(0.1514)
Male	0.3927[3]	0.1549[2]	0.0966	-0.0556	-0.2114[3]	0.0669	-0.0862	-0.2966[3]	-0.1929[3]	-0.1837[3]	-0.0886
	(0.0787)	(0.0601)	(0.0909)	(0.0950)	(0.0571)	(0.0889)	(0.0894)	(0.0477)	(0.0435)	(0.0661)	(0.0901)
Black	-0.0747	-0.0337	-0.0422	0.2375[1]	0.4792[3]	0.2300	-0.4448[2]	0.1371	0.0927	0.0807	-0.2033
	(0.1347)	(0.0959)	(0.1480)	(0.1437)	(0.1026)	(0.1450)	(0.1761)	(0.0847)	(0.0682)	(0.0949)	(0.1412)
Hispanic	-0.3389[1]	-0.1142	-0.2684	-0.2596	0.2647[2]	0.4259[2]	0.2825[1]	0.0025	0.1422	0.2047[1]	-1.0555[3]
	(0.1779)	(0.1200)	(0.1968)	(0.2247)	(0.1131)	(0.1668)	(0.1587)	(0.1048)	(0.0816)	(0.1136)	(0.2601)
HS drop-out	0.0855	0.1188[2]	0.2182[2]	0.2413[2]	0.1280[2]	0.2045[2]	0.1869[2]	0.0963[1]	0.1470[3]	0.3008[3]	0.1550[1]
	(0.0809)	(0.0591)	(0.0856)	(0.0952)	(0.0606)	(0.0935)	(0.0888)	(0.0519)	(0.0421)	(0.0612)	(0.0830)
College graduate	0.1319	-0.0423	-0.2241	-0.0968	-0.1692[1]	0.1225	-0.2535[1]	0.0406	-0.1906[3]	-0.0251	-0.3503[3]
	(0.1060)	(0.0867)	(0.1398)	(0.0952)	(0.0885)	(0.1318)	(0.1395)	(00685)	(0.0641)	(0.0976)	(0.1496)
Constant	-8.4097[3]	-8.800[3]	-9.6897[3]	-17.8567[3]	-5.9379[3]	-8.5798[3]	-8.0812[3]	-5.5237[3]	-9.3354[3]	-11.359[3]	-19.496[3]
	(0.8705)	(0.6694)	(1.0448)	(1.4771)	(0.6617)	(1.0376)	(1.0440)	(0.5257)	(0.4788)	(0.7581)	(1.6485)
ln-L	-3799.23	-5402.80	-3185.27	-2577.32	-4877.19	-2861.02	-3047.76	-6277.41	-8813.26	-5403.68	-2775.01

NOTE: Asymptotic standard errors in parentheses;
Significance: '[1]'=10%; '[2]'=5%; '[3]'=1%.

Positive coefficients in Table 3.10 indicate a higher hazard and thus poorer health. The coefficients indicate shifts in the log-hazard and thus proportional shifts in the hazard or risk of onset. For example, hypertension increases the log-hazard of heart disease by 0.4723, i.e., it increases the risk of heart disease by 100*(exp(0.4723)-1) = 60.37 percent. Table 3.11 provides the same information as Table 3.10, but with log-hazard coefficients transformed into percent changes in the various hazards (relative risks).

Table 3.11. Results of Health Transition Estimation
(Relative risks)

	Cancer	Heart	Stroke	Alzheimer's	Hypertension	Diabetes	Lung	Arthritis	ADL1+	ADL3+	Nursing home
Cancer			-10.60						13.13[2]	19.28[2]	-9.16
Heart disease			30.49[3]						15.86[3]	25.52[3]	-5.80
Stroke									29.42[3]	76.00[3]	54.03[3]
Alzheimer's									-63.01[3]	-36.93[3]	202.77[3]
Hypertension		60.37[3]	45.76[3]						26.04[3]	26.07[3]	-9.03
Diabetes		29.67[3]	30.29[2]		27.11[3]				23.63[3]	51.41[3]	35.84[3]
Lung									52.42[3]	31.78[3]	2.83
Arthritis									64.66[3]	65.73[3]	-17.73[2]
ADL>=1											150.25[3]
ADL>=3											60.13[3]
Ever smoked	16.16[1]	4.02	24.21[2]				107.07[3]		26.55[3]	11.77[1]	3.37
Under Weight		10.11	36.59[3]		-21.23[3]	-20.48[2]		-29.41[3]	-8.99[2]	12.75[1]	61.77[3]
Obese		27.52[3]	-10.35		33.79[3]	104.01[3]		30.24[3]	48.48[3]	42.98[3]	-23.05[1]
Male	48.10[3]	15.75[2]	10.14	-5.41	-19.05[3]	6.92	-8.26	-25.67[3]	-17.54[3]	-16.78[3]	-8.48
Black	-7.20	-3.31	-4.13	26.81[1]	61.48[3]	25.86	-35.90[2]	14.69	9.71	8.40	-18.40
Hispanic	-28.74[1]	-10.79	-23.54	-22.86	30.30[2]	53.10[2]	32.64[1]	0.25	15.28	22.72[1]	-65.20[3]
HS drop-out	8.93	12.61[2]	24.38[2]	27.29[2]	13.66[2]	22.29[2]	20.55[2]	10.11[1]	15.84[3]	35.09[3]	16.77
College graduate	14.10	-4.14	-20.08	-9.23	-15.57[1]	13.03	-22.39[1]	4.14	-17.35[3]	-2.48	-29.55

NOTE: Asymptotic t-statistics in parentheses;
Significance: [1]=10%; [2]=5%; [3]=1%.

All explanatory covariates are measured with a one-year lag, i.e., as of the first interview of the interview-pair. Note the very powerful cross-effects of health conditions. Neurological disorder, diabetes, hypertension, and self-reported disability all significantly increase the risk of developing a heart condition; self-reported disability increases the risk of all transitions, except for contracting cancer; et cetera. As explanatory covariates, ADLs are measured marginally. For example, the effect of three or more ADLs is found by adding up the coefficients of ADL>=1 and ADL>=3.

Men tend to have higher risks of cancer and heart disease than women and lower risks of hypertension. Blacks and Hispanics have higher risks of hypertension. Hispanics also have higher risks of diabetes. Hispanics are far less likely than non-Hispanics to enter a facility, such as a nursing home. Better educated individuals tend to be in better health. Having ever smoked increases the risk of cancer, but not by very much and only marginally significantly. We do not control for current smoking behavior. Its effects often appeared counterintuitive, and we question the accuracy of respondents' reports. In addition, inclusion of current smoking behavior would require projections of future smoking behavior for the microsimulation model. We prefer to omit this covariate.

