Predicting Medicare Cost Growth*

John N. Friedman†
Harvard University

January 14, 2010

Abstract

Spending on Medicare has grown 2.4 percentage points faster than GDP over the last thirty years. If growth at these rates were to continue, Medicare alone would account for 31% of GDP by 2082, and healthcare expenditures in total would reach 99% of GDP. In order to generate more reasonable projections, researchers and policy analysts in the government have explored a number of different approaches. This paper reviews and evaluates these methods and suggests directions in which the literature must go in order to account for other un-modeled cost drivers.

*I would like to thank Amitabh Chandra, Mike Chernew, Dana Goldman, Marilyn Moon, Joe Newhouse, and John Skinner for helpful suggestions and valuable discussion. Abdul Ahad Tariq provided excellent research assistance.

†Harvard Kennedy School, email: jfriedman@post.harvard.edu
Healthcare spending is projected to be just over $2.6 trillion in 2010, accounting for 17.7% of GDP. This burden is split almost equally between private payers and the government, making the $912 billion price tag the largest single item of the federal budget and outpacing even the cost of Social Security. In turn, Medicare is the largest single item within government health spending, projected to be $500 billion for 2010, or about 13% of the federal budget.

The prodigious level of spending makes healthcare important, but it is the fearsome growth of healthcare costs which has put healthcare reform on the national agenda. Medicare spending has grown 2.4 percentage points faster than GDP over the past thirty years, more than tripling as a share of GDP since 1960. If costs continue to grow at current rates relative to GDP, then Medicare alone will account for 8% of GDP by 2030, 14% by 2050, and 31% by the end of the 75-year budget projection window. This would imply that Medicare would be more than 50% larger than the current size of the entire federal budget. While these numbers are absurdly large, a similar projection would forecast total healthcare spending to account for an impossible 99% of GDP by 2082.1

But if healthcare spending does not continue to rise at historical rates, then what will happen instead? How drastic must reform be to avert this looming fiscal crisis? The magnitude of the situation requires us to understand the possible paths forward for healthcare spending as never before. But the sheer speed of past cost growth that demands a policy intervention has simultaneously rendered moot the most direct and widely used forecast strategy of projecting forward past growth.

In order to solve this forecasting problem, both the academic literature and relevant government agencies have devoted considerable time, creative energy, and resources to developing models of Medicare cost growth. Within the academic literature, researchers have developed a number of strategies to forecast cost growth. Below I discuss three main approaches: extrapolation, micro-simulation, and computable general equilibrium

1Source: CBO (2007).
models. I review the mechanics of each approach, as well as their strengths and weaknesses. I pay particular attention to the assumptions on the dynamics of healthcare demand, healthcare supply, and technological growth as factors driving costs. Extrapolation and micro-simulation are fundamentally statistical or actuarial in nature, while general equilibrium models focus instead on the economic dynamics of healthcare cost growth. Each forecasting approach also has different strengths over the long- and short-run.

I then discuss the particular applications of one or more of these forecasting methods in important policy contexts. Each government agency combines these three methods in different ways when producing the numbers upon which policy is formally based. For instance, the Office of the Actuary combines extrapolation with a computable general equilibrium model, while the Congressional Budget Office relies on constrained extrapolation. Drawing on the more abstract methodological discussion of the literature, I discuss the practical implications, as well as strengths and weaknesses of each approach. Since each approach from the literature presents different strengths and weaknesses, policy-makers often modify the precise forecasting strategy depending on the particular context of the forecast. For instance, micro-simulation models provide an extremely flexible tool for analyzing the total impact of a specific reform or innovation, or a group of innovations, on the entire distribution of healthcare costs and outcomes in the population. But the numerous constancy assumptions of the method tend to break down over the long run. As a result, policy agencies (such as the Agency for Healthcare Research and Quality) that focus on short- or medium-run projections tend to use micro-simulation models. In other settings, such as the Medicare Trustees Report, that focus more on the long run, modelers rely less on micro-simulation.

I then develop a number of criticisms of the forecasting literature as a whole. First, current forecasting methods tend to obscure the real consequences of the underlying assumptions driving the models. For instance, many extrapolation models build in an exogenous slowdown in long-run cost growth to prevent the healthcare costs from consuming all of GDP. But these models do not focus on the cuts that would be necessary simply to achieve baseline growth, instead assuming (contrary to past evidence) that the long-run steady state growth is just lower than it has historically been.
A first improvement would be to identify exactly what these assumptions imply for variables such as access and quality. A second and subtler step would be to think not only about the magnitude of the implied cuts but also how those changes would propagate through the model to affect other aspects of healthcare.

Despite the methodological diversity in both academic and policy forecasting literatures, a number of key determinants and implications of future healthcare spending remain either absent or un-modeled. In this paper I discuss two such issues: technology and politics. Technological progress, though perhaps accounting for the largest share of past healthcare spending growth, remains un-modeled in most forecasts. Instead, most researchers implicitly or explicitly assume that growth continues at a constant rate throughout the forecast period. But technological growth often does not proceed as a purely secular trend; rather innovations are driven by scientists and engineers and businessmen who respond to conditions in the healthcare market. Both the rate of growth and also the nature of growth matter. For instance, research focused on lowering the costs of current technologies may have a different impact on cost growth than research that discovers new technologies.

