An approach to forecasting health expenditures, with application to the U.S. Medicare system

Ronald Lee, Ph.D.
Professor of
Demography and Economics
University of California
2232 Piedmont Ave.
Berkeley, CA 94720
rlee@demog.berkeley.edu

Timothy Miller, Ph.D.
Researcher,
Center for the Economics and
Demography of Aging
University of California
2232 Piedmont Ave.
Berkeley, CA 94720
tmiller@demog.berkeley.edu

Research for this paper was funded by a grant from NIA, R37-AG11761
An approach to forecasting health expenditures, with application to the U.S. Medicare system

Objective. To quantify uncertainty in forecasts of health expenditures.

Study Design. Stochastic time series models are estimated for historical variations in fertility, mortality and health costs per capita in the US, and used to generate stochastic simulations of the growth of Medicare expenditures. Age specific health costs are modeled to depend on the number of years until death.

Data Sources/Study Setting. A simple accounting model is developed for forecasting health expenditures, using the U.S. Medicare systems as an example.

Principal Findings. Medicare expenditures are projected to rise from 2.2% of GDP to 11% of GDP by 2075. This increase is due in equal measure to increasing health costs and to population aging. We find that there is a 95% probability that Medicare costs in 2075 will fall between 5% and 26% of GDP, indicating a wide band of uncertainty. Although there is substantial uncertainty about future mortality decline, it contributed little to uncertainty about future Medicare costs, since lower mortality both raises the number of elderly, tending to raise costs, and is associated with improved health of the elderly, tending to reduce costs. Uncertainty about fertility, by contrast, leads to great uncertainty about the future size of the labour force, and therefore adds importantly to uncertainty about the health-share of GDP. In the shorter term, the major source of uncertainty is health costs per capita. The traditional projection method constructs high, medium and low scenarios to convey the uncertainty of long run forecasts. This approach
has many problems. For one, the indicated uncertainty in the earlier years of a projection is far too low, which is evident in comparing our 95% probability intervals to the high-low range of the CBO forecasts, even though after 60 years the intervals are similar.

Conclusions. History is a valuable guide for quantifying our uncertainty about future health expenditures. The probabilistic model we present has several advantages over the high-low scenario approach to forecasting. It indicates great uncertainty about future Medicare expenditures relative to GDP.

Key Words. Health expenditures, forecasting, demography, Medicare.
Introduction

Many forces will influence the future trajectory of health expenditures in the 21\textsuperscript{st} century: new medical technologies, policies regarding access to care, the costs of services, the health status of the population, population aging, and the growth of the economy. Some, like population ageing, change so slowly as to have little impact in the short-run and can be forecast with high certainty for a number of decades. Others, like health policies, or the costs of a given procedure, can change quickly, and therefore can have large impacts in the short-run as well as the long run, and are more difficult to forecast. A forecast of health expenditures will depend on the assumed future course of each of these factors. Needless to say, there is a large amount of uncertainty in the resulting forecasts.

In the first part of this paper, we discuss each of the main components and our general approach to forecasting them. We also discuss a strategy for making probabilistic forecasts, that systematically assess and communicate the degree of uncertainty in these forecasts. We believe these methods are likely to be applicable in many countries, despite differences in health care systems. In the second half of the paper, we discuss the results of applying these procedures and approaches to the specific case of the U.S. Medicare system.

Many people are deeply sceptical of long-run forecasts. Indeed, long-run forecasts of the Medicare system have been criticised as “an enterprise in comparative fantasy” by the former director of the agency responsible for issuing them (cited in White, 1999). In our view, however, if a forecast provides a good estimate of the uncertainty surrounding it, then users can decide for themselves whether longer-term
forecasts provide information that is of use to them. In fact, such an estimate of uncertainty can help us evaluate the degree to which any particular forecast scenario is “fantastic.”

A simple accounting model

In a simple accounting model, annual health expenditures can be expressed as the sum of the number of people in a given health status times the annual costs associated with that status. The number of people in a given health status depends in turn on the age-sex disaggregated population and the similarly disaggregated health status distribution. The annual costs will depend on the treatments made available to people in each health status category, and the cost of each treatment. One might further disaggregate the treatments made available into the technological progress which results in the development of new treatments and the policy decisions about which treatments to cover under health insurance. Each of these four components must be forecasted, implicitly or explicitly. Typically, policy makers will be interested not so much in the projected expenditures per se, as in their size relative to the entire economy, or GDP. This may be achieved by forecasting the difference between the growth rate of costs per beneficiary or per capita, and the growth rate of per capita income. This approach also has the advantage of filtering out the uncertainty that arises from uncertainty about the rate of growth of productivity in the future. In the following sections, we briefly discuss each component of our forecast: health status, treatments, cost per treatment, and population.
**Health status**

The need for health care depends on the health status of the population. While one could start with detailed measures of health status by age for the current population, it is far from clear how such measures should be projected far into the future. The most prominent approach has been developed by Manton and his colleagues. In a series of articles, they construct structural models for health status as a function of demographic inputs, life style behaviours, and risk factors. In order to generate forecasts from these models, one needs first to develop forecasts of the life style behaviours or risk factors, which is a challenging problem. Even then, the highly non-linear structure of the models might lead to forecast instability.