The model specifications do not control for household income since the MCBS data are of poor quality (Goldman and Smith, 2001).[9] In early model development stages, we included indicator variables that flag whether race, Hispanic ancestry, education, past smoking, and marital status were missing and imputed. Their coefficients were rarely significant, indicating that there is no systematic pattern in the missing rates of demographic covariates. We therefore need not include these indicator variables.

The estimates of Table 3.10 form the basis of the health status projection algorithms in the microsimulation model. Table 3.12 shows the estimates of the hazard model of *mortality*. The first and second columns show log-hazard coefficients; the third shows percent changes in the mortality risk. These estimates are based on MCBS data. The MCBS may or may not capture all deaths, so the next subsection compares MCBS estimates to Vital Statistics.

[9] Goldman D, Smith J. "Commentary: Methodological Biases in Estimating the Burden of Out-of-Pocket Expenses." *Health Services Research*, 35(6):1357-1370, 2001.

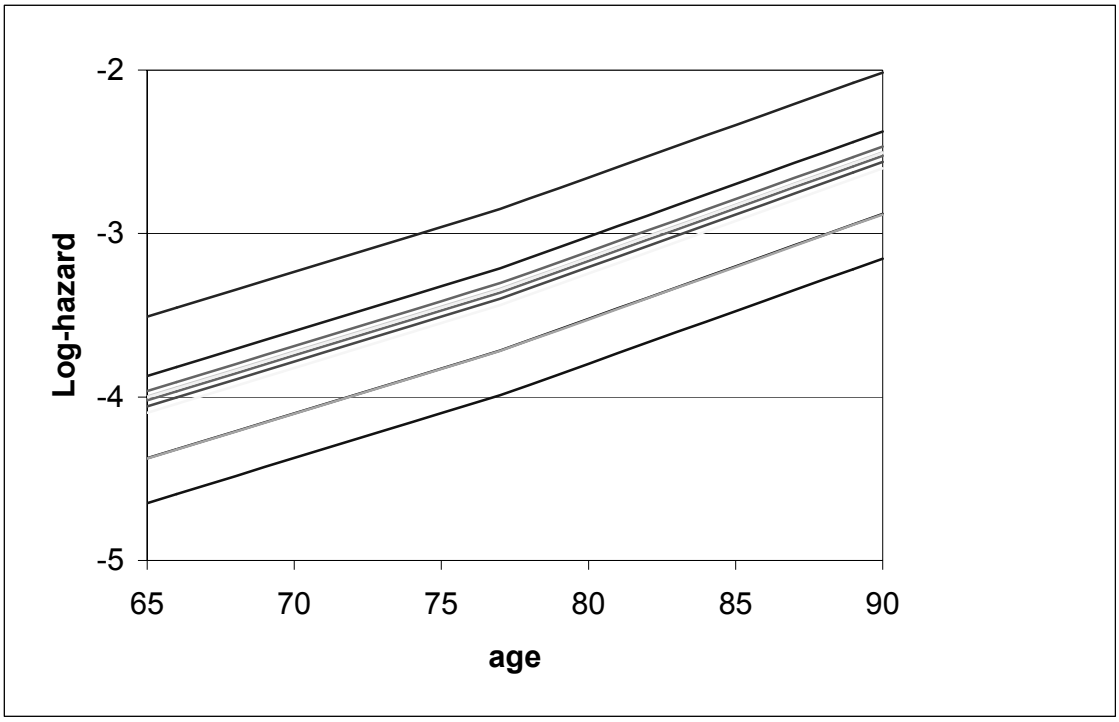
Table 3.12. Results of Mortality Estimation
(Log-hazard parameters and relative risks)

	Log-hazard coefficients		**Percent hazard changes**
	Male	**Female**	
Age<77	0.0547 ***	0.0932 ***	
	(0.0114)	(0.0130)	
Age>77	0.0641 ***	0.0707 ***	
	(0.0065)	(0.0051)	
Constant	-7.9263 ***	-11.2608 ***	
	(0.8371)	(0.9688)	
Cancer	0.3199 ***		37.70 ***
	(0.0499)		
Heart disease	0.4103 ***		50.73 ***
	(0.0450)		
Stroke	0.3785 ***		46.01 ***
	(0.0515)		
Alzheimer's	0.8654 ***		137.60 ***
	(0.0599)		
Diabetes	0.5044 ***		65.60 ***
	(0.0515)		
Lung	0.3557 ***		42.72 ***
	(0.0548)		
Arthritis	-0.2727 ***		-23.87 ***
	(0.0467)		
Hypertension	-0.0039		-0.39
	(0.0454)		
ADL>=1	0.2766 ***		31.86 ***
	(0.0551)		
ADL>=3	0.3711 ***		44.93 ***
	(0.0625)		
Ever smoked	0.1785 ***		19.54 ***
	(0.0519)		
Under weight	0.4428 ***		55.71 ***
	(0.0474)		
Obese	-0.0961		-9.16
	(0.0759)		
Black	0.0716		7.42
	(0.0760)		
Hispanic	-0.2753 **		-24.07 **
	(0.1112)		
High school drop-out	0.1172 **		12.43 **
	(0.0463)		
College graduate	-0.2564 ***		-22.62 ***
	(0.0771)		
ln-L	-7511.37		

NOTE: Asymptotic t-statistics in parentheses;
Significance: '*'=10%; '**'=5%; '***'=1%.

As before, all explanatory covariates are measured with a one-year lag, i.e., as of the first interview of the interview-pair. All health conditions increase the risk of mortality, except hypertension. Figure 3.1 illustrates the effects of morbidities on mortality risk.

Figure 3.1. Log-hazard of Mortality for Men with Selected Health Conditions



The figure illustrates several features. First, the overall age pattern is increasing, i.e., older men face higher mortality risks. There is a kink in the age pattern at age 77. Before age 77, the log-hazard increases 0.0547 (about 5.5 percent) per year. After age 77, the increase is 0.0641 (about 6.4 percent) per year; see Table 3.12. Healthy individuals enjoy the lowest mortality risks. Cancer, heart disease, neurological disorder, and disability increasingly elevate mortality risks. Their effects are to shift the age pattern parallel to the baseline (healthy) pattern. This parallel shift is a consequence of the assumed functional form. A shift in the log-hazard translates into proportional or relative changes in the hazard.

While the log-hazard of mortality appears to decelerate at higher ages, the actual pattern for any one individual may well accelerate. For example, someone may be healthy at age 65 and experience the lowest mortality log-hazard. If this person contracts, say, heart disease at age 70, he moves from the baseline curve to the heart disease curve. If further complications develop, he moves to even higher curves. The implication is that many individuals experience accelerating mortality risks.