Politics is almost entirely absent from current forecasting methods. But the fundamental question is often: What would happen in the absence of current reform? And future policy surely plays a crucial role in the proper answer. While some forecasts explicitly maintain current policy, and thus exclude the possibility of political change, politics are a crucial element in the more general question of predicting the shape of the future policy world in the absence of current change. Due to the difficulty in explicitly modeling such political behavior, formally incorporating the political sector into forecasting methods is probably impossible. However, most forecasting methods ignore the implications of the increased tax burden, and thus increased tax rates, that become necessary to support a growing healthcare sector. In addition, it will likely be political pressure from high tax rates, rather than consumer pressure from reduced non-healthcare spending, that eventually drives reform. Such politics of healthcare policy seem an important yet under-researched factor in healthcare cost growth forecasting.

The remainder of this paper proceeds as follows. Section 2 reviews the various methodological approaches developed in the literature and reflects on their advantages
and drawbacks. Section 3 then examines the particular projection methods used by various government agencies, such as the Congressional Budget Office and the Office of the Actuary in the Centers for Medicare and Medicaid Services. Section 4 comments more broadly on the elements missing from current forecasting methods, and Section 5 concludes.

APPROACHES TO MEDICARE FORECASTING

The literature on Medicare cost growth forecasting contains several distinct approaches. This section reviews these methods.

Extrapolation

Literature

The most direct approach to forecasting most economic phenomena is to use the past as a guide. Extrapolation does just this. By using past growth to predict future growth, one applies a best linear predictor model to the forecasting exercise.

In its simplest guise, extrapolation regresses a measure of healthcare cost on lagged instances of healthcare cost to predict growth. In the Medicare context, some papers analyze expenditures as a fraction of GDP, while others work with growth in expenditures. Fewer papers use total cost as the dependent variable for econometric reasons. Researchers also have estimated the time path of excess growth, which is the growth rate of Medicare above and beyond that of GDP.

There are a number of ways to include external projections into the model. For instance, the age-distribution of the population evolves in a highly regular (though non-linear) way. One way to control for such external factors is to add control variables to the right-hand side of the projection regression. By controlling directly for the number or fraction of seniors in the population, for instance, researchers essentially age-adjust their extrapolations for known deviations from current trends. Alternatively, the researcher can estimate cost per Medicare enrollee, separating enrollees by age, and then use the
external demographic projections to aggregate these per-enrollee costs back to total cost. In another example, Getzen (2000) includes both lagged inflation and growth in health employment as control variables in a simple regression. (In fact, Getzen notes that, since most short-run (i.e. one-year) growth in health expenditures comes from inflation and employment growth, one could do nearly as well predicting cost growth with only these two variables as with a long distributed lag structure).

Extrapolation, broadly defined, also includes a number of more complex time series techniques, primarily developed by macroeconomists to forecast economic growth. For instance, researchers looking to forecast several variables simultaneously can use a Vector Autoregression (VAR) model. The literature also includes a number of techniques to deal with the problems that arise when allowing a very long and flexible lag structure, or to remove seasonality or cyclicality from the data. While the details of these techniques are beyond the scope of this paper, it should be noted that researchers have ample tools here to apply to the problem of Medicare cost growth. For instance, Lee and Miller (2002) apply sophisticated time series methods to project Medicare cost growth. This paper and others also use the empirical fluctuations in past cost growth to estimate the uncertainty in cost projections, as recommended by Tuljapurkar and Boe (1998). For instance, Lee and Miller generate a 95% confidence interval with 4% to 18% cost growth of Medicare in 2050, an extremely wide range that underscores the tremendous uncertainty in these models.

Extrapolation is technically the same over almost any horizon, but the longer the time period of interest, the more one must push the model out of sample to complete the projection. This is the case whenever one uses extrapolation in any context. Getzen (2000) shows that the average mean absolute percentage error (MAPE) for total cost growth using extrapolation models grows from $\pm1\%$ over one year to $\pm3.7\%$ over 5 years. In the very long run, the range for estimates is very wide. There is a key difference here too between the medium- and long-run projections using extrapolation. Even when projecting over 10 to 20 years, historical data provides a number of twenty-year periods over which we can examine growth. Thus, we can redefine twenty years as one period, for the sake of our model, and estimate as we otherwise would. In the very long run, though, no such direct extrapolation is possible. Going back more than 40 years
runs beyond the current system of accounting for the national health expenditures; beyond 60 years medical insurance existed in a very different form, and the data are scarcely available, if at all.

In any extrapolation exercise, the key implicit assumption is that existing relationships remain constant over the projection window. If this is an acceptable assumption, the reduced form prediction functions as a sufficient statistic for all of the many complex economic dynamics that may be operating in the background. But since there are different dynamics over short- versus long- term, this assumption applies differently. For instance, in the short-run there is also no variability from population growth or technological development. Instead, economic conditions (and therefore demand for healthcare) may change, or the healthcare workforce may grow faster or more slowly. Over the long run though, population changes become very important, though predictable, while technological change is perhaps the key factor driving per capita cost growth. Assuming that technological growth remains constant over 75 years, though, is a very strong assumption. In the case of Medicare cost forecasting, as the projection window grows, the extrapolation method also runs into the fundamental issue of Medicare cost growth forecasting, which is that projected costs would grow so quickly that total spending on Medicare would grow to an infeasible share of GDP. This further reinforces the implausibility of the constant trends assumption embedded in long-run extrapolation.

In order to deal with the implausibly high level of spending projected far in the future, researchers often augment extrapolation by imposing external constraints on cost growth. For instance, Getzen (2007) begins with pure cost growth extrapolation but then includes a “Resistance Point” at which healthcare cost growth begins to slow and a “Limit Year” when growth finally reaches the rate of GDP growth (and stays there typically forever after). Modelers at the Congressional Budget Office (CBO) assume that consumption of non-healthcare goods never declines.

