Health Services Research

Key to Health Policy

Edited by Eli Ginzberg

A Report from
The Foundation for Health Services Research

Harvard University Press
Cambridge, Massachusetts
London, England
1991
Health Manpower Forecasting: The Case of Physician Supply

Uwe E. Reinhardt

The supply of physicians has long been a concern of public policy in this country. During the four decades preceding World War II, it was widely believed that the nation suffered from a surplus of physicians. Following World War II and until about the mid-1970s, it was generally assumed that there were too few. The nation then had about 160 active physicians per 100,000 population. The conventional wisdom since about the mid-1970s has been that the United States is moving, once again, toward a sizable physician surplus; the physician-population ratio has risen to 211 active physicians per 100,000 population in 1985 and is heading for a projected level of 250 by the year 2000. The view that this high ratio will then constitute a physician surplus, however, is by no means universally shared (see, for example, Schwartz, Sloan, and Mendelson, 1988).

This perennial concern over the nation's supply of physicians has spawned one of the oldest branches of health services research in this country: health manpower forecasting and the allied research supporting such forecasts. The objective of this chapter is to review the evolution of this research enterprise and to assess its role in the formation of health policy.

The first section of the chapter sets forth a theoretical framework that illustrates the difficulty of forecasting in this area. Models designed to overcome these difficulties at the practical level are described in the second section. The third and fourth sections are focused on the interplay between health manpower forecasting and the formulation of public policy on the nation's supply of physicians. Some suggestions for future research are presented in the concluding section.
A Theoretical Framework for Health Manpower Forecasting

The basic steps in forecasting the future supply and requirements of physicians are always the same. The forecast begins with a projection of changes in the size and sociodemographic composition of the population that is to be served by the manpower in question—in the present case, physicians. Next, the population forecast is translated somehow into a projection of that population's "need" or "demand" for physician services. On the basis of an assumed physician-productivity ratio (defined as the average number of physician services procured per year per physician), the projected service requirement is then translated into the corresponding number of full-time-equivalent physicians needed to produce the required services. Finally, the projected requirement of physicians can be compared with the supply of active physicians likely to be yielded by the projected population and by additional physicians recruited from abroad. This comparison points up potential future imbalances that may require remedial public policy.

Figure 9-1 illustrates this research task at a highly conceptual level. In that scheme, the vector \( \{L_{10}, L_{20}, \ldots, L_{m0}\} \) represents the entire array of health manpower needed to attain a chosen target health status for a given population at a future point in time \( t \). For example, \( L_{10} \) may denote the number of full-time physicians needed for that purpose, \( L_{20} \) the number of full-time nurses, \( L_{30} \) the number of full-time cytotechnologists, and so on.

In each setting for health care—be it a medical practice, a hospital, a nursing home, or any other setting—a particular combination of health manpower, supported by other inputs, is used to produce health services. In Figure 9-1 the quantity of each type of health service is denoted by the notation \( S \). Thus, \( S_{1} \) might represent the annual output of a particular service rendered in the hospital. The symbol \( f_{i} \) denotes the so-called production function (a mathematical relationship to be described later) that describes the transformation of the services of health manpower and of other inputs into that particular hospital service. Similarly, \( S_{k+1} \) might represent the annual volume of, say, patient visits at physicians' offices and \( f_{k+1} \) the production function for that type of service, and so on for all of the other services rendered by the health system. Finally, a particular combination of health services constitutes the medical treatment prescribed to respond to a particular medical condition. As indicated in the figure, the medical treatments dispensed by a health
Figure 9-1. Potential trade-offs in the production of health. (From Reinhardt, 1975b.)
system should be thought of as only one of many inputs determining the health status of the population, a point commonly overlooked by pundits who are quick to blame international differences in health status indicators—such as infant mortality rates—strictly on the associated health systems.

It is emphasized in this scheme that, in principle, one ought never to look at just one type of health manpower in abstraction from other types, because many of them can act as substitutes for one another. In practice, this means that acceptable levels of health care can come forth from a great variety of different health manpower-to-population ratios—for example, from quite different physician-population ratios. To economists and other health services researchers, that supposition may be instinctive. Remarkably, the notion is still being resisted at the practical level of policy, where policymakers often act as though the production of health care were ruled by technically determined, fixed health manpower-to-population ratios.

In Figure 9-1 it is assumed that one can project the morbidity likely to be suffered by a given population at some future time and that it is also possible to define and quantify a technically feasible level of health status that policymakers would like to achieve for that projected population. Although this idea seems bold, it is not just theory. In the fall of 1980, the U.S. Department of Health and Human Services announced very specific health status targets to be achieved by 1990 in its Promoting Health/Preventing Disease: Objectives for the Nation (1980). Progress toward these objectives was reported in the department's 1990 Health Objectives for the Nation: A Mid-Course Review (1986). It may be noted in passing that, according to the report, many of the midway goals had been surpassed at that point.

Potential Substitution among Different Inputs into the Production of Health

The health status achieved for a given population is the output from a multifactor health production process managed, in the first instance, by individuals themselves as they choose a particular life-style. To be sure, the most powerful determinants of an individual's health status tend to be congenital and environmental factors beyond the individual's managerial control. Even so, the individual almost always does have some
influence over positive or negative changes from his or her natural health endowment.

Only occasionally are physicians drawn into the individual's health production process. In such instances physicians act as diagnosticians who provide information on the patient's health status, as management consultants and agents who help compose the patient's demand for health services, and, typically, also as the providers of at least some of the recommended medical procedures. How physicians perform their dual role of agent for the patient and supplier of health care—particularly how financial incentives bear on that performance—has long been a matter of controversy in health manpower forecasting. The issue will be discussed later in this chapter.

Researchers have traditionally found a rather low linear correlation between the per capita use of health services or health care resources per capita, on the one hand, and observable health status in the relevant populations, on the other (see, for example, Fuchs, 1974; McKinlay and McKinlay, 1977). That low correlation should not be surprising. First, as noted, since health care is but one of many inputs into the production of health, simple first-order correlations do not control for the influence of many other factors. Second, although other things being equal the use of health services may well have a powerful positive impact on measurable health status, the input of health services is likely to be subject to rapidly diminishing marginal returns. Third, measurable health status indicators do not usually capture all positive contributions made by health care. For example, simply providing patients with accurate information on their health status may be viewed by them as a valuable contribution to the quality of life, particularly if that information confirms a good health status.

Although there have been numerous studies of the effects of particular environmental, socioeconomic, demographic, and behavioral factors on measurable health status, only a few studies have sought to embrace all these variables formally within the framework of a full-fledged health production function (see, for example, Auster, Leveson, and Sarachek, 1969; Silver, 1972; and, notably, Hadley, 1982). These studies generally have found a positive incremental association between the use of health services and measurable health status (usually, mortality rates), after control for all other factors affecting health status. For example, Hadley found that over the empirically relevant range of his observations and for the bulk of the population, a 10 percent increase in the per capita use
of medical care tended to be associated with a 1.5 percent decrease in
mortality, other things being equal (1982, p. 169). Furthermore, studies
of this sort do support the hypothesis that the input of medical care into
the health production process is subject to rapidly diminishing marginal
returns in terms of measurable health status.

As Figure 9-1 suggests, full-fledged, empirical health production func-
tions would be the logical starting point for health manpower projections.
Unfortunately, none of the empirical estimates available so far have been
sufficiently detailed and precise for that purpose. Were such estimates
available, one could derive from them normative levels of health care
utilization per capita within a larger optimization model in which other
inputs into the production of health—for example, housing or nutrition—
would be traded off against medical care in a manner that minimizes the
total social resource costs of attaining the desired level of health status.

Even without that formalism, however, it is increasingly being appre-
ciated by policymakers in both the private and public sectors that such
trade-offs ought to be made—that, for example, the rate of infant mor-
tality among low-income families could, in principle, be vastly reduced
with means completely outside the medical model. Sadly, these insights
are not easily translated into coherent policy action within the frag-
mented budgetary processes typical of modern democracies. In that set-
ting, the production of better health (as distinct from health care) is likely
to be economically efficient solely by happenstance. More typically,
health care will be either overused or underused relative to the econom-
ically efficient combination of all the factors going into the production of
health.

Potential Shifts in the Setting of Health Care Production

As suggested in Figure 9-1, any target level of medical interventions
used in the production of better health could be produced in a variety of
different settings, and with a variety of different methods within each
setting. Each possibility implies a different requirement of health man-
power.

Shifts in the setting of health care, or changes in methods of delivery,
can make earlier health manpower forecasts highly uncertain ex ante and
seem widely off the mark ex post. During the 1960s and 1970s, for exam-
ple, it was generally not foreseen how massive a shift from the inpatient
to the outpatient setting would occur during the 1980s for a great variety
of medical treatments. Table 9-1 illustrates this shift with data from the Blue Cross plans. This shift was made possible in part by technological developments that made the transfer of treatments to the outpatient setting clinically safe; but it was also accelerated by a significant increase in the degree of external, regulatory control over the utilization of inpatient health care instituted by both public and private third-party payors during the 1980s.

The ultimate impact of this shift from the inpatient to the outpatient setting on overall health manpower requirements deserves more careful investigation than it has hitherto received, because that impact may be substantial. With the empirical insights gained from such studies, a health manpower forecasting model could be used to simulate the effect of future potential shifts in the setting of health care on future requirements of physicians or other types of health manpower.

Shifts of care from the traditional fee-for-service medical practice to health maintenance organizations (HMOs) also are known to have substantial effects on the requirement for medical manpower. The factor driving health manpower requirements in that shift is not so much the setting per se, but the distinct financial incentives that accompany the delivery of health care. Under the HMO insurance contract, these incentives encourage economy in the use of medical care inputs in the production of health and also in the use of real resources in the production of health services. Chiefly by virtue of these incentives, staff or group model HMOs tend to be able to serve more than 800 patients per full-time-equivalent physician. In their recently published study of practice patterns in large prepaid group practices, Mulhausen and McGee (1989) found that a closed panel HMO appears to be able to operate with a ratio

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital inpatient admissions per 1,000 members</td>
<td>107</td>
<td>85</td>
<td>-21%</td>
</tr>
<tr>
<td>Hospital outpatient visits per 1,000 members</td>
<td>350</td>
<td>490</td>
<td>+40%</td>
</tr>
</tbody>
</table>

*Source: Blue Cross and Blue Shield Association (1989), pp. 5 and 7.*
as low as an average of 111 physicians per 100,000 population, which implies about 900 patients per physician. Even if that number did not fully reflect differences in the mix of patients served by current HMOs and the fee-for-service sector, a fully adjusted ratio would still remain far below the overall ratio already attained in the United States. Tarlov (1986) posits a ratio of 120 physicians per 100,000 enrollees for a staff model HMO, a ratio also accepted by Schwartz, Sloan, and Mendelson (1988, p. 892) in their recent manpower projection. The comparable ratio for the traditional fee-for-service sector is now closer to 450–500 patients per physician, but it is falling steadily, reaching levels below 200 in some cities.