Table 3.12 also reports the effects of demographic factors. In light of large sex differences in mortality risks, we allowed for a full sex interaction in age. The interaction terms are jointly strongly significant. Controlling for all health conditions, there is no differential mortality risk by race or Hispanic ancestry. Better educated individuals tend to live longer. Having ever smoked increases the risk of dying, even conditional on cancer.

The model does not control for marital status, even though it is highly significant and strongly predictive of men's mortality risk (and both sexes' entry rates into facilities). The reason for its exclusion is that inclusion would require an auxiliary model of marital status in order to project

future marital status for the microsimulation exercise. We intend to develop such a model in the next iteration, as we also did for the Model of Income in the Near Term (MINT) that we developed for the Social Security Administration. The estimates of Table 3.12 form the basis of the mortality projection algorithms in the microsimulation model. A correction will apply, as explained below, but that correction is minuscule.

MORTALITY

The mortality estimates of Table 3.12 are based on survival probabilities in the MCBS. While the MCBS is presumably representative of the elderly U.S. population, it is not a priori clear whether the resulting mortality rates are representative of mortality rates among all American elderly. It may be, for example, that deceased individuals could not be located and were incorrectly classified as attrited. This would bias mortality estimates down.

The Model of Income in the Near Term (MINT) that we developed for the Social Security Administration corrected for underdetection of mortality. Its mortality model was based on the 1968-93 Panel Study of Income Dynamics. We follow a similar procedure here.

The MCBS data are from 1992-1998, too short a time span to identify a longevity trend. We therefore compare the MCBS mortality data to cross-sectional 1995 Vital Statistics of the United States. We convert Vital Statistics lifetables into mortality spells[10] and estimate very simple hazard models, by sex, which only depend on age. We wish to compare these estimates to similar estimates based on the MCBS. To that end, we impose the Vital Statistics coefficients on MCBS data and estimate differential coefficients.

Table 3.14. Mortality Hazard Estimates (based on Vital Statistics and the MCBS)

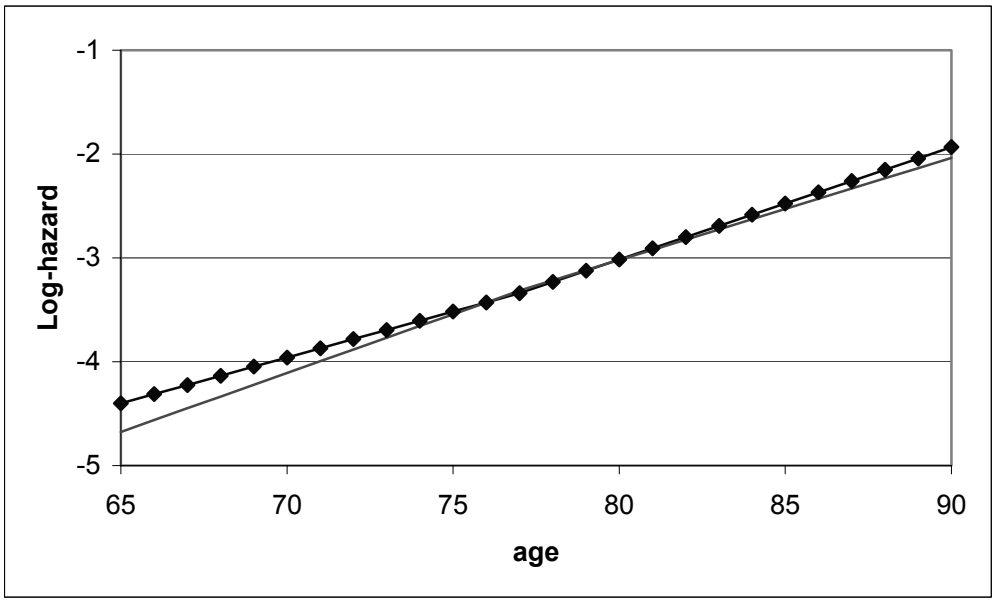
	Vital Statistics	**MCBS (marginal coeff)**
Males:		
Constant	-9.2085 ***	-0.1291
	(0.0161)	(0.8321)
Age<77	0.0819 ***	0.0021
	(0.0002)	(0.0113)
Age>77	0.0971 ***	-0.0098
	(0.0003)	(0.0065)
Females:		
Constant	-10.1469 ***	-1.8924 **
	(0.0177)	(0.9520)
Age<77	0.0884 ***	0.0249 *
	(0.0002)	(0.0128)
Age>77	0.1082 ***	-0.0098 **
	(0.0003)	(0.0048)
ln-L	-13073761.48	-8012.46

NOTE: Asymptotic t-statistics in parentheses;
Significance: '*'=10%; '**'=5%; '***'=1%.

[10] For example, the male lifetable for 1995 states that out of 85,507 men age 65, 79,037 (92.4 percent) will survive to age 70. Census data indicate that there are 5.4 million men age 65-70 in 1995. We combine this information and create two hazard spells, one for survivors and one for men who decease between their 65th and 70th birthdays. The first spell spans five years (age 65-70) and is open; it carries a weight of 0.924*5.4 million; the second also spans five years (age 65-70) but is closed; it carries a weight of 0.076*5.4 million. We do this for all age categories above age 65 and for both sexes.

The Vital Statistics coefficients are very precisely estimated due to the huge underlying population size. For males, the MCBS estimates are not significantly different from Vital Statistics estimates. For females, there is a difference in the age slope under age 77. This difference is partially compensated by a seemingly very large intercept difference, but this intercept operates at birth. At age 65, the intercept difference is only $-1.8924+65*0.0249=-0.27$. Figure 3.2 illustrates estimated age patterns for females based on Vital Statistics and the MCBS.

Figure 3.2. Log-hazard of male mortality based on Vital Statistics and the MCBS



The patterns are statistically and visually different, but are these differences substantial? A priori, it is impossible to tell and we therefore anchor mortality estimates on Vital Statistics. This is done as follows. First, we estimate mortality models that control for health conditions and demographic characteristics using MCBS data. Second, for the purpose of projections, we correct for differences between Vital Statistics and the MCBS by subtracting the marginal coefficients of the MCBS (last column of Table 3.14) from model estimates. The resulting projection parameters lead to the same aggregate mortality rates as Vital Statistics parameters, with the advantage of differentiating mortality risk by health conditions and demographic characteristics. Our simulations showed very little difference between projection algorithms based on Vital Statistics or MCBS. Stated differently, the MCBS does an outstanding job identifying deceased respondents.