Other approaches have focused on forecasts for specific diseases such as End-Stage Renal Disease (Health Care Financing Administration). Singer and Manton have also suggested the use of more general age-specific disability rates based on ADLs and IADLs as measures of the health status of the population. Using the long term care survey, they find a rate of decline for disability by age, and project it into the future. Their forecast of declining disability rates implies a substantial reduction in health care demand among the elderly – enough to offset the effects of population aging on U.S. Medicare costs over the next 75 years (Singer and Manton, 1998).

A much simpler approach neither uses nor provides information about underlying causes or kinds of ill health. It notes that health care expenditures are strongly associated with age. In this approach, it is implicitly assumed that health status is fixed in relation to age and sex, so that forecasts of population by age and sex can readily be translated into
forecasts of health status. This is roughly the method used by the U.S. Health Care Financing Administration (HCFA) for its long-run forecasts of Medicare reimbursements for hospital stays. Unfortunately, the assumption of a fixed age-schedule of health status is improbable over a long forecast horizon.

Some analysts (Lubitz and Prihoda, 1984; Fuchs, 1984) have noted that the association of health costs with age derives from a more fundamental underlying relationship between age and the fraction of people who are near to death, at least for the older population. Of the people alive at any given age, some fraction will survive for another year, and the remaining fraction will die within the year. These analysts have shown that at least over age 65, there is little variation in health costs with age for those who survive. The strong association of health costs with age is due to the strong association of age with the proportion near death. This observation then suggests that instead of holding fixed the relation between health and age, one instead hold fixed the relation between health and time until death. As time passes, life expectancy increases in the population and mortality falls, reducing the proportion of the population near death at any given age, associated with an improvement in the average health status at any given age.

Our approach in this paper derives from these empirical studies. Figure 1 shows the association between per capita US Medicare expenditures around 1990 by age and by time until death. These data reflect mainly expenditures for hospital stays and physician visits for the elderly. We must be careful in generalizing from this experience since, since the same relationship may not exist at younger ages, nor for other types of health
expenditures such as long-term care and prescription drugs. Here we will apply this approach only to forecasting Medicare expenditures, leaving open the question of whether this method might be more generally applicable.

The logic of this approach is that health care usage depends on health care needs, which in turn depend on health status. Those who are fewer years from death are presumably less healthy, so time until death can be used as a proxy for health status. For those who die from accidents or violent deaths, this proxy will be useless, but for most people it may be quite good. Indeed, historical data for the US Medicare program over the past two decades indicates that the schedule has been very stable (see Miller, 2001). This gives us some confidence that its shape (as opposed to its level) will persist in the future. Miller suggests that the remaining variation in health costs by age, after controlling for time until death, may not be stable, since it may reflect either current and mutable medical views and practices about providing health care to seriously ill people at very advanced ages or current levels of frailty among the elderly.

Any standard population projection together with the assumptions on which it is based, can in principle provide the distribution of the projected population by the projected time until death, although this information is never actually provided in published tabulations. From this distribution, one can then infer the evolution of the demographic component of health care costs, assuming that the shape of the relation of these costs to time-until-death (henceforth, the time-until-death schedule) does not change over time, although its level may change. Similarly, this kind of calculation can be used to standardize health care costs for the changing time until death distribution in
the past, thereby constructing an adjusted cost series purged of demographic influences. This adjusted series can then be forecasted by some means, and the forecasted series can then be used to multiply or shift the schedule of cost by time until death.

Neat as this argument seems, we should not be too happy with it. After all, health status is in most cases the driving force behind mortality, rather than the reverse. It would be more natural to forecast health status by age, and then to use such forecasts to derive the implied trajectory of mortality. Our forecast of mortality is only an extrapolation, with no bio-medical content, so the same could be said of the health distribution implied by our mortality projections. In defence of this approach, we emphasize that our forecasting method for mortality is fairly well established and accepted, has been tested on historical data, and performs well.

If it is problematic to derive implicit forecasts of health status from explicit forecasts of mortality, as if mortality drove health rather than the reverse, then it is equally questionable to forecast mortality as a basis for forecasting health care costs, as if health care expenditures and the services they buy, have no influence on mortality. Looked at in this way, our approach appears frankly perverse. Nonetheless, we believe it makes sense. First, there is surprisingly little evidence that health care expenditures influence the trajectory of mortality. If there appeared to be a strong relationship, then we would need to alter our approach. Second, most countries have long time series for mortality, often stretching back 100 years or more. Comparable historical data do not exist for any country, for health status or for health care services. Third, mortality has exhibited a remarkably stable and linear trend over the past century for many industrial
countries, at least when measured by the k factor of the Lee-Carter method (Lee and Carter, 1992; Lee and Miller, 2000; Tuljapurkar et al, 2000). Such trends are sustained despite variations in governmental health policy, the development of new medical technologies such as antibiotics, and the decline of infectious disease. For these reasons, we believe it is reasonable to forecast mortality independently, which we discuss below.