To illustrate the effect a shift from the traditional fee-for-service setting to a staff model HMO setting would have on total physician requirements, let us denote by $H$ the physician-enrollee ratio needed in an HMO, by $F$ the corresponding ratio needed in the fee-for-service sector, and by $T$, the overall physician-population ratio required for the nation as a whole. Then if $x$ denotes the percentage of the population served by staff model HMOs, the overall required physician-population ratio can be calculated from the expression

$$T = F - x (F - H).$$  \hspace{1cm} [1]

If $H$ were 120, for example, and $F$ 200, then the overall required ratio of physicians per 100,000 population would be 184 if 20 percent of the population were in HMOs, but only 160 if 50 percent of the population were in HMOs.

Conversely, if one lets the variable $T_a$ denote the overall supply of physicians per 100,000 population actually available at some point in time, and $F_a$ the number of physicians per 100,000 population left over for the fee-for-service sector after subtracting the physicians needed for the HMO sector, then, in terms of the variables defined above, we can express the ratio pressing in on the fee-for-service sector as

$$F_a = \left( \frac{1}{1 - x} \right) (T_a - x H).$$  \hspace{1cm} [2]

In Figure 9-2, $F_a$ is plotted on alternative values of $x$ on the assumption that the required physician-enrollee ratio for HMOs is $H = 120$ per 100,000 enrollees and that the overall average physician-population ratio available to the population as a whole is $T_a = 230$ physicians per 100,000
population (the ratio of patient care physicians per 100,000 population certain to be available in the year 2000). Clearly, with a given overall supply of medical manpower, a major shift of patients from the traditional fee-for-service sector to staff model HMOs could exert enormous economic pressure on the fee-for-service sector. For example, if half the American population were in staff model HMOs operating at a ratio of 120 physicians per 100,000 enrollees, then there would be 340 physicians left over per 100,000 population remaining in the fee-for-service sector. This implies that physicians in that sector would have to earn their gross income (currently an average of about $250,000 per year) from an average of only 295 patients, or about 100 families if one posits an average family size of about three. Physicians would either have very low incomes or would have to be able to charge very high fees, which would be improbable given the competitive presence of the HMO sector.

Potential Health Manpower Substitution within a Given Setting

Even within a given health care facility—a hospital, a nursing home, or a medical practice—one type of health manpower can often be traded off against another in the production of given sets of services. These trade-offs can be represented analytically with the aid of health service
production functions. A production function is a mathematical relationship indicating the rates of output technically attainable with alternative combinations of productive inputs. In a medical practice, for example, these inputs are the time spent by physicians and their support staff, the floor space available to them, the number of examination rooms at their disposal, and so on.

At the level of theory, a production function is generally understood to define the maximum rates of outputs technically attainable with alternative combinations of inputs, which implies that the production processes are operated as efficiently as is technically feasible. In empirical estimation, however, it is usually not known which producers in one's sample are perfectly efficient technically and which are not. An empirically estimated production function therefore usually reflects the average observed rates of output obtained with alternative combinations of inputs by a sample of different producers at one point in time.

A reliable production function estimate for a particular production process is a powerful analytical summary of that process. Such an estimate can be used to calculate the incremental contribution to output made by additions of one input to a given set of other inputs. For example, in the context of medical practice, the function allows one to calculate on average how many more patients per week a physician can see in the office if his or her support staff is increased by one person of a particular skill, or if the number of examination rooms is increased. Alternatively, the function can be used to identify the alternative combinations of physician time and the time of physician support staff capable of producing a given number of patient visits per week. In other words, it can be used to gain information on technically feasible trade-off curves between any two types of health manpower. Figure 9-3 illustrates such trade-off curves for physician-patient office visits in private medical practice.

If reliable production function estimates for all major categories of health services and all alternative health care settings were available to manpower forecasters, the latter could use these estimates to simulate alternative combinations of manpower—that is, the alternative manpower combinations \( \{ L_1, L_2, \ldots, L^m \} \) of Figure 9-1—that are capable of delivering a required set of medical treatments which, in turn, are deemed necessary to treat a given projected morbidity. Such numerical exercises would make it clear to policymakers that there does not exist just one ideal, technically determined manpower-population ratio, deviations from which, ipso facto, denote either a surplus or a shortage of
Figure 9-3. Estimated feasible trade-offs between physician time and auxiliary personnel, general practitioners. (From Reinhardt, 1975a.)

that type of manpower. For example, research of this sort on the use of health manpower in the hospital sector may possibly reveal that current fears of a chronic nursing shortage are based on the assumption of fixed nurse-to-bed ratios, totally in abstraction of the potential for health manpower substitution or capital-for-labor substitution within hospitals.

With the aid of empirical production functions, manpower forecasters could also identify the least-cost combination of manpower capable of supporting a desired set of medical treatments. Figure 9-4 illustrates that method hypothetically. In that diagram, the curved line represents an assumed, empirically estimated trade-off curve between two types of health manpower—for example, physician time and time of physician-support staff. In the technical jargon of economists, such a curve is known as an isoquant; that is, it traces out the alternative combination of these two types of manpower input capable of producing the same quantity of output. The straight lines in Figure 9-4 represent so-called
isocost curves, that is, each line represents alternative combinations of the two types of manpower inputs that cost the same.\(^2\) Combinations A, B, and C in the diagram all are capable of yielding, say, 100 patient visits per week. Combination B is relatively more intensive in its use of physician time than are combinations A and C. It is also the most expensive manpower combination, because it lies on the isocost line for $4,000 per week. Combination C, which costs $3,600 per week, is relatively more intensive in its use of physician-support staff than are combinations A and B. Finally, combination A, costing $3,400 per week, lies in between these extremes; it represents the least-cost manpower combination just capable of yielding 100 patient visits per week. Deviations from that least-cost combination would connote degrees of economic inefficiency in the use of human resources. Although it would be technically feasible to produce 100 patient visits per week with even fewer physicians than are implied by combination A, that extreme degree of economizing on the use of physician time would not be economically optimal.

*Full-Fledged Behavioral Models of the Health Care Sector*

Even under the best of circumstances, however, such simulation exercises would only reveal what would be technically feasible and economically desirable. By themselves, they would not provide one with a clue to the manpower combination likely to be used by the health system in
future years. For that purpose, the forecaster really ought to have at hand a full-fledged behavioral model of the health care sector, one example of which is shown in Figure 9-5. This figure illustrates the bare-bones structure of a behavioral health-sector model as it might be perceived through the particular prism used by economists. An arrow from one block in the diagram to another should be read as "partially determines." The figure depicts four distinct, interrelated markets: the market for health services (centered on box 11), the market for health manpower (box 23), the market for health manpower training (box 25), and the markets for other inputs into the production of health services (box 14). Not shown in the scheme, but implied by box 6, is the market for health insurance coverage, which is intricately linked to the market for health services because health care and health insurance coverage are typically jointly chosen.

The future supply of medical manpower is determined in the domestic market for health manpower training (box 25 in Figure 9-5) and the inflow of foreign medical graduates (box 29 for fully trained physicians and box 30 for physicians still requiring residency training). The inflow of foreign medical graduates (FMGs) is really the export of foreign-produced medical education to the United States. In the absence of borders and immigration laws, the latter flow would be determined in the foreign markets for health manpower training and by the decisions of the FMGs themselves. In effect, however, the binding constraint on that flow has been U.S. government policy on immigration, which needs to be anticipated to project the future flow of FMGs to this country.

Ever since World War II, the flow of graduates from the domestic market for health manpower training has been determined effectively by the supply side of that market (box 26), notably by federal and state policies on the financing of medical schools. Although the ratio of qualified applicants to available medical school places has fallen substantially during the 1980s, the market can still be said to be in a state of disequilibrium. In spite of the typically high tuition now charged by American medical schools, the demand for medical school places among intrinsically qualified candidates at this time still appears to exceed the supply of available places. In Figure 9-5, that situation would be described as one for which $E_{sd} > E_{sr}$, where subscript $t$ denotes a particular type of health manpower, in this case, physicians.

In his earlier work, Sloan (1970, 1971) had investigated the determinants of the demand for medical education (box 20 in Figure 9-5). He
Figure 9.5. A market-based health care system. (From Reinhardt, 1974.)
had found that demand to be somewhat sensitive to economic factors, such as tuition charges and expected earnings in medical practice and in rival professions. But probably because the supply of American medical graduates had been fully determined by the available capacity of medical schools for so long, there have been only a few studies since then on the factors that drive young Americans to enter medical school in favor of alternative professional careers (Bazzioli, 1986) or that drive the choice of a medical specialty (Hadley, 1977; Marder, Kletke, et al., 1988). In light of the falling ratio of medical school applications to available places, however, these facets of health manpower warrant more intensive study in the years ahead.

Among the numerous factors driving occupational choice there surely must be the pecuniary dimension, usually represented in economic analysis by the internal rate of return to medical education (roughly, the annual rate of profit a medical education returns on the investment it requires in terms of monetary outlays and forgone earnings in the next best professional alternative). Sloan (1970) found that candidates entering medical school during the later 1960s faced an annual rate of return of somewhere around 25 percent to their investment of money and opportunity costs in four years of medical school and one year of internship. More recent work (Burstein and Cromwell, 1985; Marder, Kletke, et al., 1988) suggests that the rapidly growing physician-population ratio in the meantime has served to depress these rates, although not to unattractive levels. For example, if one measures the opportunity cost of a medical career by the forgone average income that would have been earned with a career in law, then the estimated annual rate of return to a decision to enter the profession of medicine rather than of law was 9.1 percent in 1974 and 14.2 percent in 1985. Table 9-2 presents these and alternative estimates with opportunity costs measured by still other careers. In short, so far a medical career is by no means a low-yield investment in human capital. Incidentally, if one measures the opportunity cost of a medical career strictly by the average forgone earnings yielded by a baccalaureate degree, then the return to medicine would appear to be 16.8 percent in 1974 and 16 percent in 1985, although the investment base on which these rates would be earned is, of course, much smaller than the base implied by the forgone earnings of a lawyer (Marder, Kletke, et al., 1988).