CHAPTER 4. PROJECTING THE HEALTH STATUS OF FUTURE MEDICARE ENTERING COHORTS

This subtask is designed to predict the health status of each of the future entering cohorts of Medicare patients between the years 2001 and 2030. While it may be plausible to look simply at 65 year olds in the year 2000 to predict the presence of chronic conditions and disability among 65 year olds in 2001, such a procedure is likely to lead to misleading predictions for future entering cohorts. This is especially true given the presence of well-known trends in the prevalence of disease and disability among all adult age cohorts. If these trends continue, the health of 65 years olds in 2030 is likely to look considerably different from the 65 year olds today.

The measures of health status that we are most interested in here are the presence of seven of the most important, costly, and devastating chronic conditions that inflict the Medicare population. These are: heart disease, hypertension, cerebrovascular disease, Alzheimer's disease or senile dementia, cancer[11], diabetes, and chronic obstructive pulmonary disease (COPD). In addition, we project future trends in the prevalence of disability among incoming Medicare cohorts. Our measure of disability focuses upon self-reports by respondents regarding their ability to perform basic tasks of daily living, including bathing, dressing, and feeding oneself.

DATA

We use data from the National Health and Interview (NHIS), which is a large annual data set collected by the National Center for Health Statistics (NCHS). This is the right data set for our purpose because it is specifically designed to measure the population prevalence levels of a large number of chronic disease conditions and disability. Unlike another NCHS data set, the National Health and Nutrition Examination Survey (NHANES), the NHIS does not contain any physical exam or clinical data; health status is elicited from survey respondents using self-reports. However, unlike the NHANES, the NHIS is available for every year since 1957, contains large sample sizes, and uses essentially the same questionnaire in every year between 1982 and 1996. Because the survey instrument was redesigned in 1997, we use annual data between 1990 and 1996 to construct our projections. In addition, the NHIS contains extensive demographic and economic information about its respondents.

One drawback to the NHIS data relates to its sampling scheme. Rather than asking all respondents about the presence or absence of a large number of disease conditions, the NHIS randomly divides the sample into six groups. Each respondent in any given one of the six groups is asked about a different set of diseases than respondents in the other five groups. Therefore, no respondent is ever asked about the presence or absence of all of the chronic conditions considered by the NHIS. In fact, for a large subset of conditions, there is no overlap across the chronic condition questions list posed to each of the groups. However, the NHIS questionnaire also includes a list of questions regarding a small subset of chronic conditions that are posed to all respondents with some activity limitations.

[11] We included COPD and stroke in this analysis in anticipation that they might be added to the model later. A broader measure of neurological impairment beyond Alzheimer's cannot be constructed from the NHIS for the reasons articulated in Chapter 2. Further discussion of how this work feeds into the rest of the model is contained in Chapter 8 and 10.

Among the seven chronic conditions that we consider, questions regarding heart disease and hypertension are both posed to the same group of randomly selected respondents, while a question on diabetes is posed to a different group. There is no comprehensive question on cancer that is asked to all respondents. Instead, different groups of randomly selected respondents are asked about the most common types of cancer. Questions on breast and prostate cancer are posed to one group (the same group asked the question on diabetes), a question on lung cancer is posed to a second group, while a question on lung cancer is posed to yet a third group. We construct our estimates for total cancer incidence by summing over the incidence rates for each of the cancers separately.[12] Finally, a question regarding Alzheimer's disease is posed to NHIS respondents who report activity limitations.

Each year several questions regarding disability status are posed about NHIS respondents who are between 25 and 69 years old. These disability questions include "Does any impairment or health problem now keep [you] from working at a job or business?" (*Work Limitation*), "[Are you] limited in any way in any activities because of an impairment or health problem?" (*Activity Limitation*), and "Because of any impairment or health problem, [do you] need the help of other persons with – personal care needs, such as eating, bathing, dressing, or getting around the house?" (*Self-Care Limitation*). In addition, the NHIS includes a question on general health status (*General Health Status*) measured on a five-point Likert scale that is posed to everyone in the data set. Unfortunately, none of these disability and health status questions map naturally to the ADL measures that are asked in the MCBS, so they cannot be used to directly infer changes in the prevalence profiles of people unable to perform one or more ADLs (ADL 1+), or three or more ADLs (ADL 3+).[13]

In 1995, however, the NHIS included an extensive supplement that posed a version of the ADL questions to its respondents.[14] Because the usual disability and health status questions were also posed to NHIS respondents in that year, we use these data to construct map that predicts the presence of limitations in ADL from the usual NHIS questions on disability and health status. For the 1995 data, we estimate an ordered probit model that relates the total number of ADLs that respondents have difficulty performing to the Work Limitation, Activity Limitation, Self-Care Limitation, and General Health Status responses, in addition to a quadratic polynomial in age, and sex. The results of this model are presented in Table 4.1. The signs of the coefficients are consistent with common sense—older, sicker patients with more severe activity, work, or self-care limitations are more likely to report more limitations in performing ADLs. In turn, we use this model to predict ADL 1+ and ADL 3+ for each NHIS respondent in the years when these ADL questions were not asked. It is these predicted values that we use in our simulations of disability prevalence.

[12] The NHIS questions for cancer are of the form "During the past 12 months, did anyone in the family have…", so they are best interpreted as incidence rather than prevalence rates. This is in contrast to the NHIS questions for hypertension and heart disease, which are of the form "Has anyone in the family ever had…" When we sum over the incidence of the various cancer types, we implicitly assume that the incidence of each cancer type is independent of the others. This is reasonable because it is rare for cancer to emerge simultaneously at two different primary sites.

[13] These are the indicators of disability status that our MCBS-based microsimulation model currently uses.

[14] There are some differences in the ADL questions posed in the 1995 NHIS supplement and in the MCBS. These differences lead to lower estimates of difficulty performing ADLs in the NHIS compared with the MCBS, as discussed in Chapter 2.