However, these long-run assumptions are quite ad hoc. Given that we cannot assume an unchanged rate of technological growth, nor can we project using extrapolation how the slowdown will occur so far out of sample, extrapolation is best suited for the short- to medium-run. One example of this in the literature is Bhattacharya et al. (2004) who
apply an extrapolation approach to estimate the impact of the increase in the chronically
disabled on Medicare cost growth through 2030.

In the literature, extrapolation is also used to estimate subsets of relevant parameters
instead of total cost growth. In many ways, extrapolation is the basic forecasting
approach when authors prefer to concentrate their prediction efforts elsewhere. Because
most papers take some version of this partial-extrapolation approach, the better question
is not whether a paper extrapolates, but where. No matter what methodology a paper
eventually uses to estimate Medicare cost growth, it likely relies on some extrapolation.
For instance, most papers do not separately model GDP growth. In order to control for
GDP growth, then, these analyses typically rely on extrapolation (whether its own or
someone else's). Borger et al. (2008) for instance take GDP growth assumptions from
outside the model (in particular, from the Social Security Trust Fund Intermediate
Assumptions).

Discussion

One strength of extrapolation is transparency. Since the projection essentially relies
on a set of regressions, it is relatively easy to determine what factors drive different
aspects of the projection. It is also easy then to shut down a channel by imposing
restrictions on the coefficients from the regressions. For instance, suppose one wanted to
understand the impact of the growth of the elderly population on Medicare cost growth.
One would simply vary the demographic projections in the $X_{t+1}$ control variables,
comparing the projected cost growth under a high- versus low-population growth
scenario. This transparency allows the researcher to easily break down the projections
and understand which factors are more or less important.

Extrapolation relies entirely on a statistical or actuarial rather than an economic
approach. This represents both its strength and weakness. Because the method includes
no economic assumptions, one need not rely on outside estimates of parameters such as
risk aversion or discount rates. One need not know the form of the utility function or rely
on particular assumptions about economic behavior, such as perfect optimization or
market clearance. Thus, extrapolation makes no direct assumptions about demand,
supply, or any other economic factor.
The assumptions implicit in extrapolation are instead about the reduced form patterns in the world— that past patterns and relationships will stay constant into the future. Over short periods of time, this is probably a reasonable assumption. But over longer time horizons, it becomes problematic. For instance, past trends for healthcare demand are most likely reasonable approximations for future demand growth over the next ten years, but perhaps less so thereafter.

The lack of explicit economic assumptions is also a weakness of extrapolation. By implicitly assuming that existing trends stay constant, this method avoids the key question in Medicare cost forecasting: what will eventually have to change to slow the growth? This same issue also applies to extrapolation that is augmented by external constraints on cost growth. One might agree with such a restriction, but by imposing it in reduced form, the method once again avoids specifying what economic factor adjusts to meet the constraint. For instance, will demand be rationed? Will technological growth slow, either independently or because of demand restrictions?

Extrapolation thus avoids the hardest long-run questions. This method is most appropriate in the short run, and its long-run numbers may be accurate. But without detailing what drives such numbers, correct or not, extrapolation leaves the researcher unsatisfied.

Micro-simulation Models

Literature

Micro-simulation is a highly flexible, essentially non-parametric, demographic projection technique. To understand this approach, consider dividing the population into small cells defined by demographic, economic, and health conditions. One group could be 44-year-old employed white married men earning $100K per year who have prostate cancer. Another group could be 45-year-old employed white divorced men earning $80K per year with a history of prostate cancer. The groups are defined such that they are mutually distinct and form an exhaustive catalog of the population. The modeler then estimates the probability of moving from one cell to any other cell during a given year using some dataset.
These transition probabilities are the heart of a micro-simulation model. Of course, many of these transition probabilities will be zero --- an individual cannot move from being 44 years-old one year to 43 the next --- but by estimating the transition model without many parametric restrictions, micro-simulation provides an extremely flexible and accurate model of the evolution of the population over time. Instead of estimating the full transition matrix, micro-simulation models often break down the transition from year to year into several steps. For instance, the model might first estimate the probabilities of demographic and socioeconomic transitions such as income, unemployment and divorce, as a function of past demographic and socioeconomic variables and past health status. A second stage would then estimate the transition matrix between various health status conditions, including death. For instance, for an individual in any given demographic and socioeconomic state, the modeler would estimate the probability of contracting any given health condition in the next year, the probability of recovering if sick, as well as various combinations of diseases. Breaking the model into these stages severely reduces the dimensionality of the estimation.

After estimating this transition matrix, the researcher can simulate the "path" of any given cohort moving forward in time with random draws to determine various transitions. Researchers usually simulate many different paths forward for each cohort to understand the distribution of possible future paths for each cohort. Micro-simulation can also be used to study an entire population over time. Since all individuals either age or die, the modeler must augment the simulated population with new cohorts. For instance, a researcher interested in modeling the Medicare-eligible population might begin with those aged 55 in a base year 0. One could follow these individuals through to year 1 using the estimated transition matrix, while adding a new cohort of 55-year-olds to complete the population.

After projecting health status, the next step for a micro-simulation model is to project health expenditures. These models typically use large regression models to predict expenditures for individuals in each potential health and demographic status group. Depending on the complexity of the model, one may break down the cost estimates by cost categories (i.e. hospital, outpatient care, prescriptions, etc...) and payer (Medicare A, Medicare B, private payer, etc...). Finally, one must project cost growth. Some models
inflate costs at prevailing rates while others assume constant prices over projection periods.

The researcher then specifies particular assumptions for technological development, or perhaps even a single innovation to be analyzed. Innovations in this model can change either or both of the transition probabilities for health statuses and the costs associated with each health status. Some policy changes can also be modeled in this way, such as legislation to change reimbursement rates in Medicare.