Finally, it must be emphasized that, although pecuniary factors have been found to enter the decision to become a physician and to choose a
particular medical specialty, it is also clear from the available research that many other noneconomic factors enter that decision as well. Successful future research on these choices is therefore likely to be interdisciplinary.

The total supply of medical manpower (box 27 in Figure 9-5) feeds directly into the markets for health manpower in general (box 23), where it confronts the demand for health manpower (box 18). The demand for health manpower emerges from the market for health services (box 11) by the interaction of the demand for health services (box 8) with their supply (box 16), which, in turn, is shaped by the organization of the health system in general (box 17). The market for health insurance (boxes 6 and 9) obviously also has a strong influence on the market for health services.

Part of the market for health services is the market for physician services proper. Just how that market clears (that is, what factors make the quantity demanded, \( Q_d \), equal the quantity supplied, \( Q_s \), after sudden changes on either the demand or supply side of that market) has remained a mystery and, hence, a controversy among health services researchers. At issue are the questions of whether, how, and to what

\[
\text{Table 9-2. Estimated annual rates of return to the decision to enter the medical profession rather than a competing profession}
\]

<table>
<thead>
<tr>
<th>Measure of opportunity cost</th>
<th>Estimated annual rate of return to medicine above the competing profession¹</th>
<th>1974</th>
<th>1985</th>
</tr>
</thead>
<tbody>
<tr>
<td>The stream of earnings that could have been earned in:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Law</td>
<td>9.1%</td>
<td>14.2%</td>
<td></td>
</tr>
<tr>
<td>Accounting</td>
<td>13.9%</td>
<td>13.8%</td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>11.6%</td>
<td>9.9%</td>
<td></td>
</tr>
<tr>
<td>Other postgraduate education</td>
<td>20.5%</td>
<td>18.3%</td>
<td></td>
</tr>
</tbody>
</table>


a. This is the estimated average rate of return to the investment represented by the money spent on tuition and other incidental costs of medical school and by the earnings forgone by not pursuing the other professional career listed under the heading “Measure of Opportunity Cost.”
extent physicians on the supply side of that market (box 16) can manipulate the demand for their services (box 8).

In a normal, competitive market, for example, suppliers can seek to increase the quantity of their goods and services demanded by the rest of society either by lowering prices or by enhancing the quality of these goods and services, or by both means. Of course, in many markets suppliers can also seek to induce additional demand simply by persuading potential customers through truthful or deceptive advertising, although there are limits to that mechanism. The question in connection with physician services hinges on the degree to which physicians can increase the demand for their services through the analogue of advertising (see Reinhardt, 1985b, p. 190) simply by persuading their patients to accept such services, rather than through the more traditional equilibrators operating in normal markets: reductions in prices (fees) or enhancements of quality. If physicians had a fairly high degree of discretion in this respect, then a growing physician surplus would tend to be at least partially camouflaged by the delivery of added health services whose benefit-cost ratio might be trivial or even negative. I shall return to that question in a later section.

During the past two decades, much of the work of health services researchers in general, and of health economists in particular, has sought to identify empirically the nature and strengths of the causal flows in Figure 9-5. To attempt even a cursory overview of this vast body of research would go much beyond the limits of this chapter and is not my objective. Suffice it to note that this research has no parallel elsewhere in the industrialized world at this time, either in terms of its sheer volume or in terms of its technical sophistication. Much is now known with reasonable precision about the behavior of at least some parts of the American health care sector that could only be guessed at one or two decades earlier.

Alas, despite this impressive research record, it is as yet too early to fashion from this considerable wealth of empirical information one reliable, behavioral model of the entire health care sector that could be used to project health manpower shortages or surpluses decades hence.3 (It may be added, of course, that reliable long-run forecasting models of this sort do not yet exist for any other sector of the economy, nor even for the economy as a whole.)4 The nature of such a comprehensive health sector model is illustrated here only to emphasize that a health manpower forecast will pinpoint, at best, the most likely physician
requirements and supplies of an entire probability distribution of possible numbers that may actually obtain in the future. These probability distributions of possible future numbers always exhibit some dispersion of possible values about the most likely value being projected by the forecasters. That dispersion, of course, increases rapidly with the length of the forecast horizon. Thus, the very best any forecaster could ever do would be to estimate the parameters of the probability distributions that describe the range of possible future values of the projected variable.

It follows that one cannot fairly cast aspersions at health manpower forecasters if, say, the physician supply or requirement projected for a point in time ten or twenty years hence turns out, ex post, to have been wide of the mark, just as one should not ridicule those who seek to forecast bond yields or foreign exchange rates only several months hence. First, the forecasted variables themselves usually are functions of a host of other variables whose future time path is uncertain at the time the forecast is made. Second, and more important, the forecasts themselves often trigger policy actions designed to close gaps between the forecasted numbers and preferred targets. As will be seen later, this has certainly been the case in health manpower forecasting.

Pragmatic Models for Health Manpower Forecasting

Because an operational, full-fledged, behavioral model of the health sector has never been available to health manpower forecasters, they have traditionally made do with more compact models that can be viewed as rough-and-cut summaries of the larger model sketched out in Figure 9-5. Figure 9-6, which is largely self-explanatory, illustrates the more compact models that have been used in recent decades. All the variables in this figure have time subscripts, which implies that each of them should, in principle, be projected separately. The most difficult variables to project, and also the most crucial ones, are (1) the projected per capita demand for physician services at time \( t \) (variable \( D_t \)) and (2) physician productivity, defined here as the projected average output of physician services per full-time-equivalent physician at time \( t \) (variable \( Q_t \)). If one writes out the ratio \( D_t/Q_t \) as

\[
\frac{D_t}{Q_t} = \frac{\text{Projected Number of Physician Services/Population}}{\text{Projected Number of Physician Services/Physician}}
\]  

[3]
then the ratio will appear as nothing other than the physician-population ratio deemed necessary to provide the projected service requirements $D_t$.

Figure 9-7 illustrates this forecasting equation diagrammatically. The diagram represents a construct often used by economists to illustrate the interaction among variables. It is not a regular Cartesian coordinate; instead, it contains four positive quadrants pasted back-to-back so that every axis measures a positive direction for the variable in question. The heavy, solid, upward-sloping line labeled A in Quadrant I of the figure represents the projected supply of patient care physicians (the term $a, c, S_t$ in the forecasting equation). Quadrant II depicts the growth of the total population ($N_t$) as a function of time. Because that quadrant is upside down, the curve slopes downward. Quadrant III depicts the total demand (need) for physician services as a function of the population size. That line also slopes down because the quadrant is upside down. The slope of the line is the per capita demand for physician services (the variable $D_t$ in the forecasting equation), here held constant over time to facilitate the graphic illustration. Quadrant IV depicts the total number of physician services available as a function of the total number of patient
care physicians. The slope of that line represents physician productivity (the variable $Q$, in the forecasting equation), once again held constant here for ease of graphic representation.

Figure 9-7 is a convenient pedagogic device to make the entire process of health manpower forecasting graphic. One starts with a point on the time axis (some future year), reads off the projected population for that year in Quadrant II, reads off the corresponding demand for physician services in Quadrant III, translates that demand into the corresponding required number of physicians in Quadrant IV, and then compares that projected requirement of physicians with the projected supply of patient care physicians in Quadrant I to assess the supply-demand balance. Shifts in any of the curves in the diagram directly translate
themselves into changes in the projected surplus or deficit of physicians. For example, the indicated shift of the input-output line in Quadrant IV represents a gain in physician productivity; it could turn a projected physician shortage into a surplus.

*Needs-based versus Demand-based Forecasts*

Many of the forecasting exercises developed during the first half of this century simply posited normative physician-population ratios that were recommended as targets for health manpower policy. These normative ratios were then compared with projected actual physician-population ratios at future dates to identify probable shortages or surpluses. Such forecasts generally did not attempt to project separate series for per capita utilization \((D_c)\) and physician productivity \((Q_c)\) explicitly. Instead, they assumed implicitly that per capita demand and output per physician would grow at roughly the same rate over time. More recent work typically has sought to project each of the variables in Figure 9-6 separately. In these studies, the utilization series \(D_i\) is either a *needs-based* or a *demand-based* estimate. Because the choice between these two bases remains rather controversial, it merits added comment here.

The needs-based approach to health manpower forecasting is the purely epidemiological approach illustrated earlier in Figure 9-1. Under that approach, an attempt is made to project the morbidity to be treated in future years and to estimate the medical services that will be needed to respond adequately to that projected morbidity. The projected need for medical services is then translated into the number of full-time physicians required to produce the needed service, a procedure that either explicitly or implicitly assumes a projected time path of physician productivity \((Q_c)\). Classic illustrations of that approach are Lee and Jones (1933) and the Final Report of the Graduate Medical Education National Advisory Commission (1980), hereafter referred to simply by its famous acronym, GMENAC.

Under the needs-based approach, no attention is paid to the question of whether society will actually be willing to finance all of the treatments deemed to be "needed," or whether future medical manpower will actually locate itself in accordance with the geographic distribution of the projected need for medical care. Nor is allowance made for the possibility that, within limits, physicians in surplus may seek to eliminate that surplus by artificially stimulating the demand for their services beyond a
level justifiable on purely medical grounds. In short, the approach is normative and not predictive. It merely yields recommended targets.

The virtue of needs-based forecasting is that it reflects an informed consensus among medical experts on the task that will have to be achieved in the future if society wishes to own up to the much-mouthed maxim that every citizen should have access to “needed” medical care and, furthermore, if society does not deploy the available medical manpower wastefully—for example, by devoting it disproportionately to the wealthy, worried well rather than to critically ill members of, say, uninsured low-income families. A major shortcoming of the approach is that it remains, essentially, organized, wishful, expert thinking that may be quite removed from the real world.

The demand-based approach typically seeks to project the effective future demand for medical care, that is, perceived need backed up by ability and willingness to pay for that care. On that approach, the time path of future utilization is an extrapolation of observed current utilization patterns, adjusted where possible for projected changes in the sociodemographic composition of the population and for changes in insurance coverage. In effect, such forecasts seek to simulate what a full-fledged behavioral model of the health-care sector would project, without actually going into the guts of such a model. Recent examples of this type of forecasting are Schwartz, Sloan, and Mendelson (1988) and Marder, Kletke, et al. (1988).