Table 4.1. Ordered Probit Model of Number of ADL Limitations

Variable	Estimate	Statistic
General Health*		
Very Good	.0329	0.425
Good	.205	2.91
Fair	.333	4.42
Poor	.692	8.96
Work Limitation**		
Limited in kind/amount of work	-.222	-2.70
Limited in other activities	-.111	-1.27
Activity Limitation***		
Limited in kind/amount of major activity	-.0414	-0.577
Limited in other activities	-.595	-5.94
Self-Care Limitation****		
Limited in performing routine needs	-1.29	-19.8
Not limited in performing personal care or routine needs	-2.00	-32.9
No Limitations	-2.19	-18.9
Age	.00735	0.581
Age^2	-.0000375	-0.289
Male	.0449	1.19
Cut Points		
Between 0 and 1 ADL	.166	.303
Between 1 and 2 ADL	.526	.303
Between 2 and 3 ADL	.778	.303
Between 3 and 4 ADL	1.01	.304
Between 4 and 5 ADL	1.22	.304
Between 5 and 6 ADL	1.75	.308
Log Likelihood	-3854.71	
Pseudo-R2	0.383	
N	51423	

*General Health Status = Excellent is the excluded category.
**Work Limitation = Unable to perform work is the excluded category
***Activity Limitation = Unable to perform major activity is the excluded category
****Self-Care Limitation = Unable to perform personal care needs is the excluded category

Finally, in addition to NHIS data, we need information on overall and cause-specific age-mortality profiles for each year between 1990 and 1996 inclusive. We obtain these data from the annual analysis on death certificate data, Vital Statistics of the United States, conducted by the NCHS (1992-1998). In the next section, we discuss how we combine these data to obtain disease prevalence projections for future incoming Medicare cohorts.

METHODS

Our strategy to predict the health status of future cohorts proceeds in four stages. First, for each chronic disease condition of interest, we use the NHIS data to obtain age-specific prevalence information. Though the NHIS has a large sample size overall, for some age-cohorts the sample size is insufficient to produce noise-free estimates of low prevalence diseases. Thus, we introduce a

method to smooth the NHIS age-specific prevalence profiles, while at the same time accounting for trends in disease prevalence.

Second, we use a synthetic cohort-based procedure to obtain age-specific incidence rates from the smoothed prevalence profiles. In particular, we compare the prevalence of a disease in one year for one age-cohort with the prevalence rate of that disease in the next year of data (where that cohort has aged by one year). Our procedure adjusts these raw prevalence differences to account for population and disease-specific death rates.

Third, we combine information from the most recent NHIS with our estimated age-specific incidence rates to obtain our predictions about the health status of the future incoming Medicare cohorts. For example, we add the prevalence of disease among 64 year olds in 2000 to our estimated incidence rate for that disease among 64 year olds to obtain our predictions about the 2001 class of 65 year olds.

Fourth, we take our estimates of future prevalence among the entering cohort and use them to construct adjustments to the population weights of future entering cohorts with the various disease conditions.

Step 1: Smoothed Age-Specific Prevalence Rates

In order to describe the method we use to produce smooth age-specific prevalence functions—the overlap polynomial method—it is helpful to introduce some notation. The NHIS is a repeated cross section with hundreds of thousands (say, N) observations. Each observation i, taken in $year_i$, consists of information about i's self-reports regarding disease conditions and disabilities, age (age_i), and other information (X_i). In the remainder of this section, we consider one disease condition, but extending the analysis to other conditions is straightforward. Let d_i indicate whether patient i has some chronic disease. We estimate the following logit model of disease prevalence using all the years of the NHIS data between 1990 and 1996 inclusive:

$$(1) \qquad P\left[d_i = 1 | age_i, year_i\right] = \frac{1}{1+\exp\left(g_1\left(age_i;\beta_1\right)+g_2\left(year_i;\beta_2\right)\right)}$$

The g functions allow the presence of disease to flexibly vary with the year of observation and the age-cohort of the respondent. Age-cohort enters the model through g_1, which is specified using an overlap polynomial:

$$(2) \qquad g_1\left(age_i\right) = \sum_{j=0}^{K}\left(\Phi\left(\frac{age_i - k_{j+1}}{\sigma_1}\right) - \Phi\left(\frac{age_i - k_j}{\sigma_1}\right)\right) p_j\left(age_i;\beta_{1j}\right)$$

where $p_j\left(age_i;\beta_{1j}\right)$ $j = 0 \ldots K+1$ are all n^{th}-order polynomial in age_i. The knots are $k_0 \ldots k_{K+1}$, and σ_1 is a smoothing parameter, which in addition to n, are all fixed before estimation. We use first degree polynomials. Though we experimented with higher order polynomials, we find that they add to the costs of computation with no change in the final projections.

The properties of the overlap polynomial can best be appreciated when the smoothing parameters approach zero. When this is the case, $\Phi(.)$ reduces to an indicator function equal to zero if $age < k_j$ and one if $age \geq k_j$. Thus the first term of the sum, $\left(\Phi\left(\frac{age_i - k_1}{\sigma_1}\right) - \Phi\left(\frac{age_i - k_0}{\sigma_1}\right)\right) p_0$,

equals p_0 when $k_0 < age \leq k_1$, and zero otherwise. Thus between k_0 and k_1, the prevalence rate is given by p_0, which in turn depends on the parameters $\beta_{1,0}$. Similarly, between k_1 and k_2 the prevalence rate is determined by p_1, between k_2 and k_3 it is determined by p_2, and so on. Allowing positive values of the smoothing parameters eliminates the sharp discontinuity of the growth rates at the knots. In fact, one advantage of this overlap polynomial over traditional splines is that the function and all its derivatives are automatically continuous at the knots without imposing any parameter restrictions.[15]

In addition to an overlap polynomial for age, we also include another overlap polynomial, g_2, for year to flexibly allow for changes in the age-prevalence relationship over time. Here, the knots are $m_j, j = 0 \ldots M$, the smoothing constant is σ_2, and q_j are the polynomials. As before experimentation led us to use first order polynomials in year.[16]

$$(3) \qquad g_2\left(year_i\right) = \sum_{j=0}^{M}\left(\Phi\left(\frac{year_i - m_{j+1}}{\sigma_2}\right) - \Phi\left(\frac{year_i - m_j}{\sigma_2}\right)\right) q_j\left(year_i; \beta_{2j}\right)$$

While (1) does not include any covariate information regarding i, such information can readily be incorporated into the analysis by replacing (1) with the following:

$$(1') \qquad P\left[d_i = 1 | age_i, year_i, X_i\right] = \frac{1}{1 + \exp\left(\left(g_1\left(age_i; \beta_1\right) + g_2\left(year_i; \beta_2\right)\right) X_i\right)}$$

This framework can also be adapted to allow for interactions between age and year effects:

$$(1'') \qquad P\left[d_i = 1 \mid age_i, year_i, X_i\right] = \frac{1}{1 + \exp\left(\left(g_1 + g_2 + g_1 * g_2\right) X_i\right)}$$

The object of the maximum likelihood logit estimation is to obtain consistent estimates for β_1 and β_2—$\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. In this version of our estimates, for the sake of simplicity we use equation (1) rather than (1') or (1''). In future drafts, we will generalize our estimates to account for more interactions.