Demography

Population projections are available for virtually every country in the world, either from a national statistical agency or lacking that, from the United Nations. We will briefly discuss some issues that arise, since future population aging is expected to be an important force causing rising health care expenditures.

First consider mortality. We believe that official statistical agencies have systematically under-projected mortality declines and gains in life expectancy. In Lee and Miller (2000) we analysed all past mortality projections by the Office of the Actuary of the US Social Security Administration. We found that in every decade except the 1980s, the Actuary under projected mortality decline. Giving each decade’s forecasts an equal weight, the mean predicted life expectancy at birth minus the actual was -.25 years after 10 years; -1.2 after 20 years, and –2.7 years after 30 or 40 years. A recent study by the US National Academy of Sciences (2000) confirmed an earlier study by Keilman (1997) in finding through ex post analysis of projection performance that official population projections and those of the UN consistently underestimated the size of the future elderly population, particularly at the higher ages, due to under prediction of mortality decline.
This tendency may well be continuing. A recent paper by Tuljapurkar et al (2000), using the Lee-Carter (1992) forecasting method we will describe below, forecasted mortality for the G7 countries out to 2050, and compared results to official forecasts. Tuljapurkar et al forecasted life expectancy at birth that was 1.5 to 4 years or more above the official forecasts by 2050. The analysis of mortality trends from age 80 to 100 by Kannisto et al (1994?) indicates more rapid decline at these advanced ages than is projected by official agencies in OECD countries.

In the forecasts we present in the second half of the paper, we use the Lee-Carter method (Lee and Carter, 1992; Lee, 2000; Lee and Miller, 2000). This extrapolative method uses statistical time series analysis applied to long historical time series of age specific death rates to derive an index of the intensity of mortality which can then be forecast, and the implied age specific rates and life expectancies can then be calculated. The method produces a probability distribution for each rate and for life expectancy itself for each future year, based on the variability and fitting errors observed in the past. Extensive historical testing of this model on the historical US data, and more limited testing on historical data for Japan, Sweden, France and Canada, confirms that it would have produced accurate forecasts in the past, relative to the estimated probability bounds.

The Social Security Administration forecasts life expectancy in the US to reach 83.5 in 2075 in their middle scenario, with low and high scenario forecasts of 79.5 and 87.1. The Social Security forecasts are also used by HCFA and by CBO. Our forecasts show significantly more rapid decline in mortality than these official forecasts based on
expert opinion. We forecast sexes combined life expectancy in 2075 of 86 years, with a
95% probability interval of 82 to 90.

Now consider fertility. The record of the US Census Bureau in forecasting
fertility in the past has not been good (see Figure 3 in Lee, 1999) but it is doubtful that
anyone else could have done better. We have not found statistical time series analysis
useful for generating a mean forecast for fertility in the US (see Lee, 1993). In the results
reported later, we have estimated the variance and autocovariance structure of
“innovations” to the fertility process from time series analysis. In our forecasts, we have
constrained the long run mean to equal a pre-specified trajectory, and superimposed the
error process derived from the time series analysis. This approach yields fertility
forecasts with means corresponding to the pre-specified trajectory, but with uncertainty
about that mean. For the mean trajectory, after careful consideration, we chose the middle
projection of the Social Security Actuary, or a TFR of 1.95 births per woman, which we
believe reasonable.

The long-run mean TFR of 1.95 may seem too high, given that the average TFR
in Europe stands at 1.4 and is as low as 1.1 in some industrial nations. However,
transitional effects of delayed childbearing may be partly responsible for these low levels,
enough to depress rates by 10 to 20% below completed cohort fertility (see, for example,
Bongaarts and Feeney 1998 for a discussion of this effect). In the US, similar transitional
effects seem to have been partly responsible for the low TFRs (1.7) seen in the 1970s and
1980s. Surveys indicate that women in Europe, like women in the US, expect to have
about 2 children. It seems possible, therefore, although far from certain, that the
European situation reflects factors that, although quite persistent, are not permanent (Lee, 2000). We have no recommendations for forecasting fertility in OECD nations other than the US.

Immigration can also have an important influence on demographic change, including population aging, although its effect depends strongly on whether and how quickly the fertility of immigrants and their descendants converges to that of the native population. In our population projections we have assumed immigration at the rates projected by the Office of the Actuary. Implicit in the fertility assumption (see above) is the expectation that fertility of immigrants converges fairly quickly to that of natives, as has been the case in the past in the US.