In contrast to the needs-based approach, which is purely normative in the sense that it suggests what ought to happen in the future, the demand-based approach is purely positive in the sense that it merely seeks to predict the future. The demand-based approach implicitly allows for the possibility of physician-induced demand for health services since, in its purest form, it does not even distinguish between medically necessary and unnecessary procedures. It tacitly regards treatments accepted by patients as treatments demanded by them.

For example, a needs-based forecast would project a surplus of a given specialty of physicians if the projected supply of these specialists exceeded the number required to meet the projected need for their services. By contrast, if it could reasonably be conjectured that these specialists are likely to spend a significant part of their time on medically dubious or outright unnecessary procedures—a conjecture entirely justified within the American context (Brook and Vaiana, 1989)—then a demand-based forecast might characterize that very same situation as a
projected physician shortage, if the supply of physicians is judged inadequate to meet the projected demand for both appropriate and inappropriate procedures. Similarly, if the projected supply of physicians is judged to be inadequate to meet the need for medical intervention projected by medical experts, then the needs-based approach would project a physician shortage. The demand-based approach, on the other hand, might well characterize that very same situation as a projected physician surplus if some of the projected need of care is not likely to be backed up by a willingness to finance that care, and if the projected supply of physician services exceeds the projected demand for those services (that is, perceived need of care that is backed up by a willingness to finance that care).

This difference in methodology—the purely normative versus the purely positive basis of forecasting—has caused much confusion in the debate of health manpower policy, and it is likely to have contributed to the current controversy over the question of whether or not the United States now faces a physician shortage or a surplus. Those who, like the GMENAC Report, point to a sizable future physician surplus tend to lean on the needs-based approach, while those who, like Schwartz, Sloan, and Mendelson (1988), doubt the emergence of a surplus tend to rely on a demand-based approach.

In fact, of course, neither approach is inherently superior to the other, nor are they mutually contradictory. The two approaches merely respond to quite different policy questions. One seeks to yield normative targets, and the other seeks to predict what would be likely to happen if policy were not geared deliberately to move the health sector toward the normative target. Ideally, a comprehensive health manpower forecast ought to embrace both approaches.

Early Health Manpower Forecasts and Their Impact on Health Manpower Policy

Probably the first formal attempt to assess the adequacy of the nation's physician supply was undertaken, during the period 1908–1910, by the Carnegie Foundation for the Advancement of Teaching. The final report of that study, known today as the Flexner Report, noted in its introduction that “taking the United States as a whole, physicians are four or five times as numerous in proportion to population as in older countries like
Germany" (Flexner, 1910, p. x). Implicitly, the Flexner Report took Germany's lower physician-population ratio as more nearly the proper ratio. The Flexner Report subsequently became the blueprint for a major reform of medical education in the United States, which led, \textit{inter alia}, to a substantial reduction in the number of American medical schools.

The notion that the United States was beset by a physician surplus continued into the 1930s. In a report written by the Commission on Medical Education, organized by the Association of American Medical Colleges and supported by a number of prestigious foundations, it was concluded that

if the United States had the same ratio of physicians to the population as England and Wales, there would be about 82,500 doctors in this country. If the ratio were the same as in Germany, the total would be 79,000; France, 73,000; Norway, 70,000; Sweden, 42,500. The actual number at present in the United States is 156,440 . . . An adequate medical service for the country could probably be provided by about 120,000 active physicians . . . On such an assumption there is an oversupply of at least 25,000 physicians in this country. (Final Report of the Commission on Medical Education, 1932, cited in Ginzberg, 1987, p. 6)

Very shortly thereafter, however, Roger Lee and Lewis Jones argued in \textit{The Fundamentals of Good Medical Care} (1933) that as of that year the United States actually had 13,000 fewer physicians than could be considered adequate on medical grounds. This study is the first attempt at a proper needs-based approach, relying upon detailed assessments of morbidity and expert opinion on the care needed to address that morbidity properly. The study sought to alert policymakers that, in spite of the prevailing protestations over physician surpluses at the time, millions of Americans actually went without financial or physical access to adequate health care.

The Lee-Jones study was followed, after World War II, by the Ewing Report in 1948, the Mountin-Pennell-Berger report in 1949, the President’s Commission on the Health Needs of the Nation in 1953, the 1958 report of the Department of Health, Education, and Welfare consultants, known as the Bayne-Jones report, and the report of the Surgeon General’s Consultant Group on Medical Education in 1959, more commonly known as the Bane Report. Each of these reports based their forecasts on some normative physician-population ratio, which typically was the
ratio prevailing in the more richly endowed regions of the country. Alternatively, it was recommended that, within the forecast horizon, every region in the United States be brought to at least the average national physician-population ratio prevailing at the time of the forecast. Given that approach, it is not surprising that each report continued the theme first sounded in the Lee-Jones Report, namely, that the nation was facing a severe physician shortage.

Each of these reports called for a major expansion of medical school capacity, and each called for direct federal assistance to both students and institutions to facilitate that expansion. The Bayne-Jones report, for example, had called for an increase in the number of first-year medical school places from 6,800 in 1955 to 8,700 in 1970. The subsequent Bane Committee report called for an expansion in first-year places from 7,400 in 1959 to 11,000 in 1975. Little thought appears to have been given in these reports to the long-run effects of large increases in medical school capacity on the supply of active physicians, since, for decades to come, the number of entering practitioners would continue to outnumber vastly the number exiting from medical practice.

*The Political Response to the Predicted Physician Shortage*

Neither the Eisenhower administration, on receiving the bulk of these reports, nor the Congress was willing to embark upon the bold new direction recommended by the forecasters. Hitherto, the federal government had supported medical education only indirectly and stealthily through the awarding of research grants, a time-hallowed method in higher education. The sundry manpower commissions recommended the direct and open intrusion of federal financing into medical education, which was vigorously opposed by organized medicine (Ginzberg, 1987, p. 19).

The dire forecasts therefore collected dust until the years of the Kennedy Johnson administration, when they became the analytic foundation for the pathbreaking Health Professions Educational Assistance Act (HPEAA) signed into law in 1963. The HPEAA provided the substantial, direct federal assistance called for by the postwar health manpower commissions. It offered matching federal grants for the construction of medical schools as well as loans and grants to students of medicine, osteopathy, and dentistry. Together, the original act and its subsequent amendments of 1965, 1968, and 1971 helped literally to double the
capacity of American medical schools within a very short time, in part with the powerful incentive of capitation grants, which made each school's federal grants directly proportional to the size of its student body.

This massive expansion in the capacity of American medical schools was accompanied by an equally massive influx of foreign medical graduates, who were eagerly invited into the country not only to augment the overall supply of physicians quickly but also to staff hospitals and clinics in locations shunned by American physicians—for example, in rural areas or in the low-income sections of major cities. Throughout the first half of the 1970s, close to a third of the 50,000 to 60,000 residency positions in graduate medical education in the United States were filled by foreign medical graduates. By the mid-1980s, that percentage had declined once again to about 17 percent, although of a larger total of 75,000 residencies (Marder, Kletke, et al., 1988, fig. 4 and table 4).

Passage of these generous HPEAA amendments was undoubtedly facilitated by yet another round of health manpower forecasts coming forth during the latter part of the 1960s, each forecasting sizable shortages of physicians as of the mid-1970s. In 1967, for example, the prestigious National Advisory Commission on Health Manpower issued a much-publicized report recommending a continued expansion of first-year enrollments in existing medical schools and the establishment of new schools. Close on its heels followed a forecast by the U.S. Department of Labor, Bureau of Labor Statistics (1967), which offered similar advice. That forecast, in turn, was followed by a U.S. Public Health Service (1967) report recommending that the United States attain, by 1975, the physician-population ratio attained by the best-endowed region in the United States in 1966. Using this standard, the Public Health Service projected a shortage of 50,000 as of 1969, a figure that was cited also in the 1970 Manpower Report of the President, whence it found its way into the Carnegie Commission's highly influential report Higher Education and the Nation's Health: Policies for Medical and Dental Education (1970).

The Carnegie Commission posited a direct linkage between health manpower and better health and, on that thesis, recommended that the number of medical school entrants or their equivalent be increased by 50 percent, from the 10,800 estimated for 1970–71 to about 15,300 by 1976 and about 16,400 by 1978 (1970, pp. 2 and 43). That bold recommendation had been fully realized by 1980. Between 1966 and the early
1980s, the number of American allopathic medical schools (conferring the M.D. degree) increased from 89 to 127, and first-year enrollments in these schools from 8,900 to 17,300. In 1987 there were still 127 allopathic schools, with 16,779 first-year places and 15,872 graduates (Marder, Kletke, et al., 1988, tables 2.11 and 2.12). In addition, enrollment in osteopathic schools (conferring the D.O. degree) increased rapidly as well, from 623 first-year places in 1970–71 to 1,724 in 1986–87 (Council on Graduate Medical Education, 1988, vol. 2, table 3, p. 18).

Second Thoughts among Policy Analysts

As policymakers opened the floodgates to American and foreign-trained would-be entrants into medical practice in the United States, health services researchers began to wonder whether there could be too much of a good thing (see, for example, Dickinson, 1948; Ginzberg, 1960, 1966). Protected from the relentless political pressure to produce quick solutions to perceived social problems, the research community had the luxury of taking a long-run perspective. Figure 9-8 illustrates that perspective.

The lower of the two curves in Figure 9-8 represents the projected physician-population ratio likely to be generated by projected graduations as of 1973. It assumes a gradual expansion of medical school capacity from about 10,000 graduates in 1973 to about 16,000 in 1990. The curve includes foreign-trained medical graduates (FMGs) already practicing in the United States as of 1973, but it excludes all FMGs entering thereafter. In effect, the curve significantly underestimates the physician-population ratio that would actually obtain under a sustained influx of FMGs. The upper curve in the figure projects the physician-population ratio that would obtain if the annual number of graduates were increased to 25,000 by 1990, as was occasionally being proposed at the time (see, for example, Gerber, 1967). Once again, that supply projection excludes any influx of FMGs after 1973, which is, of course, not realistic. The point of the graph, however, is not to make projections of the actual physician-population ratio likely to obtain, but merely to illustrate the long-run implications of policies that sought to meet a perceived physician shortage quickly through vast increases in the capacity of American medical schools (and by opening the borders to large numbers of FMGs). In the longer run, such a policy may overshoot the target by a wide margin.
Aside from the long-run implications simply of medical school expansion, researchers also pointed to the enormous differences one observes across the United States in prevailing physician-population ratios (see Table 9-3). It was noted that these differences were substantially offset by differences in the number of annual physician-patient visits in medical practices and hospitals, which, in turn, appeared to be explained in part by more extensive reliance on physician support personnel in areas with low physician density and high annual rates of patient visits per physician (see Table 9-4). Although differences in the quality of the physician-patient visit might explain some of the observed differential in weekly visits, it was reasonable to attribute a good part of it to genuine differences in physician productivity.