Using these estimates, it is easy to generate age-prevalence profiles representative for any particular year. Let $\rho_{t,a}$ be the disease prevalence among a-year olds in year t. Then,

$$(4) \qquad \rho_{t,a} = \frac{1}{N}\sum_i P\left[d_i = 1 \mid age_i = a, year_i = t; \hat{\beta}_1, \hat{\beta}_2\right]$$

In the next section, we combine these estimates of disease prevalence with information on population and cause-specific death rates to derive yearly age-incidence curves.

[15] After some experimentation, we choose $k_0 = -\infty$, $k_1 = 25$, $k_2 = 35$, $k_3 = 45$, $k_4 = 55$, $k_5 = 65$, $k_6 = 75$, $k_7 = \infty$, and $\sigma_1 = 25$.

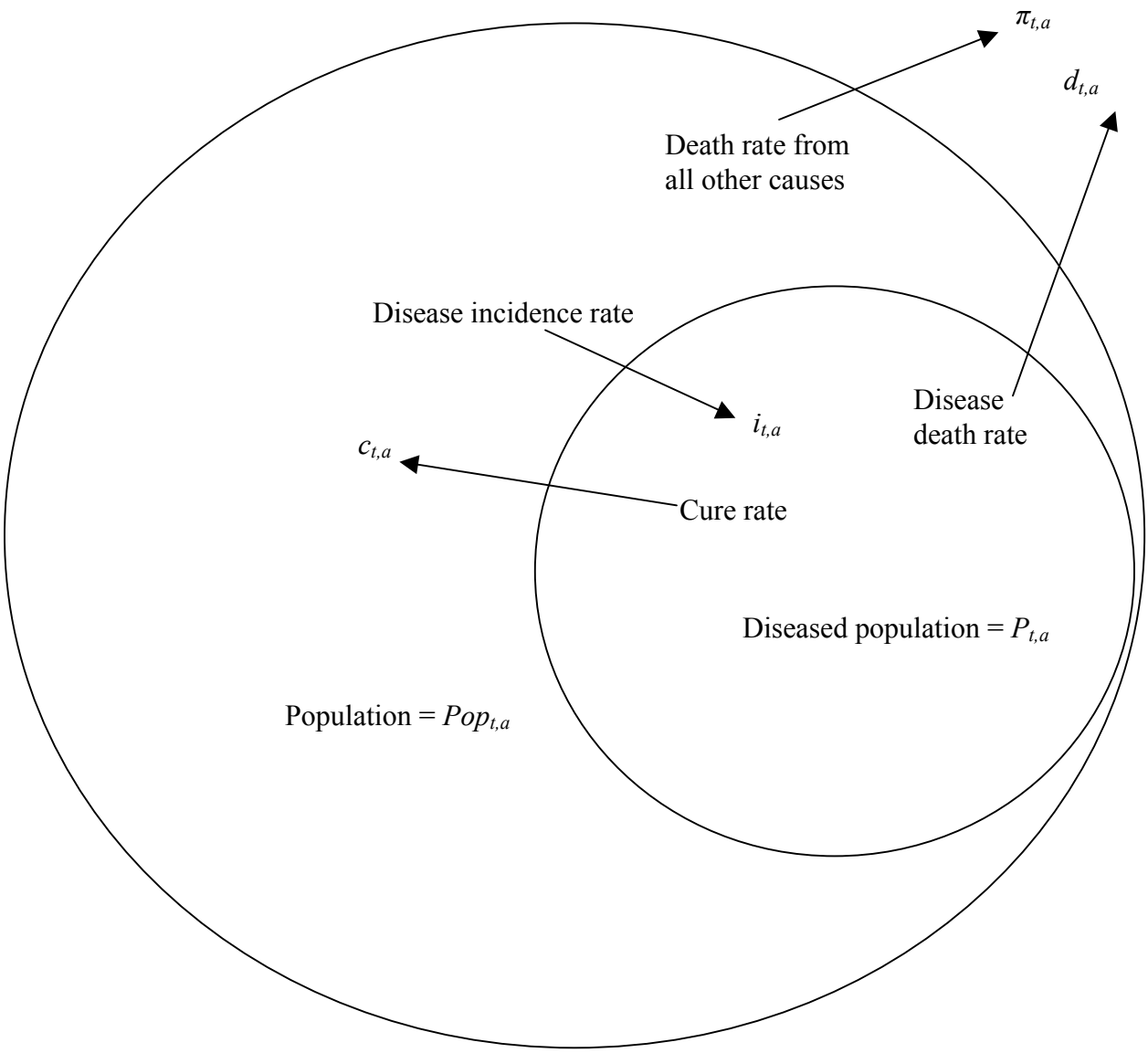
[16] After experimentation, we choose $m_0 = -\infty$, $m_1 = 91$, $m_2 = 93$, $m_3 = 95$, $m_4 = \infty$, and $\sigma_2 = 4$. In all our analyses, $year_i$ is entered as $year_i - 1900$.

Step 2: Estimating age-incidence profiles

The purpose of this section is to develop a simple model relating the prevalence of a disease in one period to its prevalence in the next period. We use a synthetic cohort approach to estimate an age-incidence profile for each disease from the prevalence estimates that we derive in the previous section. In the basic structure of our model, cohorts age from year to year and transition between health and disease. Because the NHIS is a nationally representative survey, a - 1 year old respondents in year t - 1 are presumably drawn from the same population universe as a year olds in year t, except aged by one year. Broadly speaking, we derive our estimate of age-specific incidence rates by comparing successive prevalence rates.

In our model, the population transitions between health and illness from year to year. Figure 1 illustrates all the possible transitions for one disease. At time t, the size of the population who are aged a is given by $Pop_{t,a}$. The size of the age a diseased population at time t is given by $P_{t,a} < Pop_{t,a}$. The $Pop_{t,a} - P_{t,a}$ patients without the disease condition, who are inside the large circle but outside the smaller circle, die from all other causes at a yearly rate given by $\pi_{t,a}$ and they develop the disease condition at the age- and year- specific incidence rate $i_{t,a}$. The $P_{t,a}$ patients inside the smaller circle die from the disease at a yearly rate given by $r_{t,a}$ and are cured at a rate given by $c_{t,a}$.