Costs
Recent research on costs of treating heart attacks (Cutler et al, 1998) has found that medical price changes have risen at a rate close to that of the general price level, while a cost of living type measure, adjusted for quality changes, has actually risen about 1% per year more slowly than the general price level. However, technical progress in medicine constantly replaces older procedures with newer, more effective ones, which are typically more expensive. There is debate over the extent to which the new, more expensive procedures or drugs are really superior to the old, and it appears that policy could exert a powerful influence at this point by slowing the adoption of the new and costlier treatments.
We will not attempt to model and forecast this complicated process. Instead, we will forecast based on observed past trends in actual expenditures which combine the effects of changes in prices, and changes in the quantity and quality of services consumed. As discussed earlier, future per capita costs at each age depend on the health status of those at that age, and the cost disaggregated by health status. In our approach, health status is proxied by time until death (with or without disaggregating by age), and so the relevant cost measure is cost by time until death, rather than cost for treating specific medical conditions, or costs for specific procedures. However, Cutler and Sheiner (2000) found that after controlling for final year of life, disability conditions has an independent effect on health costs, suggesting that a multifactoral model of costs may be more appropriate. We retain the simple time-until-death approach, due to the difficulties in forecasting these disability conditions.

We assume that the schedule of costs by time until death has a fixed shape, and we further assume that the level of costs by time until death is determined by multiplying this fixed schedule by a shifting index of costs. Historical estimates of this index can be formed by standardizing the per capita costs in the past for changes in the distribution of time until death in the population. Then the index can be modelled and forecast using some method such as time series analysis.

Rather than forecasting health costs per capita themselves, it is common, at least in the US, to calculate and forecast the gap between the rate of increase in health costs and the rate of increase in per capita income or wages. If the growth rate of health costs per capita (or per beneficiary) were no greater than the growth rate of per capita income,
then expenditures on health care, as a proportion of GDP, would not increase or would increase solely due to changing demographic composition, depending on exactly how the health cost index is defined. Increases at this rate would be consistent with constant or modestly increasing tax costs.

The US Medicare system was instituted in 1965, and after erratic variation during the start up period, a useable series of growth rates is available since 1970 or so through 2000. These differential rates of growth between per-beneficiary health costs adjusted for health status and per-capita GDP, are plotted in Figure 2. It can be seen that the growth rates fluctuate considerably from year to year, which in good part reflects changes in Medicare policy. These year to year fluctuations may not be representative of the longer run uncertainties arising from policy and from other sources. Furthermore, we are interested in modelling the underlying trend in health costs due to factors relating to changing technology or medical practice and not those introduced by policy changes in the particular health program. Also, it provides a long time series and is available for many countries. For these reasons, the latest Medicare Trustees Technical Report suggests that more general data for national per-capita health expenditures be used rather than the specific data on Medicare or any other program. This growth rate series is also plotted in Figure 2. Its mean is somewhat lower, and its variability is also somewhat less, compared to the Medicare index.

In order to forecast this index, we apply time-series methods to the historical series. As with fertility, we think that the historical series can provide useful information about the variability from year to year in the growth rate of costs, but not necessarily
about the long run mean rate of increase. For the mean, we think it is important to take into account the views of experts who have made judgmental forecasts reflecting detailed quantitative research into the forces driving cost increases. We take our mean forecast trajectory from the latest forecasts by the Congressional Budget Office (issued in October, 2000), which were based on the suggestions of the Medicare Trustees Technical Advisory Panel, suggestions which are now incorporated in their official report. One might well say that we have out-sourced the single most important component of our forecast. This is true, and we can only claim some modest value-added on our part, by developing the mechanics of the time-until-death framework, and by putting the forecast in a stochastic context.

We forecast costs using a constrained mean time-series model so that the average differential over all simulations mimics the long-run CBO central trajectory. This trajectory starts at the current gap, and trends downward until it reaches a gap of 1.1% over per capita income growth in 2025. Our estimated variance structure for the errors in the fitted time series model is then assumed to hold for the forecast period as well. Based on the autocorrelation and variance estimated from the time-series of US cost rates, we forecast a 95% probability interval for the average differential growth over the period of 1999 to 2075 of 0.5 to 2.1%. Figure 3 presents the 95% probability interval for the annual differential (solid lines) and the more narrow 95% probability interval for the cumulative average differential. If we ignore the effects of population aging and changing health status in the population, we can forecast a 95% probability interval for
US health expenditures as a percentage of GDP in 2075 as 20% to 70% with a median value of 36% in comparison to its 1999 level of 14%.

Policy

For long run forecasts, the present day details of program structure should not be emphasised; instead forecasts should be based on more fundamental forces influencing expenditures on health care. As for projecting policy changes that may occur in the future, the utility would depend on the purposes for which the forecasts are to be used. If they are to be used as a guide to health policy formation, then it would not be appropriate to make policy change endogenous or to assume that it occurs. The forecasts could then be used to show the long term financial consequences of maintaining current policy, or of modifying it in pre-specified ways. In this case the forecasts are conditional rather than unconditional, since they take the policy assumption as a given.