The notion that the physician shortage projected for the 1970s might be partially solved by encouraging greater physician productivity had been put forth as early as 1949 by the economist Frank G. Dickinson in his critique of the Ewing Report of 1948. This idea was formally incor-
Table 9-3. Physician-population ratios, patient loads, and medical fees by size of county, United States, 1970

<table>
<thead>
<tr>
<th>Demographic county classificationa</th>
<th>Physician-population ratiob</th>
<th>Weekly patient visits</th>
<th>Fee for an initial office visit (in dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>Office</td>
</tr>
<tr>
<td>Non-metropolitan:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10,000–24,999</td>
<td>51</td>
<td>223</td>
<td>167</td>
</tr>
<tr>
<td>25,000–49,999</td>
<td>64</td>
<td>217</td>
<td>164</td>
</tr>
<tr>
<td>50,000 or more</td>
<td>87</td>
<td>192</td>
<td>153</td>
</tr>
<tr>
<td>Metropolitan:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50,000–499,999</td>
<td>107</td>
<td>194</td>
<td>150</td>
</tr>
<tr>
<td>500,000–999,999</td>
<td>141</td>
<td>167</td>
<td>140</td>
</tr>
<tr>
<td>1,000,000–4,999,999</td>
<td>150</td>
<td>138</td>
<td>114</td>
</tr>
<tr>
<td>5,000,000 or more</td>
<td>191</td>
<td>124</td>
<td>109</td>
</tr>
</tbody>
</table>

Source: Reinhardt (1975a), table 2-6.
a. Numbers refer to inhabitants.
b. Number of nonfederal physicians in patient care per 100,000 resident population as of December 31, 1970.

The analysis was incorporated into the analysis of the National Advisory Commission on Health Manpower (1967) and was the central theme of Rashi Fein’s *The Doctor Shortage* (1967).

Figure 9-9 illustrates the sensitivity of projected physician requirements to assumptions about the future growth in physician productivity. The vertical axis in that diagram represents the required future physician-population ratio 10 and 20 years hence, under varying assumptions about the future ratio of the per capita demand for physician services at future time t (variable $D_t$) to the number of physician services produced per physician in year t (variable $Q_t$). As noted earlier in connection with Figure 9-6, this ratio $(D_t/Q_t)$ is the projected required physician-population ratio.

Now suppose initially that the forecaster had assumed the future growth in physician productivity to be just offset by future growth in the per capita demand for physician services and that, on this assumption, a ratio of 185 full-time physicians per 100,000 population was deemed adequate to serve the projected needs of the population. If, however, the future growth in physician productivity outpaced the growth in per capita
Table 9-4. Relationship between number of aides per M.D. and practice hours, hourly patient visits, and expenditures on medical supplies per visit

<table>
<thead>
<tr>
<th>Number of aides per M.D.</th>
<th>0 to 0.5</th>
<th>0.51 to 1.00</th>
<th>1.0 to 2.00</th>
<th>2.01 to 3.00</th>
<th>More than 3.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total practice hours/week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General practitioners (N = 371)</td>
<td>54</td>
<td>55</td>
<td>59</td>
<td>60</td>
<td>62</td>
</tr>
<tr>
<td>Pediatricians (N = 133)</td>
<td>49</td>
<td>50</td>
<td>50</td>
<td>56</td>
<td>55</td>
</tr>
<tr>
<td>Obstetricians/gynecologists (N = 101)</td>
<td>49</td>
<td>51</td>
<td>55</td>
<td>53</td>
<td>61</td>
</tr>
<tr>
<td>Total patient visits/practice hour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General practitioners</td>
<td>2.2</td>
<td>2.6</td>
<td>3.2</td>
<td>3.8</td>
<td>4.0</td>
</tr>
<tr>
<td>Pediatricians</td>
<td>1.8</td>
<td>2.6</td>
<td>3.3</td>
<td>3.8</td>
<td>3.9</td>
</tr>
<tr>
<td>Obstetricians/gynecologists</td>
<td>1.7</td>
<td>1.9</td>
<td>2.5</td>
<td>2.8</td>
<td>3.2</td>
</tr>
<tr>
<td>Medical supplies/office visit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General practitioners</td>
<td>$0.55</td>
<td>$0.45</td>
<td>$0.56</td>
<td>$0.55</td>
<td>$0.66</td>
</tr>
<tr>
<td>Pediatricians</td>
<td>.56</td>
<td>.33</td>
<td>.49</td>
<td>.50</td>
<td>.66</td>
</tr>
<tr>
<td>Obstetricians/gynecologists</td>
<td>.41</td>
<td>.48</td>
<td>.43</td>
<td>.62</td>
<td>.78</td>
</tr>
</tbody>
</table>

Source: Reinhardt (1975a), table 4-9.

a. Includes solo practitioners only. N denotes number of respondents.
b. Excludes time spent on reading, research, and teaching.
c. Patient visits include visits to patients in the home or in hospitals.
d. Reported expenditures on medical supplies and drugs and small instruments per visit at the physician's office.

demand by one percentage point per year, only 167 physicians per 100,000 population would be needed ten years hence, and only 151 twenty years hence. On the other hand, if physician productivity lagged growth in per capita demand by one percentage point, then about 203 physicians per 100,000 population would be needed ten years hence and 226 twenty years hence. It may be noted that, during the late 1960s and early 1970s, health services researchers estimated the annual growth in physician productivity during the period 1945–1968 to have been anywhere between 2 percent and 4 percent, a range that bracketed the estimated annual growth in per capita demand (Reinhardt, 1975, chap. 3).
Research on Physician Productivity

In the late 1960s, the available empirical information on physician productivity was very rough indeed. It consisted in the main of long-term trends in the average real annual gross billings per physician or in average annual patient visits per physician. Such estimates, however, were considered treacherous by the research community, because there is never any assurance that historical trends will replicate themselves in the future. For example, it might have been thought in the midst of the perceived physician shortage during the 1960s that all potential sources of gains in physician productivity would by then have been fully exploited by the health sector. To gain a better idea of the potential for future increases in physician productivity, researchers deemed it essential to

Figure 9-9. Sensitivity of future physician requirements to growth in physician productivity. (From Reinhardt, 1975a.)
explore the determinants of that important parameter more formally. Work on that issue began in earnest toward the end of the decade.

As noted earlier, empirical information on the determinants of physician productivity is obtained most easily through econometric estimation of physician-service production functions. Such estimates were first published by Smith, Miller, and Golladay (1972) and by Reinhardt (1972). The former researchers styled the physician's practice as a multi-product firm and based their estimates on linear activity analysis, that is, a so-called discrete production function model. That model allowed them to represent the output from a medical practice as an entire array of some 40 distinct physician services. On the other hand, the model forced them to estimate empirically a separate input-output coefficient for each pair of some 7 types of health manpower and some 40 types of physician services. They accomplished this task through work sampling techniques, a variant of time-and-motion studies. Once the large set of input-output coefficients was estimated in this way, the model was then manipulated with linear programming techniques to identify the degree of health manpower substitution that seemed technically feasible within a modern medical practice, and also to identify least-cost combinations of health manpower for given sets of output of the 40-odd physician services embraced by their study.

Reinhardt (1972, 1975) used a continuous production function specification whose parameters were estimated empirically from the data yielded by a large, nationwide cross section of medical practices. In that production function specification, the output from the physician's practice was represented by a single index: the number of weekly patient visits at the physician's office. That output index, of course, is much cruder than the multidimensional representation used by Smith, Miller, and Golladay (1972). On the other hand, the method is easier to implement in practice and also much less sensitive to errors in measuring particular observations. As already illustrated in Figure 9-3, a continuous production function can be used both to identify technically feasible trade-offs among types of health manpower and to identify the economically most efficient health manpower combination capable of producing given rates of patient visits.

Both of these production function studies indicated that there was much more leeway than had traditionally been thought by health manpower planners for substituting nonphysician labor for physicians within medical practices. Both studies also suggested that this potential did not
appear to have been fully exploited by the representative American physician at that time—that, from a strictly economic point of view, the prevailing pattern of medical practice in the United States appeared to be wasteful in the use of medical manpower.

Indeed, the research suggested further that, had American medical schools not been expanded so rapidly during the late 1960s and early 1970s, even rather large future increases in the per capita demand for patient visits—increases far in excess of those that actually obtained by the mid-1980s—could easily have been met simply through greater reliance on physician substitutes within medical practices (see, for example, Reinhardt, 1975, chap. 7). Rumors of a pervasive physician shortage during the late 1960s seem to have been much exaggerated, just as the current nursing shortage may one day be found to have been exaggerated, once research identifies the full potential for health manpower substitution within the hospital.

Unfortunately, the very swiftness of the expansionist health manpower policy during the late 1960s had rendered these research insights on the physician shortage moot. Had these insights been available half a decade earlier, they might well have given policymakers pause and altered the course of health manpower policy in the ensuing years. To be sure, those expansionary policies cannot fairly be characterized as an uninformed overreaction on the part of policymakers to the dire health manpower forecasts pouring forth during the 1950s and 1960s. These policymakers were not reacting solely to the forecasters. Additional political pressure for a major expansion of medical schools came from two other sources which might well have overpowered even earlier warnings of a potential physician surplus.

First, the passage of the Medicare and Medicaid programs in 1965, after years of vehement political debate, must have made policymakers nervous. No sensible politician, having voted for these ambitious programs, would have wanted to stand accused later on of having promised vastly expanded health care benefits to the aged and the poor without thinking about the health manpower needed to realize this grand promise. In part the rapid expansion of medical school capacity during this period simply reflected normal risk aversion in the political arena.

A second source of political pressure for a major expansion of medical school places, however, undoubtedly came from the parents of the large and relatively well educated baby boom generation. These parents desperately sought lucrative career opportunities for their numerous off-
spring, not only in medical schools but in all areas of higher education. This parental pressure, incidentally, was not confined solely to the United States; it was manifest in every other country beset by a postwar baby boom. Most of the European nations, for example, expanded their medical schools during the 1970s even more rapidly than did the United States (see Schroeder, 1984; Reinhardt, 1985a), and they now report physician-population ratios far in excess of ours. Unlike Americans, who still enjoy the luxury of debating whether or not the United States actually does face a physician surplus, European policymakers are now anxiously exploring how to cope with the massive physician surpluses that everybody agrees they now face.