**RAND Future Elderly Model**

The largest and most commonly used micro-simulation model in the literature is the RAND Future Elderly Model (FEM), developed by Dana Goldman and coauthors. This model uses the Medicare Current Beneficiary Survey (MCBS) and the Health and Retirement Survey (HRS) to estimate both demographic and health transitions, as well as costs. The model begins with individuals at age 41, and so after each simulated year the researchers refresh the model with a new cohort of 41-year-olds.

Goldman et al. have applied the FEM model to study a wide variety of impacts on health care costs. Some studies have focused on the impacts of particular treatments such as telomerase inhibitors and a potential cancer vaccine. Other studies focus instead on demographic impacts such as increases in education or the rise in the Hispanic population in the United States. Finally, researchers have applied this model to policy changes such as the impact of the potential 1997 Balanced Budget Amendment (Goldman et al., 2004). Lakdawalla, Goldman, and Shang (2005) analyze the impact of obesity on Medicare spending using the FEM model. Joyce et al. (2005) considers the impact of improved health status among the elderly on Medicare costs and argues that the ten most prominent innovations are more likely to raise costs than to lower them.

**Discussion**

Micro-simulation models are well suited to study the effects of policy or research innovations that impact particular aspects of the healthcare system. Since prices are fixed (or relatively fixed, at best), this analysis is a partial equilibrium exercise. The constant price assumption therefore applies best for smaller interventions, while larger system-
wide interventions likely have significant general equilibrium effects that this method will not capture. Micro-simulation methods also offer the best model of the future demographic transition to a permanently older population in the US. Lee and Miller (2002) project that demographic changes account for half of the growth in Medicare spending (in absolute terms) through 2050. Furthermore, micro-simulation models account for heterogeneity in demographic transitions, as some sub-groups in the population may experience larger increases in life-expectancy than others (a fact demonstrated in CBO 2008). Understanding these forces is no small feat.

Although micro-simulation may not capture the impact of innovations on prices, the method has several key advantages in capturing not only the direct but also the indirect effects of prospective changes. To see this, suppose we are interested in the effect of new cancer treatments on elderly Americans. The direct impact will be an increase in the treatment rate for that particular cancer, but there are many indirect effects (both positive and negative). For instance, the incidence of cardiac arrest would increase as fewer individuals die from cancer. More generally, as the population survives more disease, we should find that prevalence of many diseases increases (Crimmins 2004). Extrapolation models could account for this indirect effect, but a researcher would have to control explicitly for it. Instead, micro-simulation models automatically correct for the offsetting effect on heart attacks. Similarly, one could model the simultaneous impact of many different healthcare reforms or technological innovations by simply changing multiple elements of the transition matrix at once. The model automatically adjusts for the many potential interaction effects.

Micro-simulation models also provide a natural approach to predicting not just the average but the entire distributional impact of a given intervention. Since the model follows each individual through her entire life, no additional work is needed to analyze any aspect of the entire population.

Micro-simulation models implicitly rely on extrapolation to estimate and predict many aspects of the healthcare market. The model adjusts for demand shifts due to demographic transitions or stipulated technological innovations, but not the broader general equilibrium effects on prices. Supply of healthcare resources is assumed to stay constant or progress at current rates within the study period. Similarly, the model does
not endogenously generate technological progress, and so any forecast generated through micro-simulation must make an exogenous assumption here. For instance, Cutler (2005), argues that the technological progress analyzed by Goldman et al (2004) focuses too exclusively on cost-increasing rather than cost-saving technological innovations. But since these assumptions are outside the model, micro-simulation per se does not tell us what kinds of technological progress to expect, only the consequences of a predetermined path of research improvement.

The model also implicitly assumes that the supply of healthcare remains unchanged (although Goldman et al. are working to incorporate supply-side responses into the RAND FEM model). As a result, long-term forecasting involves the same out-of-sample issues as extrapolation. (The RAND FEM model usually sticks to projections only over 25 years or so rather than the full 75-year budget window).

**Computable General Equilibrium Modeling**

**Literature**

Whereas micro-simulation and extrapolation are statistical rather than economic models, computable general equilibrium (CGE) models provide a diametric alternative for Medicare cost growth forecasting. This approach directly models the key economic relationships in healthcare cost growth: demand for healthcare, supply of healthcare, and technological growth. For instance, consumers optimally demand healthcare and non-healthcare goods to maximize utility. Firms providing medical care optimally hire workers based on demand. Prices in the medical market equate supply and demand. These models were developed largely in response to the actuarial nature of the models in use in the 1990s for cost projection that did not capture any of key incentives effects that may determine healthcare cost growth.

The assumptions in CGE models are economic parameters. For instance, one must specify the discount rate, the elasticity of substitution for demand of medical versus non-medical goods, the intertemporal elasticity of substitution, the elasticity of substitution in the healthcare production function between technology and market inputs, and the health production function technology share parameter. Most researchers calibrate their models
to match historical patterns of healthcare spending and healthcare employment and therefore assume values of these numbers, rather than estimate them independently; however, these assumptions rely on long literatures investigating the relevant parameters. These models can then vary the input parameters to generate robustness checks on the outputs.

Without constraints, CGE models often feature multiple equilibria that generate a wide variety of cost predictions. To resolve these indeterminacies, researchers impose constraints on the model to match other parameters from the literature. For instance, one might choose the cost projection path displaying an effective price elasticity of healthcare demand to match empirical estimates. Similarly, one might try to match the income elasticity of healthcare demand. Other options to constrain such models are to impose particular time patterns on the equilibrium cost projection (such as the GDP +1 assumption of the Medicare Trustees, see below).