If instead the forecasts are to be used as a component of comprehensive government budgetary forecasts (for example, see Lee, Tuljapurkar, and Edwards, 1999), then it would indeed be useful to make assumptions about future policy changes, and in that sense to make unconditional forecasts, intended to represent the forecasters best estimate of what the future will begin, including policy changes. In attempting to model such future policy interventions, we might consider the limit imposed by tax burdens. We might look to international experience to determine possible limits. In particular, we would explore whether the recent slowdown and stabilization in government health care
experienced by numerous OECD nations is related to the overall tax burden in those countries.

**Dealing with the Uncertainty of Long Term Forecasts**

**The Scenario Approach**

As noted earlier, there is a great deal of uncertainty about long term forecasts, particularly over a horizon so long as 75 years. Traditionally, this uncertainty has been conveyed though the use of high cost and low cost scenarios. The scenario approach begins by choosing high, medium and low trajectories for each of the key components of the forecast. Experts may be consulted at this stage. For health costs, there would be trajectories for the demography (fertility, mortality, immigration), and then perhaps trajectories for health care costs by age and sex, combining implicit assumptions about health status, treatments offered, and costs of these treatments. The next step is to decide how to bundle together the high or low values of each trajectory. Typically, a high cost scenario would be composed of the low trajectories for fertility, mortality and immigration (resulting uniformly in an older population), with the high trajectory for health care costs by age. This would yield the highest ratio of costs to GDP, but not the highest absolute cost—for this, the high trajectories for fertility and immigration would be used. Similarly, a low cost scenario is constructed. Then the projections corresponding the trajectories in the two scenarios are taken to bracket the likely outcomes in the future. This approach has certain advantages. Expert opinion is readily incorporated. One can easily study the impact the assumptions about each component, by varying one trajectory
at a time. This may yield analytic insights. The approach is easily explained to policy
makers and to the public.

However, the scenario approach also has several serious drawbacks. First, what
does it mean that the high-low range is “taken to bracket the likely outcomes in the
future”? With what probability are they expected to contain the true outcome? Why are
they not twice as wide or half as wide? Only rarely is there an attempt to attach
probabilities to scenarios. (However, see Lutz, et al. 1996, 1998 for attempts at assigning
probability distributions to expert opinion).

Second, the scenario approach implicitly make extremely strong and highly
unlikely assumptions about the correlation of forecast errors. It necessarily assumes
perfect correlation of the component trajectories with each other and across time. For
example, in the “high cost” scenario, fertility is always lower than in the central forecast
(a negative forecasting error), every year (so there is perfect intertemporal correlation of
errors); per capita health costs are always higher than in the central forecast (a positive
forecasting error), every year (again, perfect intertemporal correlation of errors); and
negative forecasting errors for fertility are always perfectly associated with positive
forecasting errors in per capita costs, so there is a perfect negative correlation of
forecasting errors for these two variables. Similar statements can be made about any pair
of key variables in the forecast. These assumptions rule out possibilities such as that
fertility might start out low, and then become high, while mortality might do the opposite,
with health care costs per beneficiary also wandering around. Some resulting aggregate
cost forecasts could pass out of the high-low bounds of the scenarios.
Third, is the high-low range intended to contain annual values or is it calibrated to be no wider than necessary to contain long-run trends? It cannot satisfy both of these criteria at once, since the short-run forecast errors may be large from year to year, but tend to cancel over time. Error cancellation is ruled out by the implicit assumption of perfect intertemporal correlation of errors. In this sense, the scenario approach gives probabilistically inconsistent indications of uncertainty. Since year-to-year randomness may dominate the actual outcomes for the early years of the forecast period, the scenario based forecasts risk having the actual costs fall outside the high-low range soon after publication. In the case of scenario based population forecasts, sometimes the actual number of births has fallen outside the high-low range before the publication date of the forecast.

Fourth, Similarly, since there can be either no cancellation of errors across key variables or no reinforcement, depending on whether a perfect negative or positive correlation is assumed, the probability for an outcome to fall within the range for any single variable is inconsistent with the probability that the entire forecast will contain the true outcome.

Ex Post Analysis of Forecast Performance
In some cases, forecast performance and the probability coverage of the high-low range can be evaluated ex post, and used as a basis for assigning probabilities ex ante. Such studies are very useful. However, this kind of ex-post analysis of forecasts is generally restricted to institutional forecasts such as those issued by the United Nations or by the
U.S. Social Security Administration, which have compiled a long record and which generally use fairly consistent methods. Examples of such ex-post analysis of population forecasts include Keyfitz, 1981; Stoto, 1983; Keilman, 1997; and NAS, 2000; and of Social Security mortality forecasts, Lee and Miller, 2000.