From Physician Shortage to Surplus: The GMENAC Report

Long before the ratio of active physicians to population actually began to rise in response to the massive expansion of medical school capacity during the late 1960s, the writings of health services researchers persuaded policymakers that the gates leading to the practice of medicine might have been opened too far, with possibly undesirable consequences for health care in the United States. Sometime toward the mid-1970s, talk of an overall physician shortage gave way to talk of a physician surplus, and interest in physician productivity gave way to concern over a perceived maldistribution of American physicians across geographic regions and across medical specialties and concern over the prospect that a physician surplus might drive up the cost of medical care beyond levels justifiable on purely medical grounds.

As early as 1973, Assistant Secretary for Health Charles Edwards had warned Congress that the nation faced the prospect of a major physician surplus (Ginzberg, 1987, p. 33), and in April of 1976 the Secretary of Health, Education, and Welfare appointed the Graduate Medical Education National Advisory Committee (GMENAC) with the mandate to review the state of American graduate medical education and to assess the supply of and demand for medical manpower in the United States.

Long before the GMENAC issued its report in late 1980, however, this shift in perceptions of the health manpower situation had already influenced federal legislation. The 1976 amendment of the Health Professions Educational Assistance Act, for example, explicitly mentioned the prospect of a physician surplus. The act curtailed the capita-
tion grants for medical schools and the flow of foreign medical graduates into American medical practice, and it provided instead direct financial incentives designed to redistribute physicians both geographically and across specialties. Chief among the new policy instruments was the National Health Services Corps (NHSC), an imaginative scholarship program designed to direct graduating physicians into rural areas or urban institutions characterized by a shortage of medical manpower in return for federal financing of their medical education.

In the meantime, under the leadership of Alvin R. Tarlov, and with the active support of the Carter administration, the GMENAC launched what is surely one of the most ambitious health manpower forecasting exercises ever undertaken in the history of American health services research. The product of that research effort was reported in September of 1980. In a nutshell, the report predicted an overall surplus of 70,000 physicians by 1990 and one of 145,000 physicians by 2000, although for some primary care specialties the GMENAC actually projected future shortages.¹¹

Not surprisingly, Congress and the rest of the nation greeted the GMENAC report with very little interest. Although the prospect of a physician shortage tends to strike fear into the hearts of the public—and hence of politicians—it is just not clear to anyone then and now whether a physician surplus actually is harmful to anyone other than physicians’ financial health. Indeed, it has been argued, notably by neoclassical economists (Harris, 1986; Rossiter and Wilensky, 1987; Feldman and Sloan, 1988), that from the patient’s perspective, the greater the supply of competing physicians, the better. Furthermore, by the early 1980s the nation had persuaded itself of the view that the best public policy in most matters is for the government not to do anything at all—to let the private markets work out whatever imbalances might develop in the economy.

On the other hand, the GMENAC report quickly triggered a passionate debate among stakeholders in American medical education and in the research community. One would have expected opposition to the report from the medical education establishment, whose flow of public subsidies might be reduced by Congress on the strength of the report. Remarkably, however, the American Medical Association and some medical specialty societies (particularly the American College of Surgeons) also sharply attacked the report, although one would have thought that they might welcome any policy to reduce the number of physicians seeking their income from a patient pool projected to grow much less rapidly
than the supply of physicians. Perhaps, in setting their policies, the leaders of organized medicine trade off the size of their members' average income against the size of their organizations' political base, which finds strength in number. Perhaps there may have been yet other, and possibly loftier, reasons for organized medicine's reaction to the GMENAC report, reasons not easily grasped by an economist's one-track mind.

In American political debate, it is customary never to couch one's reaction to a particular study in terms of its potential impact upon one's own economic or political position. Instead, the proper etiquette is to phrase that reaction strictly as a seemingly objective, scientific evaluation of the methodology used in the study. And thus it is that the GMENAC study, easily the most imaginative and sophisticated health manpower forecast ever produced in the United States, has been subjected to an unusual onslaught of vehement criticism on purely methodological grounds, an onslaught that continues to this day (see, for example, American College of Surgeons, 1989).

To develop its forecast, the GMENAC simultaneously launched three distinct modeling efforts: (1) a Physician Supply Model; (2) a Graduate Medical Education Model; and (3) a Physician Requirements Model. In its principal structure, the Physician Supply Model was standard issue. It went beyond the state of the art mainly in its attempt to break down the overall supply of physicians into numerous specialties and subspecialties, which, incidentally, required the development of a separate Graduate Medical Education Model. Unfortunately, the finer the forecaster's breakdown into medical specialties, the more visible will be forecasting errors that would wash out in less disaggregated forecasts. By its bolc attempt to travel down that route, the GMENAC obviously enhanced the probability of being proved wrong, ex post, by actual events. Even so, one should not fault the GMENAC for taking this bold methodological step; one could at most fault it for setting perhaps more store by the accuracy of its supply forecasts than was warranted at the time.

The GMENAC's Physician Requirements Model represents a classic needs-based approach, although an unusually elaborate one. Through a lengthy interplay of its own analytical staff with a set of Delphi panels composed of clinical experts, the Committee projected (1) the future population of the United States, broken down by sociodemographic groups, (2) detailed information on the morbidity likely to emerge from that population, (3) the set of medical services likely to be required to
respond to that morbidity *effectively and efficiently*, and, finally, (4) the number of physicians in each specialty implied by a proper clinical response to the projected morbidity. On the basis of the requirements model, the GMENAC projected a requirement of 498,250 full-time-equivalent physicians by the year 2000. Its Physician Supply Model had projected supplies under four different scenarios, yielding numbers ranging from a low of 573,000 to a high of 684,000 physicians by the year 2000, suggesting an overall physician surplus anywhere between 75,000 and 186,000 physicians in that year (see GMENAC, 1980, fig. 2, p. 10).

Although the use of Delphi panels is sometimes frowned upon by purists in the scientific community—and they are an easy target for critics who are adversely affected by recommendations emerging from such studies—one may alternatively laud the technique as an attempt by researchers to reach out to the users of scientific research. A standing complaint among these users has been that, in structuring their analyses and in extracting policy recommendations from them, researchers too often are insensitive to the constraints, irrationality, and other imperfections said to beset the so-called "real world." A judicious use of carefully chosen Delphi panels can go a long way toward meeting that objection.

If one seeks fault with the GMENAC's Physician Requirements Model, however, one can find it in two major areas. First, in the translation of the projected need for medical services into the number of physician hours needed to meet these needs, the GMENAC implicitly assumed fixed input-output coefficients. To be sure, the Committee sought to use for this purpose what it thought of as optimal input-output ratios. Even so, the use of a fixed set of such ratios tends to obscure from view the enormous flexibility that a health system really has in responding to the health needs of a nation, a point made at length in earlier sections of this chapter. Within fairly broad limits, one can actually be more relaxed about the future health manpower situation than seems widely assumed among policymakers. Second, although the GMENAC study went boldly into the area of disaggregating the supply of physicians by specialty, it failed to extend that disaggregation to gender. Perhaps that omission is understandable within the context of a time period—the late 1970s—in which any allusion to possible differences between men and women immediately triggered charges of sexism. A disaggregation of the future supply of physicians into gender, however, is important.
The proportion of women physicians in the overall supply is growing rapidly. In 1969–70, only about 8 percent of medical school graduates were women. By 1986–87, that proportion had reached 32 percent (Marder, Kletke, et al., 1988, table 2.12). It is likely to rise further in the future. Moreover, there appear to be sizable differences in the practice patterns of men and women physicians. These differences had been noted as early as the mid-1970s (see, for example, Kehrer, 1976). In her analysis of hourly rates of patient visits among American physicians, Langwell (1982) found that, after adjusting for specialty, women physicians saw an average of only 60 to 70 percent as many patients per practice hour as did their male counterparts. Using data from the AMA's annual survey of 1981, Reinhardt (1985a) found that, on average, women internists tended to see only 81 percent as many patients per practice hour as did their male counterparts, devoted about 92 percent as many weekly hours to medical practice, and practiced only 92 percent as many weeks per year. Overall, women internists in the sample saw only about 70 percent as many patients per year as did their male counterparts.

Incidentally, similar tendencies are manifest elsewhere in the world. In their study of physicians in the Canadian province of Quebec, Berry and colleagues (1978) found that, after controlling for other practice inputs, female physicians saw 20 to 30 percent fewer patients per hour than did their male counterparts. Similar results were reported in a study of women physicians in France published by the Centre de Recherche pour L'Etude et l'Observation des Conditions de Vie (see CREDOC, 1983). Table 9-5 presents some rather striking data adapted from that study. It should be noted that physician fees are uniform across France, so that differences in patient billings cannot reflect gender-driven differences in professional fees.

The factors that have driven these observed differences in practice patterns are not yet well understood, nor is it known how stable these differences are likely to be over time. That question should receive sustained attention in health services research in the future, as the proportion of women in American medical practice grows. In the meantime, some health manpower forecasters have begun explicitly to adjust their projections for these differences. In their paper “The Projected Physician Surplus Reevaluated,” for example, Jacobsen and Rimm (1987) assert on the basis of an earlier study (Jussim and Muller, 1975) that, during their lifetime, female physicians will be only 60 percent as pro-
Table 9-5. Differentials in workload of men and women physicians in private practice, France, 1981

<table>
<thead>
<tr>
<th>Specialty and age</th>
<th>Patient billings/year (in francs)</th>
<th>Visits and consultations/year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men</td>
<td>Women</td>
</tr>
<tr>
<td>General practice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 35 or less</td>
<td>262,000</td>
<td>138,000</td>
</tr>
<tr>
<td>35 to 44 years</td>
<td>358,000</td>
<td>187,000</td>
</tr>
<tr>
<td>45 to 54 years</td>
<td>349,000</td>
<td>183,000</td>
</tr>
<tr>
<td>55 to 64 years</td>
<td>303,000</td>
<td>171,000</td>
</tr>
<tr>
<td>65 years or older</td>
<td>201,000</td>
<td>113,000</td>
</tr>
<tr>
<td>All general practitioners</td>
<td>305,000</td>
<td>157,000</td>
</tr>
</tbody>
</table>

Source: Adapted from CREDOC (1983), table 46, p. 93.