Such constraints notwithstanding, the projections still provide a wide range of potential cost growth paths. These paths all tend to have a concave time pattern, relative to methods that rely on extrapolation, since the CGE model correctly captures the endogenous growth reduction as healthcare takes up a larger of income in the future.

Empirically detailed CGE models are relatively recent to the literature due to the computational requirements to solve the model. Warshawsky (1994,1999) presents a two-sector model, for instance, that breaks the economy into the health and non-health sectors. Hall and Jones (2007) also calibrate a general equilibrium model of health expenditure, though with the goal of projecting optimal rather than actual spending. The most recent CGE model is that of Borger, Rutherford and Won (2008). This paper focuses on the demand and supply factors involved in healthcare provisions; the paper assumes a constant rate of expansion of the state of medical knowledge, i.e. constant technological growth. The paper derives equilibrium conditions from the maximization decisions of consumers (who are workers and capital owners), medical firms and non-medical firms.

Discussion

CGE models present the only approach in the literature that takes seriously the
economic relationships involved in healthcare spending. If healthcare cost growth will slow one day as a result of natural economic factors, then this type of model is the only current option for predicting such a change. Furthermore, these economic relationships are almost surely the dominant dynamic factor in the very long run other than the demographics of population growth, which can be modeled separately. Therefore, an ideally specified CGE model fits the needs of long-run prediction well. Since these economic factors are less important in the short run, the approach is less suited to the projections there.

Nevertheless, CGE models have a number of weaknesses. First, the models tend to be very opaque. It can be extremely difficult to disentangle the precise link between input factors and cost growth projections, since the general equilibrium effects can operate in highly non-linear ways. Furthermore, these models tend to be quite sensitive to particular assumptions (often about quite uncertain parameters).

Another drawback of these models is the lack of heterogeneity in the population. In order to solve the model, the researcher must simplify away many if not all of the differences in preferences or institutions across the population. Many CGE models explicitly use a representative agent approach. Thus, these models are not well suited to answer questions that focus on subpopulations or relative healthcare developments. In fact, no one actually gets sick in the model, they just optimally consume more health care as a result of expanding budgets and technological possibilities. For instance, it would be difficult to trace the impact of the spread of HIV/AIDS on the population since it directly affects only a small fraction of the population.

A related drawback is that CGE models tend to abstract away from many of the institutional or economic concerns that make healthcare so empirically and conceptually challenging. For instance, Borger et al. (2008) assume that a representative agent has a constant elasticity of substitution (CES) utility function for medical and non-medical consumption. The model does not deal with the moral hazard from insurance, adverse selection and its effect on healthcare prices, or supply-side incentive issues (such as capitation versus cost-plus reimbursement). In other words, these models treat health care like a standard good, despite the fact that we have known since at least Arrow (1963) that it is not. These papers do calibrate the model to match the actual history of medical
spending, which of course does reflect all of these phenomena. Therefore the CGE models may begin at the right level of cost. But if there are any interactions of future demand or supply changes with any of the subtler economic issues involved in healthcare, these models will not account for it. Similarly, these models cannot allow for changes in the institutional environment, since the model assumes away all but the most basic features of the healthcare market.

POLICY APPLICATIONS

I have now discussed a range of methods developed in the literature for projecting Medicare cost growth. Each has its own strengths and weaknesses. Such forecasts are not made in a vacuum, though; real policy decisions with billions of dollars of funding at stake rely on one or more of these approaches to set policy. I now review the methods currently used by various official agencies to project future costs.

CMS / Office of the Actuary

By law, the Centers for Medicare and Medicaid Services (CMS) must produce 75-year projections of Medicare expenditures for the annual report of the Medicare Board of Trustees. The Office of the Actuary (which is responsible for such projections) splits the forecast into three periods: the short run (0-10 years), the medium run (11-24 years) and the long run (25-75 years).

For the first ten years, the Office of the Actuary projects each category of medical spending forward. These projections are essentially demographically adjusted extrapolations of past cost growth. For the final 51 years of the forecast (the long run), the projection occurs in two steps. First, the forecasters assume that the excess growth in healthcare costs will be 1% on average (an assumption denoted “GDP +1”). The forecasters then estimate the CGE model of Borger et al. (2008) to solve for the time path of this growth, constraining the solution to the model with the GDP +1 assumption. During this long-run period, all subcategories of healthcare are assumed to have the same
growth rate relative to GDP. The forecasters then fill in the medium term by linearly interpolating the evolution in growth rates between year 10 and year 25. Since growth rates differ across the various subcategories in year 10 but not year 25, the medium-term projections involve a convergence of growth rates relative to GDP. The Office of the Actuary takes as inputs to this model a number of assumptions on the future rates of GDP growth, the real wage differential, the real interest rate, total fertility rates, and death rates. By varying these numbers, the actuaries produce a high-cost and low-cost scenario, in addition to the headline number. However, all of these scenarios apply the same GDP+1 rule to average excess cost growth over the long-run period between 25 and 75 years into the future.

The GDP+1 assumption resulted from a recommendation made by the 2000 Medicare Technical Review Panel. Medicare costs have never grown slower than GDP+1 over any extended period, though regressions isolating permanent drivers of cost, such as technological growth, tended to produce estimates between 0.8 and 1.5 above GDP. Furthermore, the GDP+1 assumption remains consistent with continued growth in non-healthcare spending throughout the 75-year window, and is thus not obviously too large. The use of the Borger et al. (2008) CGE model (also known as the OACT CGE model) was recommended by the Medicare Technical Panel in 2004. However, it is important to reinforce that the CGE model does not help with the total cost growth forecast, rather it simply reallocates growth earlier in the time path (U.S Department of Health and Human Services, 2009).