**Probabilistic Forecasts Based on Constrained Time Series Analysis**

A growing literature has developed the use of modified time series methods for demographic forecasting (Alho, 1990; Alho and Spencer, 1985 and 1990; Bell, 1997; Cohen, 1986; Gomez de Leon, 1990; Lee and Carter, 1992; Lee and Tuljapurkar, 1994) and for forecasting the finances of the US Social Security system and other government budgets (Lee and Tuljapurkar, 1998 and in press; Lee, Tuljapurkar and Edwards, 1999). In this paper, we extend this approach to forecasting Medicare expenditures.

In the preceding sections, we have indicated how time series models may be fit to historical data series, including the data series of health care or Medicare expenditures in the US, expressed relative to the growth rate of per capita income or of the wage rate. In some cases, forecasts were generated from the fitted time series model; in others, the mean forecast was taken from subjective expert forecasts, while the forecast error structure was taken from the fitted time series model. Autocorrelations and cross-correlations estimated from the historical data series replace the assumption of perfect correlations used in the scenario approach. (In this application, cross correlations have been found to be zero, but this is not always the case. Autocorrelations are typically found to be quite high in all applications, including this one.)
The general forecast strategy for each key variable is to assume that age specific rates in any period can be represented as the product of a single time-varying index and a fixed, time-invariant age profile (or time until death profile) based on cross-sectional data (earnings, fertility, mortality, health costs by time till death, etc.). The single time-varying index is then forecast, and forecasts of age or time-until-death specific rates are derived by multiplying this time invariant schedule by the index. In this way, forecast models are assembled for fertility, mortality, and costs by time until death. Immigration is treated as deterministic, and productivity growth does not affect the outcome, which is expressed relative to GDP. Alternatively, it would be possible to forecast absolute health costs by forecasting productivity growth. Forecasts of trust fund balance would require forecasts of interest rates, as is done for applications to Social Security (see earlier references).

In principle, once the model is fully spelled out as outlined above, an analytic solution could be found for the probability distributions of all items of interest in the future. A solution of this sort is presented in Lee and Tuljapurkar (1994) for the population forecast. In practice, however, a far simpler and more flexible approach is to use stochastic simulation to generate a large number of sample paths, one thousand for example, each representing a different future. Then the distribution of the quantities of interest across these one thousand sample paths can be calculated, probability intervals can be found, and so on. For example, a 95% probability interval is one which contains the middle 95% of sample paths for some quantity in some year. This is the procedure we follow for Medicare. It is important to realize that although a probabilistic forecast for each key variable could be made and presented, no such forecasts are used in the forecast.
for Medicare costs as a share of GDP. Instead, the collection of stochastic models, one for each key variable, is used as the basis for the stochastic simulation.

As an example of a stochastic demographic forecast, Figure 4 shows forecasts of the old-age dependency ratio based on the Lee and Tuljapurkar (1994) approach. The Figure plots the first 5 of the 1,000 sample paths, which are shown as solid lines. Also shown in the figure are the median and 95% probability interval for the old-age dependency ratios based on 10,000 sample paths among which the five shown. Because we assume that health status and hence the demand for health are associated with years-before-death, we forecast the population by both age and years-before-death for each sample path.

Forecasts generated through time series analysis are essentially extrapolative. Extrapolation has the advantage of objectivity (although subjective judgements about depth of historical period and form of model are certainly required), and of a well established body of experience in a wide range of applications. However, also has the disadvantage that it cannot predict profound breaks with the past, even when there may be evidence that would lead us to expect such breaks in the future. Scenario forecasts are the appropriate means to present visions of the future such as a doubling of the human life span. However, such profound changes in a key component of the forecast sometimes would imply a world so different than our own as to render meaningless the basic premises underlying the forecast. We should also note that the historical record does contain many unexpected events such as the invention of antibiotics, the AIDS
epidemic, the baby boom, and therefore shocks of this magnitude are reflected in our
probabilistic forecasts of the future.

Results of the probabilistic forecast of the US Medicare system

The U.S. devoted 13.6% of its GDP to health care in 1998, more than any other OECD
nation (OECD, 2000). Most of this was funded privately (55%), either out of pocket, or
through personal or employer-provided health insurance. However, public financing has
been steadily increasing since about 1965 when only 25% of health expenditures were
publicly funded (HCFA, 1998). This expansion mainly reflects the founding and growth
of the federal Medicare insurance program. Medicare provides basic health coverage for
hospital stays, physician visits, and certain outpatient services to those 65 and older. It
is currently one of the largest federal programs, representing one tenth of the Federal
budget. As a result of recent legislation imposing some temporary spending freezes,
Medicare spending as a percent of GDP has declined for the last two years. But looming
on the horizon is a large demographic transformation ushered in by the retirement of the
baby boomers and the possibility of dramatic increases in their longevity. Our
probabilistic forecast is an attempt at a systematic treatment of these uncertainties in a
health expenditure forecast.