Figure 1: Population Transitions



Because there is no immigration of people into the population in Figure 1, the total size of the population aged $a + 1$ at time $t + 1$ will equal the size of the population aged a at time t, minus the people who die either from the disease condition or from other causes. The transition equation linking the population size of a given cohort from one year to the next is then given by:

$$(5) \qquad Pop_{t+1,a+1} = Pop_{t,a} - \left(Pop_{t,a} - P_{t,a}\right)\pi_{t,a} - P_{t,a}r_{t,a}$$

Dividing through by $Pop_{t,a}$, we write (5) in terms of the population age-specific prevalence of the disease, $\rho_{t,a}$, and the cohort growth rate $\gamma_{t,a}$:

$$(6) \qquad \gamma_{t,a} \equiv \frac{Pop_{t+1,a+1}}{Pop_{t,a}} = 1 - (1 - \rho_{t,a})\pi_{t,a} - \rho_{t,a} r_{t,a}$$

We are interested in how the number of chronically diseased people within a fixed cohort, who are age *a* at time *t*, changes as that cohort ages. This formula will allow us to relate incidence rates to changes in the prevalence rates that we calculate in Step 1, above. The number of people with chronic diseases in that cohort at $t + 1$ will equal all of those with the disease in the previous year save those who are cured or died, plus all the health people in the cohort who develop the disease. Therefore, the number of chronically ill within a fixed cohort evolves according to the following equation:

$$(7) \qquad P_{t+1,a+1} = P_{t,a} + (Pop_{t,a} - P_{t,a})i_{t,a} - P_{t,a} r_{t,a} - P_{t,a} c_{t,a}$$

Again, we divide through by $Pop_{t,a}$ to express (7) in terms of population prevalence rates:

$$(8) \qquad \gamma_{t,a} \rho_{t+1,a+1} = i_{t,a} + \rho_{t,a}(1 - i_{t,a} - r_{t,a} - c_{t,a})$$

Finally, we rearrange (8), solving for $i_{t,a}$ to write the age-incidence curve as a function of successive measurements of disease prevalence:

$$(9) \qquad i_{t,a} = \frac{\gamma_{t,a} \rho_{t+1,a+1} - (1 - r_{t,a} - c_{t,a})\rho_{t,a}}{1 - \rho_{t,a}}$$

We use information from equation (4) to generate estimates of disease prevalence rates, $\rho_{t+1,a+1}$ and $\rho_{t,a}$. We use information from Vital Statistics to generate information on disease specific death rates $r_{t,a}$ and on overall death rates $1 - \gamma_{t,a}$. Data on disease specific cure rates are nowhere available from any single consistent source. Consequently, in our calculations we assume that $c_{t,a} << r_{t,a}$. Because we are considering only chronic diseases with low cure rates, this assumption should not introduce too much error.[17]

Finally, taking linear combinations over *t* of $i_{t,a}$ generates age-incidence profiles that are representative for the period over which the linear combination is taken. Thus, in this framework it is easy to incorporate information about trends in disease or disability, at least to the extent that such trend evidence is present in the successive NHIS years that we use. Let the linear combination of age-incidence profile be i_a.

Step 3: Projecting the health status of future Medicare entering cohorts

Once the prevalence and incidence functions are calculated for each disease separately, we generate our projections for the health status of future entering cohorts of Medicare enrollees. The essential idea behind our projection is that for any given future year, we know how old the entering

[17] Indeed, for some conditions, this is true by definition. For example, the NHIS asks respondents whether a doctor has ever told them that they had a heart attack. There is no cure for heart disease if it is defined in this way; once a doctor tells a respondent that he has had a heart attack, the respondent should always respond yes to this question.

Medicare cohort is today. For example, writing in the year 2000, we know that the 65 year olds of 2001 are currently 64 years old; $\rho_{2000,64}$ gives the prevalence of chronic disease among this cohort, and i_{64} gives the predicted proportion of those without disease in that cohort who will develop the disease between ages 64 and 65 (among those who are disease free at 64). The disease prevalence for 65 year olds in 2001 is given by a direct application of equation (8):[18]

$$\rho_{2001,65} = \frac{1}{\gamma_{2000,64}}\left(i_{64} + \rho_{2000,64}\left(1 - i_{64} - r_{2000,64}\right)\right) \tag{10}$$

Recursive application of equation (8) to different cohorts in the NHIS data yields predictions regarding the prevalence of this disease condition for the entering cohort of any future year *y* (as long as the cohort is alive at the time of the latest NHIS). Thus, for our disease prevalence estimates for 65 year olds in 2002, we combine the disease prevalence numbers for 63 year olds in 2000, which we observe directly, with our incidence estimates:

$$\begin{aligned} \rho_{2001,64} &= \frac{1}{\gamma_{2000,63}}\left(i_{63} + \rho_{2000,63}\left(1 - i_{63} - r_{2000,63}\right)\right) \\ \rho_{2002,65} &= \frac{1}{\gamma_{2000,64}}\left(i_{64} + \rho_{2001,64}\left(1 - i_{64} - r_{2000,64}\right)\right) \end{aligned} \tag{11}$$

Similarly, our projections for the year 2003 start with the disease prevalence of 62 year olds in 2000, and recursively apply the incidence rates i_{62}, i_{63}, and i_{64} in three applications of equation (8). By starting with progressively younger cohorts, and applying the recursion formula more times, we generate projections of disease prevalence for each year between 2001 and 2030. In principle, this method could be used to project disease prevalence for any future year, as long as the group of people who will be 65 in that year are alive today.[19]

Step 4: Constructing population weight adjustments from prevalence projections

The three steps we have described up to now allow us to construct projections of future disease prevalence one disease at a time. While such univariate projections are independently interesting, they are insufficient for a project focused on predicting future Medicare expenditures. Elderly patients can have more than one chronic disease, and it is simply untrue that medical expenditures on a patient with two chronic diseases will equal the sum of expenditures on two patients, each with one chronic disease. In order to construct plausible estimates of total future

[18] For simplicity of exposition, the formula uses prevalence and incidence formula based upon the 2000 NHIS (which obviously has not yet been completed). The actual calculation for the 2001 entering cohort starts with prevalence estimates for 60 year olds in 1996, and use the predicted incidence formulae for 61, 62, 63, 64, and 65 year olds to generate the predicted 2001 prevalence. We do not use the 1997 and 1997 NHIS because the survey instrument changed in 1997, and it is not clear that the data after the change are directly comparable with the data after the change.

[19] As we mention in footnote 18, the discussion in the main text maintains the existence of the 2000 NHIS, which in reality has not been released at the time of this writing. Because the latest NHIS year we use is 1996, we start with disease prevalence rates of the 60 year olds from that year to construct our year 2001 projections. Similarly, we use 59 year olds from that year to construct our year 2002 projections, and so on.

Medicare expenditures, then, we need some estimate of the frequency with which chronic diseases jointly occur, as well as their frequency in isolation. This frequency distribution over the joint occurrence of chronic diseases can then easily be converted into predicted population weights for the incoming Medicare cohorts. Our purpose in this section is to describe the methodology we use to infer this joint frequency distribution.