Why did the Medicare Technical Panel settle on GDP+1? Given the low accuracy of long-run forecasting given current methods, the panel believed that the transparency of a GDP+X rule trumped the added realism of a more complex rule. The committee most likely also chose X in this rule to be a round number (or at least a multiple of 0.25%), and so was left with only a few choices. Choosing X = 1% seemed like the best such rule, according to members of the 2000 Technical Panel. Certainly the existing method has the advantage of transparency, as well as long-run consistency. Other methods might allow the long-run growth rate, an imprecisely estimated parameter at best, to bounce around from year to year in ways that would confuse changes in this projection with other policy changes.
There are a number of drawbacks to this approach. The model assumes that excess cost growth across the different areas of Medicare begin to converge in the 11th year out, sharply reversing the existing trend of divergence. Next year, though, this convergence process will not begin until year 12 (from today's perspective), which is of course inconsistent. Therefore NPV Medicare expenses for years 11-75 ratchet upwards from year to year as the actuaries factor in an additional year at current excess growth rates.

The approach of CMS also obscures the true costs of the assumptions necessary to limit projected growth. The model assumes that cost growth must slow from current rates to GDP+1 on average during the period between 10 and 25 years in the future. But the model does not specify how this slow-down would occur. Would the growth rate of demand for medical services slow as prices rose? Would the supply of doctors increase so as to exert a pressure on cost counter to the current growth rates? While these changes may accurately reflect the future realities if no policies change (see below), the automatic adjustment obscures the cost that would somewhere be borne in order to generate such savings. Depending on the source of this slowdown, we might consider the resulting slower growth a good thing or a bad thing. For instance, if supply of doctors expanded to decrease the rate of cost growth, that is more likely welfare enhancing than a fall-off in consumer demand growth due to increased prices or a fall in technological growth.

The obscuring of policy choices is worsened by the policy-neutral baseline used by CMS (and other non-partisan agencies). By law, these forecasts must assume current legislation is to hold. But current law is updated every year, for instance by raising Medicare physician reimbursement rates. The current law benchmark thus represents a hopelessly optimistic projection of actual law. Thus, the projections assume future unspecified cost savings on top of an optimistic baseline, giving doubly the impression that less must be done to actually bring Medicare cost growth in check. While CBO typically provides a scenario that includes realistic policy updates, the Office of the Actuaries does not (or at least does not publicize it like the CBO does).

**Congressional Budget Office**

The Congressional Budget Office (CBO) operates under a similar current law
environment as the Office of the Actuary, but the forecasting methodology differs greatly. CBO instead relies more heavily on a pure extrapolation of current growth rates in cost. As a result, CBO projects Medicare costs to grow much faster than does the Office of the Actuary.

CBO applies two brakes to the otherwise unchecked forecasted growth of Medicare. First, CBO assumes that unspecified policy changes would reduce Medicare cost growth by one-fourth of the reduction in non-Medicare growth. For instance, if spending growth rates for health care outside of Medicare (or Medicaid) fell from 2% to 1%, the growth rate for Medicare would fall to 1.75%. The only brake on cost growth in the CBO forecasts is the requirement that non-health consumption never decline. The projection hits this bound, and so by the end of the 75-year window all income growth is channeled into increased healthcare spending.

In the end, Medicare grows much faster in CBO's projections than in those from the Office of the Actuary, especially in the long-run window. Cost growth 75 years out still remains 1.1% above GDP growth. The implications on the flow cost of Medicare at the end of the window are substantial. For instance, in the base scenario presented by the Office of the Actuary, gross Medicare expenditures grow from about 4% of GDP today to 12% in 2082, while CBO projects growth to 17% of GDP, by that time period.

On balance, CBO's forecasts represent a less optimistic approach than the Office of the Actuary, but both approaches rely heavily on constrained extrapolation. As a result, many of the same critiques from above apply here as well. For instance, by imposing reduced form restrictions to limit growth, the model obscures the true costs of the status quo. To be fair, CBO does not deny this, and in fact it explicitly writes in the most recent detailed report on Medicare spending that under the scenario that CBO presents, the slowdown in excess cost growth would not be painless and would not occur simply through improved efficiencies given the current structure of the health sector (CBO, 2008. But writing such language in a methodological report is different than explicitly projecting either sharp healthcare rationing (under their baseline forecasts) or declining non-health spending (should the reforms needed to generate their numbers not occur).
The Agency for Healthcare Research and Quality (AHRQ) is not charged directly with long-term cost projection. Therefore the agency focuses instead on projecting the more detailed impact of various reforms over a shorter time horizon. In keeping with this mission, the agency maintains a micro-simulation model for this purpose, managed by the Agency for Health Care Policy and Research. The model is similar in concept to the RAND FEM model, but instead uses the Medical Expenditure Panel Survey (MEPS) data to estimate health expenditures conditional on health status. (MEPS is administered by AHRQ).

It should be noted that, unlike MCBS, MEPS provides spending information for all of healthcare in the United States rather than just Medicare spending. Unfortunately, the quality of the data falls off substantially due to presumed under-reporting. For instance, the Agency for Health Care Policy and Research must significantly re-weight the data (beyond the “nationally representative weighting” in the data) to match the aggregate NHE figures from CMS in the baseline case and get comparability of estimates. This divergence also presents interesting issues, beyond the scope of this paper, in the difference between Medicare cost growth and spending and the expenditure for all health care (e.g., Boccuti and Moon (2003)).