Productive as male physicians. Combined with an assumed increased proportion of female physicians by the year 2000 (they seem to have assumed 16 percent), they conclude that the physician supply projected for the year 2000 by the GMENAC must be multiplied by a factor of 0.937 just to account for the increasing proportion of female physicians in the pool. After further adjustment for an unanticipated decline in physicians entering practice and a general decline of physician productivity (calling for a further multiplicative adjustment factor of 0.940), the authors reduce the physician surplus projected by the GMENAC from 145,000 to 39,000, and even further to 6,000 if foreign medical graduates were barred from entering American medical practice after 1986 (see their Exhibit 2, p. 53).

Similarly, in their recent reevaluation of the GMENAC forecast, Schwartz, Sloan, and Mendelson reduce the projected effective number of physicians in the year 2000 by 4,000 on the assumption that women physicians will work an average of 8 percent fewer hours per year than men physicians (1988, p. 895). Remarkably, these authors do not make any adjustments for the often observed differences in the hourly visit rates of male and female physicians. Perhaps they assumed that this particular differential would vanish over time. As Frank A. Riddick, Jr., a member of the AMA's Council on Medical Education, remarked before the Council on Graduate Medical Education (established by Congress in 1985 under Public Law 99-272): “AMA data indicate that female physi-
cians have lower productivity than male physicians, working 90 percent as many hours per week in patient care activities and having 75 percent as many patient visits. However, recent studies indicate that differences in the productivity of male and female physicians have decreased in recent years” (Riddick, 1987, p. 163; italics added).

The Lingering Controversy over Medical Manpower

Since the GMENAC Report appeared in 1980, two questions have dominated the debate on health manpower policy. First, does the United States actually face a surplus of physicians? Second, even if there were such a surplus, who would actually be harmed by it? Research has not been able to answer either of these questions conclusively, although not for want of trying.

Does the United States Face a Physician Surplus?

At this time, policymakers sincerely interested in obtaining a bearing on the future supply of and demand for medical manpower in this country could forge their perceptions from a wide range of forecasts, spanning the set from a predicted overall shortage of physicians by the year 2000 to a massive and growing surplus. One of the key factors driving the differences among these forecasts has been the assumed growth in enrollment in health maintenance organizations (HMOs). As noted earlier, HMOs are able to operate with much lower physician-to-population ratios than does the rest of the health system (see especially Figure 9-2). A second major variable surrounded by great uncertainty is future growth in the demand for physician services, which, of course, depends significantly on society’s willingness to continue financing the large increase in outlays on physician services experienced in the past two decades.

In a revisit of the GMENAC forecast that he directed, for example, Alvin R. Tarlov concluded in 1986 that the physician surplus in the year 2000 would be likely to exceed even the earlier alarming forecast of a massive physician surplus (see Tarlov, 1986). Noting the rapid enrollment of the population in HMOs during the 1980s and assuming such plans to be able to operate with some 120 physicians per 100,000 enrollees (versus 225 or more in the traditional fee-for-service system), Tar-
lov projected a surplus of about 180,000 physicians if 20 percent of the population were enrolled in HMOs by the year 2000, and a surplus of 220,000 physicians if as many as 40 percent of the population were then enrolled in HMOs.

A similar prediction was made in the same year by Steinwachs and colleagues (1986). The authors concluded that, as a result of the deeper penetration of HMOs into the American health system than had been assumed in the GMENAC study, 20 percent fewer primary care physicians than projected by GMENAC would be needed for children and 50 percent fewer for adults.

At the other end of the spectrum is the recent forecast authored by Schwartz, Sloan, and Mendelson (1988), who project an impending shortage of physicians by the year 2000. These authors' forecast is summarized most easily with the help of the simple forecasting equation illustrated earlier in Figures 9-6 and 9-7. That equation is reproduced here for easier reference; its variables are defined in Figure 9-6.

\[ X_i = a_i c_i S_i - (D_i/Q_i) N_i. \]

First, the authors quite properly define the projected supply of physicians not just as those who will be professionally active in the target year (the product \( a_i S_i \) in equation [3]), but only as those likely to be active in patient care (that is, the product \( a_i c_i S_i \)). This adjustment results in a deduction of 112,000 physicians from the 700,000 or so physicians projected to be professionally active by the U.S. Bureau of Health Professions (1988b). Apparently, according to the authors, the parameter \( c_i \) (the proportion of professionally active physicians actually engaged in patient care) has been tacitly and erroneously set equal to 1 in most other forecasts.

Next, Schwartz, Sloan, and Mendelson apparently differ with others on the projected increase in the per capita demand for physician services (the growth in variable \( D_i \) of equation [3]). Most other studies posit much lower growth rates in that variable than do these authors (see, for example, the U.S. Bureau of Health Professions, 1988b).
In order to sidestep the possibility that data from the fee-for-service sector reflect utilization induced by physicians for purely pecuniary reasons—that is, in order to get as close to a true needs-based approach as they can—the authors base their projected physician requirements on the historical experience of the Kaiser Foundation Medical Plan in Southern California, which, presumably, offers physicians no incentive to provide patients any but truly needed health services. Furthermore, in order to avoid having to make explicit assumptions about future physician productivity (the variable $Q$, in equation [3]), the authors simply project the future growth in the physician-to-population ratio (that is, in the ratio $D/Q$, in equation [3]). In other words, they simply proceed with the more compact forecasting equation

$$\frac{(D/Q)_{2000}}{(D/Q)_{1983}} = (1 + g)^{17},$$

where $(D/Q)_{2000}$ denotes the required physician-to-population ratio in the year 2000, $(D/Q)_{1983}$ the physician-to-enrollee ratio in the Kaiser Plan in base year 1983 (1 physician per 840 patients or 119 physicians per 100,000 enrollees), and variable $g$ is the projected growth rate in the required physician-to-population ratio.

Clearly, the estimated future growth of the required physician-population ratio (variable $g$ in equation [5]) is the crucial driver in this forecasting exercise. It is, in effect, the difference between the projected future growth in the per capita demand for physicians' services and the projected future growth (positive or negative) of physician productivity. As such, it is a highly problematic variable, one for which it is not easy to acquire an intuitive grasp.

Using data from the period 1965–1983, and after eliminating the effect strictly of increases in the percentage of enrollees aged 65 and over in the Kaiser Plan, the authors estimate the baseline growth rate $g$ for Kaiser to have been 1.03 percent per year over the period 1965–1983 (versus 2.20 percent for the fee-for-service sector). The authors then assume that this rate of growth in Kaiser's physician-enrollee ratio will persist until at least the year 2000 and, moreover, that this growth experience will be representative of all HMOs in the United States. Finally, the authors adjust that baseline growth rate further for the growth in the percentage of the aged in the U.S. population and assume that this
percentage will apply to the HMO sector as well as to the rest of the health sector. This adjustment for the aging of the population per se brings the estimated growth rate in the required physician-enrollee ratio in HMOs (variable \( g \) in equation [5]) to 1.60 percent per year. In other words, Schwartz, Sloan, and Mendelson project that by the year 2000, HMOs in the United States will require a ratio of

\[ 156 = 119 (1 + .016)^{17} \]  

[6]

physicians per 100,000 enrollees, or about 1 physician for 640 patients.

Like Tarlov (1986), the authors assume that, by the year 2000, 44 percent of all Americans will be in tightly managed, competitive health plans—possibly an overly conservative assumption. In contrast to Tarlov, however, the authors assume that only about 20 percent of the population will be in staff model HMOs like Kaiser, and that the rest will be in independent practice associations (IPAs) and preferred provider organizations (PPOs). The latter type of arrangements are assumed to use 10 percent more physicians per enrollee than a bona fide staff model HMO, a quite reasonable and possibly conservative assumption. Finally, it is assumed that the remainder of the population will be served, in the year 2000, by the traditional fee-for-service system at a rate of 480 patients per patient care physician (or 208 patient care physicians per 100,000 population). On that basis, and after adjusting further for overall population growth (the variable \( N_t \) in equation [4] above), they project a requirement of 592,000 patient care physicians by the year 2000, against a supply of 585,000. In other words, the authors project a physician shortage of 7,000 for the year 2000.

Schwartz, Sloan, and Mendelson observe that, depending upon the particular set of assumptions one makes about the future growth in per capita demand, the number of persons enrolled in HMOs, and a number of other factors, the future supply of physicians could either fall short of projected demand by as many as 83,000 physicians or exceed that demand by as many as 40,000 physicians. This, of course, is a rather disturbing range of possibilities for a forecasting horizon of only 10 years and for a projected overall supply of somewhere around 600,000 physicians. In reacting to that wide range, however, policymakers should be mindful that not every number within a range of estimates offered by a
conscientious forecaster is believed by the author to have an equal probability of occurring. The extremes in the range offered by Schwartz, Sloan, and Mendelson clearly are believed to be improbable by these authors, even though they cannot be completely ruled out a priori. The authors' "best estimate"—one they believe to have the highest probability of occurring—is the projected shortage by the year 2000 of some 7,000 physicians.

Ernest P. Schloss (1988), in a companion piece published in the same issue of the *New England Journal of Medicine*, comes to a similar conclusion, although his forecast is much less explicit and detailed. In the light of recent socioeconomic and demographic factors, he argues, the United States is likely to face a physician shortage within the next three decades, and he warns policymakers not to overreact to predictions of an impending physician surplus.

In between these extreme forecasts are a number of others that generally point to a physician surplus. As already noted, after adjusting the GMENAC forecast for a number of factors, including the ever larger presence of women physicians in the overall supply, Jacobsen and Rimm (1987) concluded that the physician surplus by the year 2000 would be more likely in the range of 6,000 to 39,000, depending on the number of foreign medical graduates entering the physician pool. In its *Sixth Report to the President and Congress on the Status of Health Personnel in the United States* (1988a), the U.S. Bureau of Health Professions projects a physician requirement of 637,000 physicians by the year 2000 and a supply of 708,600, pointing to a surplus of 71,600 (table 2-11, p. 2–22). In its critique of the study by Schwartz, Sloan, and Mendelson (1988), the Bureau judges the growth in future demand posited by these authors (variable \( g \) in equation [5] above) to be excessively high (U.S. Bureau of Health Professions, 1988b; especially table 3) and also questions the high number of physicians assumed by the authors to be in activities other than patient care.