As we mention above, in this document we focus on seven of the most expensive to treat chronic disease conditions that afflict the elderly, in addition to a measure of disability. The disease conditions include heart disease, hypertension, cerebrovascular disease, Alzheimer's disease, cancer, diabetes, and COPD. For the purpose of this section we define a set of index variables $d_i^* = \left\{d_i^{*1}, d_i^{*2}, ...d_i^{*8}\right\}$, where the superscript indexes over each of the eight disease and disability conditions, and *i* indexes over each member of some future Medicare incoming cohort. We redefine $d_i = \left\{d_i^1, d_i^2, ...d_i^8\right\}$ to be a set of indicator variables such that $d_i^j = 1\left(d_i^{*j} > 0\right) \forall j$, where $1(.)$ is the indicator function.[20] The analysis up to now allows us to estimate $\rho = \left\{\rho_{65}^1 = P\left[d^1 = 1\right], \rho_{65}^2 = P\left[d^2 = 1\right], ...\rho_{65}^8 = P\left[d^8 = 1\right]\right\}$, but does not allow us to infer $P\left[d^1, d^2, ...d^8\right]$.

The critical missing ingredient is information on the joint incidence of these seven conditions and of disability in the population of interest. In principle, there are $2^8 = 256$ different combinations of our chronic diseases that incoming Medicare cohorts can have. In practice, however, many cells are likely to be sparsely populated. For example there are, fortunately, few unfortunate folks in the cell where $d^j = 1 \ \forall j = 1...8$. The most densely populated cells tend to those where $d^j \prod_{j \neq k} \left(1 - d^j\right) = 1$ for some $j = 1 \ldots 8$; that is, those cells whose inhabitants have exactly one chronic condition. Also, some combinations of chronic conditions are quite important from an epidemiological and medical point of view, such as diabetes and heart disease, or hypertension and cerebrovascular disease.

Unfortunately, the NHIS does not allow us derive an estimate of this joint distribution without further assumptions. As we describe in the Data section above, the particular sampling scheme used by the NHIS never asks respondents about the presence or absence all disease conditions at the same time. The consequence of this data limitation is that using the NHIS we cannot derive the frequency of combined occurrence for some chronic conditions, including some important combinations (such as diabetes and heart disease).

To circumvent this difficulty, we augment our NHIS marginal prevalence estimates with information from Medicare recipients aged between 65 and 70 years. We examine recipients in the 65-70 year age range, because if we were to restrict the sample to just 65 year olds, our sample size in the MCBS database would be too small to allow an accurate estimation of the correlation across the prevalence of disease conditions. Let the correlation matrix in *d* measured in this Medicare population be denoted by Σ. Because the disease variables are each dichotomous variables, for any *j* we have that:

$$(12) \qquad Var\left(d^j\right) = \rho_{65}^j \left(1 - \rho_{65}^j\right).$$

[20] For the sake of notational simplicity in this section, we suppress the *t* subscript that reflects which future incoming Medicare cohort that *i* belongs to. For the same reason, we henceforth drop the *i* subscript as well.

Let $\Lambda = diag\left(\sqrt{Var\left(d^1\right)}, \sqrt{Var\left(d^2\right)}, ...\sqrt{Var\left(d^8\right)}\right)$. We assume that the joint distribution over *d* is generated by:

$$d^* \sim N\left(\Phi^{-1}(\rho)\Lambda, \Lambda\Sigma\Lambda\right) \tag{13}$$

Here, Φ^{-1} is the inverse of the standard normal cumulative density function applied element by element to the ρ vector. Both ρ and Λ are estimated from the NHIS data using the procedure we describe in sections 0, 0, and 0, whereas Σ is estimated from an entirely different data source, MCBS, but is representative of the same population as the NHIS. The main attraction of the normality assumption is that it allows a significant reduction in the number of parameters we need to characterize the distribution over *d*. Instead of 256 numbers, one for each possible combination of *d*, we represent the distribution with 8 numbers for the univariate prevalence estimates and the $\binom{8}{2} = 28$ numbers for the correlation matrix. We show below that the normality assumption on the joint distribution of d^* allows us to accurately recover information on first two moments of the *d* distribution.

Under assumption (13), we can reproduce the observed marginal prevalence rates as the mean of the *d* distribution. To show this, we note first that all the diagonal elements of Σ are equal to one, since it is a correlation matrix. With a slight abuse of matrix notation, this implies that

$$diag(\Lambda\Sigma\Lambda) = diag\left(\Lambda^2\right) = diag\left(Var\left(d^1\right), Var\left(d^2\right), ...Var\left(d^8\right)\right) \tag{14}$$

Given (13) and (14), we have for each disease condition *j* that:

$$d^{*j} \sim N\left(\Phi^{-1}\left(\rho_{65}^j\right)\sqrt{\rho_{65}^j\left(1-\rho_{65}^j\right)}, \rho_{65}^j\left(1-\rho_{65}^j\right)\right) \tag{15}$$

The population prevalence of disease *j* is given by:

$$P\left[d^j = 1\right] = P\left[d^{*j} > 0\right] = P\left[\frac{d^{*j} - \Phi^{-1}\left(\rho_{65}^j\right)\sqrt{\rho_{65}^j\left(1-\rho_{65}^j\right)}}{\sqrt{\rho_{65}^j\left(1-\rho_{65}^j\right)}} > -\Phi^{-1}\left(\rho_{65}^j\right)\right] \tag{16}$$

Therefore,

$$P\left[d^j = 1\right] = 1 - \Phi\left(-\Phi^{-1}\left(\rho_{65}^j\right)\right) = \Phi\left(\Phi^{-1}\left(\rho_{65}^j\right)\right) = \rho_{65}^j. \tag{17}$$

In addition to the marginal probabilities of the d distribution, (13) and (14) allow us to infer second order moments, which are simple functions of the first moment—see (12). In addition to these two moments, with the joint normality assumption over d^* we can now specify the joint probability distribution over *d*, $P\left[d^1, d^2, ...d^8\right]$, based upon known information:

$$P\left[d^1, d^2, ...d^8\right] = \int_{\frac{d^1-1}{d^1}}^{\frac{d^1}{1-d^1}} ... \int_{\frac{d^8-1}{d^8}}^{\frac{d^8}{1-d^8}} d\Phi^8\left(d^{1*}, d^{2*}, ...d^{8*}\right) \tag{18}$$

where $\Phi^8\left(d^{1*}, d^{2*}, ...d^{8*}\right)$ is the cumulative density function of the 8-variate normal distribution shown in (13).