Such a micro-simulation approach of course does little on long-term healthcare growth rates, nor does it account for the economic implications of the various reforms studied. (In fact, the agency takes growth rate projections from the Office of the Actuary model described above). But the model does allow for studies of heterogeneous effects across the population, and thus provides the capacity for more subtle impact analyses. For instance, this approach would allow the agency to study the impact on healthcare costs of obesity, which some have estimated at $147 billion per year (Finkelstein et al. 2009), accounting for 27% of past spending increases (Thorpe et al. 2004). This is an excellent example of an agency using the best modeling tool for the specific job at hand, here a nuanced prediction of the distributional impact of a number of current or proposed health reforms.

Department of Veterans Affairs
The Department of Veterans Affairs (VA) provides complete coverage for nearly 13% of the population (according to the 2000 census). As a result, the VA maintains a micro-simulation model crafted explicitly towards projecting the healthcare costs of veterans. This model is known as the Enrollee Health Care Projection Model (EHCPM). Of all the policy alternatives, this is most similar to the RAND FEM model. It includes two of the three components of the RAND FEM model (health status projections and health expenditure projections), but instead of modeling incoming Medicare cohorts it models new cohorts of veterans.

Unfortunately, the precise algorithms within the VA model are considered proprietary, so it is difficult to know more closely how the methods used differ from those at AHRQ or RAND. RAND evaluated the VA projection model in Harris, Galasso and Eibner (2008) and concluded that the model provided accurate and appropriate short-run modeling of VA costs and potential reforms, but was less useful for long-run planning. This conclusion follows from the strengths and weaknesses of micro-simulation models discussed above.

Figure 1: This figure plots historical and projected Medicare spending, net of enrollee premiums, as a fraction of GDP. CMS projections are adjusted from Table III.A2 of the 2009 Medicare Trustees Report.
CBO projections are from Figure 1-2 of the June 2009 Long-Term Budget Outlook. Extrapolation of current growth rates are the author's calculations.

Summary

Figure 1 summarizes the major governmental cost forecasts discussed thus far in this Section. On it I plot the path of Medicare spending as a fraction of GDP from 1965 through 2083. The first 44 years of the sample are the actual fractions of GDP spent, according to the National Health Expenditure data. I then plot three distinct projections forward: extrapolation of past growth rates relative to GDP growth, the CBO projection, and the projection from the CMS Office of the Actuary. If current trends continue, Medicare spending (net of premiums) would exceed 30% by the end of the 75-year projection window. CBO and CMS each project a slowdown in growth even absent policy change, so that Medicare would roughly triple under the CMS projections and quintuple under the CBO projections. These three projections methods are actually quite close for the first 20 years, after which the differing assumption on long-term growth begin to bite.

UNMODELED FACTORS IN HEALTHCARE COST GROWTH

This paper has reviewed a number of methodologies for forecasting the cost growth of Medicare. These approaches have different strengths and weaknesses. Some approaches, such as CGE models, better capture the economic forces at play in determining healthcare cost growth. Others take a more reduced-form statistical approach. But across these various approaches, there are some important factors that are never fully addressed. This section discusses some of the most important unmodeled factors in forecasting Medicare cost growth.

Technology

Many papers regard technological change as the key driver of healthcare costs.
Newhouse (1992) identified technological growth as explaining one-half to two-thirds of healthcare cost growth. Okunade and Vasudeva (2002) find similar numbers to Newhouse after modeling more intricately the time series dynamics of different factors, while Smith, Newhouse and Freeland (2009) compare cost growth across a range of countries and conclude that technological development drives between one-quarter to one-half of worldwide cost growth, a somewhat smaller figure than in the US. Similarly, Cutler (2004) argues that technological improvement, while providing large net benefits, has driven most of cost growth in the past half century, while Cutler et al. (1998) show that the costs of treating heart attacks has fallen by roughly 1% annually as the even as the quality of treatment has improved.

Yet despite the almost universal agreement of the importance of technology, none of the modeling strategies outlined above, either in the academic literature or the policy world, make any attempt to actually forecast technological growth and its implications for Medicare costs beyond simple extrapolation. (Ironically, of all the economic forces that the OACT CGE model incorporates, technological cost is one of the exogenously driven factors). The RAND FEM model, as well as others, can model the impact of particular technological developments, or even a range of such developments at once. But these models make no attempt to model the economic feedback on the rate of technological growth.

What is more, the forecaster must not only predict the level of technological development but also its type. Weisbrod (1991) decomposes technological change into those innovations which cheapen existing procedures or products and those that develop new procedures or products. Interestingly, Chandra and Skinner (2009) argue that both can be either good or bad for cost growth, depending on the usage. New developments could increase cost growth if previously untreated patients may be saved or if already treated patients move to a more expensive (though of course hopefully more effective) course of treatment. More counterintuitively, innovations that cheapen existing technologies may also add to healthcare cost if the decline in price is offset by an increase in quantity, either because we are on an elastic part of the demand curve or because the improved procedure also offers fewer side-effects or restrictions. In particular, Chandra and Skinner argue that technologies that impose few side effects, and thus can be applied
to everyone (or even multiple times to a given patient, as in the case of medical imaging), accounts for much of the cost growth in healthcare. A third form of technological improvement to consider is administrative or information technology that may both increase the quality of care and lower the administrative burden. For instance, despite an ongoing multi-billion dollar push to digitize medical records, for instance, few models incorporate the costs or savings involved here.