Using the techniques discussed in the previous section, we assemble probabilistic
population forecasts for the US from 1999 through 2075. These provide us with
population counts by age and time-until-death. Total earnings are simply the product of
these population age counts and the age schedule of earnings. Similarly, total Medicare
expenditures are the product of population time-until-death counts and the time-until-death schedule of costs. This schedule of costs is defined by a fixed shape whose level is set by a index of costs. As previously discussed, this index represents the cumulative average differential between the rate of increase in health costs and the rate of increase in per capita income. The forecast trajectory of this index is different for each of the 1,000 sample paths. However, the use of a constrained-mean time-series forecast insures that the average rate of increase over all sample paths is equal to the CBO’s baseline assumption (long run average differential of 1.1%). Having generated 1,000 possible Medicare futures, we can now analyse the probabilistic forecast of Medicare finances.

We compare our results to recent long-run forecasts issued by the Congressional Budget Office (CBO), the federal agency responsible for providing Congress with economic and budget analysis; and the Health Care Financing Administration (HCFA), the federal agency responsible for administering the Medicare program. We believe the assumptions underlying their forecasts are suspect (too slow a mortality decline, too unhealthy a population).

Figure 5 shows the median Lee-Miller forecasts of Medicare expenditures as a share of GDP based on our preferred model. We project Medicare to rise from 2.2% of GDP in 1999 to 11% in 2070. The intermediate scenario forecasts by the CBO (2000) and HCFA (2000 and 2001) are plotted in the same figure. In addition, we plot an alternative model in which a fixed age schedule of costs rather than time-until-death is used. The retirement of the baby boom is evident in the large increases in expenditures seen in all forecasts starting around 2010. The HCFA year 2000 forecast begins to level off around
2040 after the last of the baby boomers have retired, while the other forecasts continue to show rapid growth. This is because HCFA had assumed that growth in per-beneficiary costs would gradually slow to the rate of increase in per-capita GDP or average wages, so that only demographic change affects their longer run forecasts. However, their year 2001 forecast assumes a more rapid long-run rate of increase in which per-beneficiary costs are assumed to grow 1% faster than per-capita GDP. Lee-Miller and CBO both use a long-run rate of 1.1% growth in excess of productivity growth (or average wages).

In our preferred model, we use time-until-death as the measure of health status in the population. As seen in Figure 5, this substantially reduces the projected increase in expenditures in comparison to the model that uses a fixed age schedule of medical costs. Despite the fact that our faster rate of mortality decline leads us to project more elderly, our forecast is quite similar to that of the CBO (which uses a fixed age schedule of costs). Apparently, the cost increase due to increasing numbers of elderly is offset by the decreasing proportion near death at each age. In this way, the time-until-death assumption moderates the impact of mortality decline. The assumed relationship of health status and costs to time until death means that when there is an error in forecasting mortality, so that the true declines more rapidly, leading to more elderly people qualifying for Medicare, these people are also healthier than expected, and require less costly care.

Both CBO and HCFA issue multiple scenarios to represent forecast uncertainty. In their most recent project, CBO (2000) presented six alternative forecasts based on “optimistic” and “pessimistic” scenarios for 3 key variables: productivity growth, growth
in per-beneficiary health costs, and population aging. CBO’s optimistic and pessimistic scenarios for growth in health costs along with its baseline forecast are shown in Figure 6. Through 2010, all scenarios are assumed to follow the same trajectory. Beyond 2010, the scenarios slowly diverge as their annual growth rates in per-beneficiary health costs are assumed to move toward different long-run averages: 0%, 1.1%, and 2.1% for the optimistic, baseline, and pessimistic scenarios. Also plotted in the figure are the median and 95% probability interval from a Lee-Miller forecast which reflects only per-beneficiary health cost uncertainty (using a deterministic population forecast). The CBO range has a 35% probability coverage in 2025, doubling to 70% in 2040. This highlights a typical problem in scenario forecasts: they tend to understate the uncertainty in the early years of the forecast, since intrinsically they assume a gradually spreading fan.

The Lee-Miller forecast shows a large amount of uncertainty stemming from the future trajectory per-beneficiary costs. By 2040, the 95% probability bounds encompass tax burdens 1.9 to 4.8 times their 1999 level. Even over a short time horizon, we see in 2010, tax burdens from 0.98 to 1.5 the 1999 level are within the 95% confidence band.