Finally, after reviewing all of the studies available on this topic, the Council on Graduate Medical Education concluded in its *First Report* on July 1, 1988: "From the data and testimony it has received, the Council has concluded that there is now or soon will be an aggregate oversupply of physicians in the United States. The Council notes, however, that there are significant uncertainties which could change this assessment" (vol. 1, p. xxi). But the Council hastened to add: "There is conflicting
evidence as to whether an oversupply of physicians would necessarily lead to socially undesirable consequences” (vol. 1, p. xxi). Finally, the Council recommended, wisely in my view, that “at the present time, the Federal Government should not attempt to influence physician manpower in the aggregate” (vol. 1, p. xxi).

How Harmful is a Physician Surplus?

The Council’s caveat and its recommendation raise the question: who actually is harmed by a physician surplus? Alternatively, it may be asked: are the social consequences of a surplus as serious as those of a shortage? Unfortunately, the research community has not reached a consensus on these questions, in spite of many serious efforts to do so.

In its testimony before the Council on Graduate Medical Education, the American Medical Association argued that a physician surplus could “lower the quality and raise the cost of physician services” (Riddick, 1987, pp. 164–165). Quality of care is thought to suffer when physicians performing complicated procedures do not perform them frequently enough to maintain their skill, or when specialists are driven by competition to render primary care services for which they are not properly trained. Costs are driven up, according to the AMA testimony, because there appears to be a high correlation between increases in real physician expenditures and increases in the physician-population ratio, although the AMA was quick to add that the added utilization need not be inappropriate. In any event, the AMA left the Council with the remarkable proposition that “market forces cannot be relied upon to adjust the growth rate of the U.S. physician supply, because the U.S. health care system does not operate in a free market” (Riddick, 1987, p. 170) and vaguely hinted at measures to stem the rise in the number of American physicians without actually calling for direct controls on that growth.

Within the research community itself, the frequently observed positive correlation between the physician-population ratio and per capita spending on physician services has engendered one of the most tenacious, heated, and, so far, futile debates. One school of thought (for example, Evans, 1973, 1974; Reinhardt, 1978, 1985b; Rice, 1984; and, apparently, Sloan, 1983, 1984) sees in that correlation support for the intuitively appealing hypothesis that physicians respond to the economic
pressure triggered by higher physician density by inducing in their patients a higher level of demand for physician services than would genuinely be demanded by fully informed patients technically capable of evaluating the medical merits of their care. That excess utilization, according to this school of thought, is not only costly to the rest of society; it can also be harmful to the patient's health. Other researchers (Sloan and Feldman, 1978; Harris, 1986; Rossiter and Wilensky, 1987; Feldman and Sloan, 1988) have argued, however, that a mere correlation of this sort implies nothing about causation and would, in fact, be data-compatible with hypothetical models that explicitly preclude the physician's ability to induce demand for his services at given levels of fees faced by patients. It is hard to argue with that objection, of course, although by itself such theoretical argumentation clearly cannot not settle the matter conclusively.

Econometric studies designed to test these rival hypotheses with appeal to nonexperimental data have so far not been able to settle the matter one way or the other either, and for a very good reason: such analyses have not been able to distinguish properly between physician services merely accepted by often frightened and ill-informed patients, and services actively demanded by the patient. Although it is plausible to argue that initial patient visits to a physician are purely patient-initiated, that argument does not extend to the many ancillary services that can be packaged by a physician into such a patient-initiated visit. As Mark Pauly has wisely concluded on this point: "For many reasons it is likely that we [economists] will never fully resolve the demand creation/information imperfection question... Thus the fee-for-service market will remain a mystery as far as empirical predictions are concerned" (1988, p. 229).

Unwilling to wait for a resolution of this issue by the experts, and possibly quite innocent of this learned debate, policymakers almost everywhere on the globe appear to have let their policies on this matter be guided by their own intuition. That intuition appears to have sustained the thesis that physicians can, indeed, create demand for their services within broad limits, and that they will usually do so to counteract downward pressure on their incomes. Upon that thesis, for example, has rested the demand by the White House and by the Congress, in 1989, that the introduction of the Resource-Based Relative Value Scale (RBRVS) be accompanied by a global cap on Medicare's expenditures on
physician services. That cap is meant to control what is euphemistically
called the “volume problem,” which however is clearly understood by all
to refer to the likelihood that surgeons and other proceduralists will
react to the impending reduction in their fees through increases in the
recommended use of their services. Thus, policymakers tread boldly
beyond the point on the road at which health services researchers
remain locked in erudite disputation.

Is Health Manpower Forecasting Worth the Trouble?

It has been the purpose of this chapter to describe to the noninitiated
the task of health manpower forecasting, in theory and in practice, and
to assess the role of that research effort in the formulation of the nation’s
health policy. The second section of the chapter was intended to convey
to the reader the sheer complexity of this research enterprise. In effect,
the forecaster who seeks to project the size of the physician surplus or
shortage in some future target year attempts nothing less than project-
ing for the user of that information the conditional mean (or modal value)
of a fairly wide probability distribution of possible values for the variable
“projected supply of physicians minus projected physician requirements
in future year t.” The shape and position of that probability distribution
are conditional upon a whole host of influences that must, first, be quan-
tified and, second, projected into the future as well. In short, the task
is daunting.

It is no more daunting, however, and no more unsuccessful in prac-
tice, than the labor market forecasts regularly issued by the U.S. Bureau
of Labor or the macro-economic forecasts regularly issued by the Office
of Management and Budget and the Congressional Budget Office.
Indeed, the task is no more daunting, nor any less successful in practice,
than the annual financial accounting reports and the financial forecasts
upon which modern business firms base their strategies and upon which
pension and mutual funds place their billion-dollar bets.

Time after time, when health manpower forecasters in the past have
submitted precise point estimates of the future health manpower situa-
tion, they have found themselves to be widely off the mark, ex post. On
the other hand, when forecasters have sought to cover their flanks, ex
ante, by hedging their point estimates with entire ranges of possible,
alternative scenarios, these ranges have been sufficiently wide to frus-
trate policymakers. It may therefore fairly be asked whether the benefits from health manpower forecasting warrant its cost at all.

In exploring that question, one will soon discover that the cost of the enterprise has actually been rather modest by comparison with other efforts to structure information for decision makers. The highly elaborate GMENAC exercise, for example, probably cost no more than $3 million or so, a number that pales in comparison with the vast sums spent annually by the typical American corporation to produce its financial reports. As I have argued elsewhere, these corporate reports are much cruder and certainly no more reliable than even the crudest health manpower forecasts produced by health services researchers (Reinhardt, 1981, pp. 1155–56).

In the New York Times of March 22, 1990, for example, it was reported that “the Shawmut National Corporation [a bank] said yesterday that after examination by Federal banking regulators it had revised its 1989 earnings to show a loss of $128.9 million. In January, it had reported a profit of $201.7 million” (p. D1). Not in the entire history of health manpower forecasting in the United States have researchers ever been so widely off the mark with their numbers—even for twenty-year forecasts—than were this bank’s highly paid accountants and auditors. Yet major revisions of this sort are not at all uncommon among the people who structure financial information for the business world, information upon which powerful decisions may be based. To make this point even more forcefully, one may ask whether there is in the history of health manpower forecasting anything as egregiously wanting as the work of the high-priced accountants who produced, and the high-priced auditors who judged as “fair and accurate,” the financial reports of, say, the Lincoln Savings and Loan Association—reports that seduced the present Chairman of the Federal Reserve and former economic consultant Alan Greenspan to pronounce that institution safe, sound, and beyond the need for government regulation, all at an eventual cost to the American taxpayer of some $2 billion or so? Or, are health manpower forecasts usually any wider off the mark than are the longer-range macroeconomic forecasts of the U.S. Treasury Department or the President’s Council of Economic Advisors?

In placing health manpower forecasting into this wider perspective, the issue of its cost quickly loses significance. There remains the question, however, whether the enterprise yields any benefits at all, or whether it possibly might do more harm than good. To think about this
question, one might open for health manpower forecasting an imaginary T-account and credit to it the good thought to flow from that research, while debiting to the account the potential pitfalls of the effort. Upon completion of that exercise, one may conclude that the negative side of health manpower forecasting lies not so much in the production of the forecasts themselves, but in the use to which such forecasts have been put at the level of policy.

The first credit one would enter for health manpower forecasting is the routine baseline inventorying of the already available health manpower that the exercise forces upon the researcher. These inventories by themselves should be useful information bases for policymakers. Next, there is surely merit in having available even simple, computer-based models that permit one to explore "what if" questions through simulation exercises. Coupled with good human judgment on what scenarios are more probable than others, such exercises can identify future pressure points that can be eliminated through judicious policy action. Third, the exacting requirements of health manpower forecasting serve as a perennial incentive to the health services research community to produce empirical information on the many behavioral links illustrated earlier with the aid of Figure 9-5. Information on these behavioral links is important in many policy initiatives other than forecasting per se.

Finally, health manpower forecasting could become potentially more useful if it shifted its focus away from the traditional, narrow preoccupation strictly with numbers of physicians to what Jeffrey Harris (1986) has called "intermediate targets"—time series on physician incomes and the rate of return to medical education, on wait-times to appointments and other measures of access, on the bookings of physician-recruitment firms, on rates of incidence of surgical procedures, and so on. In other words, Harris really calls for the metamorphosis of health manpower forecasting per se into a better coordinated discipline of health manpower analysis.

Health manpower forecasting can be less than helpful if the forecasts are carelessly rendered or communicated to the users—for example, if simple point estimates are proffered as gospel truth, without the necessary caveats, or if the projections are deliberately biased to trigger a desired policy action. And even carefully rendered forecasts can do harm if policymakers place undue reliance on numbers and fail to temper these numbers with other information, experience, and good judgment. The history of health manpower forecasting during the 1950s and 1960s sug-
gests that the forecasts produced then may have been too simplistic and too alarmist, possibly seducing policymakers into an excessively expansionist health manpower policy (although, as noted, there were other powerful forces pushing in that direction).

One may describe health services research as an attempt to enhance the accuracy of the folklore upon which policymakers base their decisions. Properly structured, conveyed, and used, health manpower forecasting is one bit of the vast information needed to improve that folklore, but it will never be the one sufficient bit of information, and it will never replace that mysterious ingredient fundamental to all good policymaking: judgment. As Eli Ginzberg (1989) has sagely observed on this point: “In a pluralistic society such as the United States, which does not have a national system of health care, it may be futile to pose the question whether the ration will have too many, too few, or just the right number of physicians a decade or two in the future. All we can hope to do is to address selected facets of the supply problem as they force themselves onto the nation’s agenda. To do more is likely to lead to frustration; to do less is to stockpile problems for the future.” Well said!
9. Health Manpower Forecasting