This is undoubtedly a difficult area for prediction. But economic research has shown, in a variety of contexts, that research innovations respond to important economic aspects of the market. Many researchers have followed the macroeconomics literature and treated technological growth as a residual remaining after controlling for other factors (i.e. the Solow residual). But medical research is more concentrated than technology for the entire economy, and some researchers have made large strides in understanding the interplay between market demand, institutional conditions, and technological growth. For instance, Acemoğlu and Linn (2004) show that the development of pharmaceutical products responds to the underlying size of the market, using demographic shifts over time for identification. Similarly, Acemoğlu and Finkelstein (2008) demonstrate that the Medicare prospective payment system affected hospitals’ use of capital vs. labor-intensive technologies. Finkelstein (2004) shows how innovation of vaccines responds dynamically to market conditions. A number of papers have also specifically explored the profit increases and resulting increase in R&D spending by pharmaceutical firms following the passage of Medicare Part D in 2003 (Friedman (2009), Lakdawalla and Sood (2009), Blume-Kohout and Sood (2008)).

Another approach is to use expert “Delphi” panels to rate the likelihood of particular technological improvements in the future. Experts in medical subfields meet to discuss and then individually estimate the probabilities of specific technological improvements. By integrating over these estimates, one can implement a meta-distribution of future technology. This approach is very close to the “Gamma Discounting” approach of Weizman (2001) to estimate the appropriate discount rate for the social welfare function. Shekelle et al. (2005) take this approach in the subfields of cardiology, cancer, the biology of aging, and neurological disease. This analysis combines expert panels and a literature reviews with the RAND FEM model to estimate not only the likelihood but also
potential impact of a host of future technologies. Ex post, the predictions generated by these panels were fairly accurate over the past five years, for instance correctly predicting the spread of implantable defibrillators. Further into the future, though, the expert panels come to less of a consensus and so the predictions become far more diffuse. While not a panacea, this approach could be a valuable complement to the use of economic factors to predict technological growth, especially in the short- to medium-term.

Politics

A second important omitted factor for Medicare cost growth is the political process. Unlike technological growth, which models treat (however unsatisfactorily), politics is left entirely out of the picture. Furthermore, some of the forecasters (such as CBO and CMS) have an explicit mandate to evaluate the long-term consequences of current policy, which by definition excludes the possibility of future adjustment. However, even CBO assumes that un-modeled future policy changes will drive some of the reduction in cost growth rates.

Politics could enter into cost growth forecasting in a number of ways. Of course, potential future political changes do not resolve the fundamental inconsistency of current growth rates with long-term affordability. But the political reality suggests two alternative ways in which research and forecasting could help predict how and when the current system will finally change.

The most important political implication of current policy is the rising tax burden required to support the government share of healthcare expenditures. Currently, government spends about 6.2% of GDP, or 22.0% of all tax revenues taxes collected on Medicare and Medicaid. Under even the optimistic projections from the Medicare Trustees, this fraction would rise to 23% of GDP by 2082, more than the entire current federal budget. Such growth would require substantial increases in the tax burden. This is a related, though conceptually distinct brake on the growth of healthcare costs to that in other models. CBO and other forecasters relate the growth of healthcare costs to the demand for other services, and thus it is individuals' demand for other goods that eventually limits healthcare growth. Here, it is the political reality of a sharply increasing
tax burden that limits the growth. This political brake on the system may occur even while demand for healthcare is rising, since ever increasing marginal tax rates have an independent effect on both the economic and political situation. Unfortunately, most CGE models do not model the government sector, and so they do not account for the tax burden of rising healthcare costs.

There is some evidence of such political limitations of healthcare cost growth. Getzen (1992) shows that, in cross-country panel regressions, it is income rather than the fraction of the population that is elderly that explains the variation in healthcare spending. Chandra and Skinner (2009) point out that healthcare growth rates since 1985 are negatively correlated across countries with the level of the tax burden at the beginning of the period. Of course, such cross-sectional comparisons may suffer from substantial omitted variable bias. Marginal tax rates have also fallen sharply in the United States in the past 30 years, although the total tax burden has remained largely unchanged as the base for taxation has broadened. Relative to European countries, the United States is an outlier for having such a low tax burden. This might suggest that the United States has a ways to go before the tax burden pressures seen in Sweden and Denmark kick in to restrain health care spending growth. On the other hand, the particular political history of the United States may imply that these political pressures would appear instead much sooner if tax preferences forbid a return to the pre-Reagan-era top tax rates.

CONCLUSION

The problem of forecasting Medicare cost growth is an important yet difficult endeavor. Researchers in the academic literature have developed a number of techniques for forecasting, including extrapolation, micro-simulation, and computable general equilibrium models. The many government agencies tasked for projecting healthcare expenditures in different settings each choose from these broad methods in their particular cases. But regardless of the specific case in which these methods are applied, or even the details of the particular method used, one always returns to the same deep question that arises whenever we attempt to predict the future path of health care: what
will be different in the future? Since healthcare cannot continue to grow as it has, an answer to this question should lie at the heart of any long-run forecast. Forecasters should not only answer this question but the answer clear so that others can judge the model on its merits.

I end on a positive note. It is impressive that, despite the many problems with existing methodologies, each agency matches its projection approach quite well with the purpose of their projections. For instance, the VA and AHRQ use their projections for a more detailed but short-run estimate of healthcare costs. Appropriately, these two agencies use micro-simulation at the heart of their forecasting models. On the other hand, long-term fiscal solvency is the goal of the CBO and CMS projections, and so the focus is on very long-run projections of cost growth. The 75-year budget window for entitlement programs imposes a unique challenge on modelers at CBO and CMS. Accordingly, these agencies use extrapolation constrained by either restrictions on the decline of non-health goods or a CGE model. The CGE model itself is a recent addition, and modeling the economic dynamics better as our understanding of healthcare and our computational ability increases should allow CMS (and perhaps CBO) to make fuller use of this approach.
REFERENCES