The effect of demographic uncertainty is seen in Figure 7 which plots the median and 95% probability interval for a Lee-Miller forecast with only demographic uncertainty (per-beneficiary costs are assumed to follow the trajectory of the CBO’s baseline forecast). The CBO’s optimistic and pessimistic population scenarios are also plotted. Demographic uncertainty contributes little to our uncertainty about future Medicare expenditures over the medium-term. The use of time-until-death as a measure of health status filters out much of the uncertainty from mortality.
Having separately examined the impacts of uncertainty about future population and growth in per-beneficiary health costs, we now turn to examine the full probabilistic Lee-Miller forecast of Medicare expenditures. In Figure 8, we have plotted the median and 95% probability intervals for the full probabilistic forecast as well as those forecasts representing only uncertainty about growth in per-beneficiary costs or only demographic uncertainty. One striking feature is the wide band of uncertainty near the end of the forecast. In 2070, tax burdens which are 2.5 to 8.7 times their 1999 level are contained within the 95% probability bounds. We also note that the uncertainty is highly asymmetric, which is a typical outcome for stochastic forecasts which involve the multiplication of uncertainties, and which consequently tend to a log normal probability distribution. For Medicare, the 95% interval extends only 6% downward from the median, but the upside risk is much greater, extending upwards by 15%.

Further, we note that uncertainty from costs only in our stochastic forecast captures almost all the uncertainty until 2060 or so; thereafter, the demographic uncertainty begins to matter, particularly for the upside risk. By contrast, demographic uncertainty alone contributes only a small amount until 2040 or so, when uncertain fertility begins to have a growing effect on the labor force and therefore on the tax base and GDP. By 2075, demographic uncertainty alone would give nearly 60% of the width of the probability interval. Despite this significant contribution from demographic uncertainty, uncertain costs alone would give about 80% of the probability interval. Clearly, one would not do badly by treating a population forecast as deterministic, and costs by time until death as uncertain. This result does not come from the structure of the
forecast, but rather from the great variability in the rate of increase in the health costs per beneficiary in the past.

Conclusions

Long term forecasting is fraught with uncertainty, and this is particularly so for forecasts of health care costs. This paper has tried to make two main contributions. One is to treat the uncertainty of long term forecasts for health costs more systematically than in the past. The second is to draw on the apparently stable relation of health care costs to time until death as a means of adjusting the forecasts for the changing health status of the population at each age. In this way, the steadily declining mortality in the population is seen not only as leading to the survival of additional elderly for whom health care must be provided, but also as reflecting steadily improving health status of the elderly at each age, and therefore implying declining costs of health care by age, other things equal.

For our central forecast, we have drawn on the expert official forecasts, but instead of adjusting for the age composition of the population we have adjusted for its distribution by time until death. Our stochastic forecasts have assumed that these central forecasts are known without error as the true future “expected values”, and we have superimposed on them errors whose size and structure are extrapolated from the historical data as fit by time series models. For this reason, our probability intervals should be taken to describe a kind of lower bound for the uncertainty of the forecasts.

To summarize our key findings:

1. Medicare expenditures are currently 2.2% of GDP in the US. We project them to rise to 11% of GDP by 2075. This trajectory stays close to the CBO forecast until it
terminates in 2040, because our forecast of more rapid mortality decline has offsetting effects on projected health costs, as explained earlier.

2. We find that there is a 95% probability that Medicare costs in 2075 will fall between 5% and 26% of GDP. This is a very wide band of uncertainty, but we believe it is appropriately so. The upward uncertainty is more than twice as great as the downward uncertainty.

3. Relative to the approach based on health costs by age, we find that the time-until-death framework leads to lower forecasts of health costs. By 2075, the difference in costs as a share of GDP is 2.6 percentage points (13.7% versus 11.1%), and the difference is average annual growth rate over the 75 year period is about 0.3% per year. This is because increasing longevity postpones Medicare expenditures to later in life.

4. There is substantial uncertainty about the future course of mortality decline. An important feature of the time-until-death approach is that it in effect filters out much of the uncertainty arising from uncertain survival, since there are offsetting effects of mortality decline and the health improvement that causes it. There is no such filtering out of uncertainty about fertility and its effect on the size of the labour force.

5. Because of intrinsic features of the scenario approach, the implied level of uncertainty for short term projections is far too low relative to our estimates, with very low probability coverage; but after 60 or so years it should have roughly 95% probability coverage. The CBO forecast therefore risks being far wrong in its first 20 years or so, and being dismissed on that basis, even if it is sound for the long run.
References


Figure 1: Medicare Costs per Beneficiary by Age and Time-until-death

Source: Author’s calculation from data provided by James Lubitz, based on retrospective costs of 1990/91 decedants.
Figure 2: Annual growth in health costs per capita and in Medicare costs per beneficiary in excess of GDP per capita.
Figure 3: Annual growth in health costs per capita in excess of GDP per capita:
Median forecast with 95% probability interval.
Figure 4: Probabilistic forecasts of the old-age dependency ratio
Figure 5: Lee–Miller, CBO, and HCFA forecasts of Medicare as a share of GDP
Figure 6: Medicare as a share of GDP with Lee–Miller probability interval and CBO range reflecting only health cost uncertainty.
Figure 7: Medicare as a share of GDP with Lee–Miller probability interval and CBO range reflecting only demographic uncertainty.
Figure 8: Medicare as a share of GDP with Lee–Miller probability intervals: Full probabilistic model, only health cost uncertainty, and only demographic uncertainty.