

CHAPTER 1

HEALTH AND THE USE OF HEALTH CARE

INTRODUCTION: THE SCOPE OF HEALTH CARE ECONOMICS

The health of people, individually and in groups, is influenced by a very wide range of factors, of which health care is only one. Moreover, while health care may be decisive for the health of a particular individual in particular circumstances, it is generally accepted that for most people, most of the time, health status is primarily dependent on sanitation, diet, shelter -- the complex of factors summarized in the federal White Paper of 1974 (Lalonde) as "lifestyle" and "environmental." And the great historical improvements in mortality and morbidity, life expectancy and health status experienced in now-developed countries appear to owe much more to improvements in these areas than to the progress of health care narrowly defined. "Tell me how a people die, and I will tell you how they live."

The economics of health, which Culyer (1981) defines as the application of the discipline and tools of economics to the subject matter of health, accordingly encompasses the full range of two-way causal relations between the health status of individuals and groups and their economic activities -- production, distribution, and exchange. This might include, for example, the study of the impact of unemployment and bankruptcy rates on anxiety and mental health, or of anti-malarial campaigns on agricultural labour productivity, or of the interaction between patterns of income or wealth distribution and health status.

The economics of health *care*, which is the subject of this book, is only a part of this much broader field. It restricts attention to a particular set of goods and services which have somehow been identified as having a special relationship to health status, and to the activities associated with their production and consumption. Like any other branch of economics, it studies the processes and institutions which govern the allocation of scarce social resources to these activities, the choice of techniques of production and mix of outputs, and the distribution of health care among the end users -- consumers or patients.

But unless this special relationship, and health status itself, are defined narrowly, the economics of health care can easily become the economics of everything. The famous World Health Organization definition of health, "a state of complete physical, mental, and social well-being," not "merely the absence of disease or infirmity," is a splendid call to arms, but bears a striking resemblance to the economist's concept of utility or welfare. The only difference is that utility is unbounded, being constrained only by the resources available in particular circumstances. Health, on the other hand, is implicitly bounded by some state of "complete ... well-being," beyond which one cannot go; but we are not, in fact, expected to reach that limit, in this life at least.

But if "health" is equated with welfare or well-being, then clearly there is no sphere of human activity and, in particular, no form of economic activity which does not have health as its principal concern. Economics is frequently defined as the study of the allocation of scarce resources among competing wants so as to "maximize" (in a way left undefined) the satisfaction of those wants. But if all wants are encompassed in health, then all economics is health economics.

The attractions of disciplinary imperialism are tempered by awareness of the dental and digestive problems which result from biting off more than one can chew. Accordingly, insofar as they have considered it at all, health economists have generally accepted, explicitly or implicitly, a narrower conception of health status as "absence of disease or infirmity." This restriction, however, does not suffice to identify health care as a set of commodities, or activities, which are uniquely related to health. Since sanitation, food, home and work environment, activity levels, play at least as important a role as health care in influencing the absence of disease or infirmity, it is clearly impossible to define health care as all activities or commodities influencing health, even narrowly defined, without slipping back into the "economics of everything."

The economic literature appears to display two different responses to this difficulty. One approach is simply to sidestep the definitional issue, and to accept as the proper content of health care economics whatever conventional usage has come to label health care -- hospitals, doctors, drugs, and all that. Health care economics studies the behaviour of health care providers and users, whoever those may be. The development of a sub-discipline differentiated from economics in general then presumably rests on certain characteristics peculiar to the processes of resource allocation in the production and distribution of these commodities. But there is no explicit recognition of any special relationship of health care to health.

The weakness of this approach, apart from the problems of identifying boundaries which are common to any definition, is that it tends to blur important distinctions between health care and commodities in general. It thus provides no logical grounding for the characteristics which make health care "different" and which motivate not only a special sub-discipline of economics but also, in almost every society, the very special set of institutions which govern and regulate its production and distribution. Why are some activities health care, and others not, and why are the former treated differently?

The alternative, which is followed in this book, is to define health care as that set of goods and services which consumer/patients use solely or primarily because of their anticipated (positive) impact on health status. It is health, as a status, rather than health care, as a commodity, which is of value to its users. Indeed, the direct effects of most, if not all, health services on their users' well-being, independent of their anticipated health effects, is negative. Dentistry, drugs, diagnostic and therapeutic interventions, hospital stays, are frequently uncomfortable or frightening in and of themselves. Few, indeed, would choose to purchase them in the absence of an expected health benefit. And those who do -- drug addicts, say, or sufferers from Münchhausen's Syndrome, who derive pleasure from medical care and counterfeit illness to get it, are generally regarded as unwell. Anyone who seeks care when he is not sick, is sick.¹

This formulation does not, of course, exclude preventive services from health care. Users of such services may be healthy at present, but they are obviously contemplating a possibility of some future deterioration in their health status. Preventive care must be believed to reduce the probability or severity of such deterioration, or it is not preventive care. Nor do we exclude the obvious variations in amenity levels surrounding various forms of care, from spartan to luxurious, assembly-line to humane and personal. There are different degrees of satisfaction or dissatisfaction associated with health care, given that one is ill (or more generally believes that a particular form of health care is needed to maintain or improve health). What our formulation implies is that faced with a choice between illness plus any level of associated care, and not being ill at all, the representative consumer prefers not to be ill. Further, if ill, she regards health care which is not expected to improve health (e.g., interventions for teaching purposes) as a

"bad," not a good. And amenities for which this is not true -- colour T.V. sets or steak-and-champagne dinners in hospitals with low occupancy rates -- are not health care.

Health care is thus clearly demarcated from other commodities which influence health, but which are consumed for their perceived direct satisfactions rather than their health effects. Nutritional value may be a consideration in food purchasing, but if it were the only or the predominant consideration we would eat much more skim milk powder and beans, and much less steak. Indeed some foods, and particularly such substances as tobacco and alcohol, are consumed despite their known harmful effects on health.² But anyone using health care which she believed to be harmful to health would be thought irrational, to say the least.

Explicit recognition of the role of health status in the definition of health care will provide significant assistance in our understanding of why health care as a commodity is "different," and why societies develop peculiar types of institutions for directing its production and distribution. Moreover, health status, unlike utility or well-being, is identifiable and to a degree at least measurable by observers external to the consumption process. It thus provides a standpoint and a source of additional information for the evaluation of resource allocation processes in health care, which is lacking in economic analysis in general. The economic analysis of health care can draw on the rich medical literature on efficacy and effectiveness which focusses on the relation between health care and health status. We shall show, throughout this book, that such a linkage powerfully extends the range of economic analysis and permits it to grasp important dimensions and issues of health care policy and institutional design which are inexplicable, and ignored, in a framework which treats health care as similar to any other commodity, assumed to bear directly (and positively) on the user's utility or well-being without the necessity of any mediating concept of health status.

In particular, explicit recognition of the mediating role of health status enables us to come to grips with the concepts of over- and under-utilization, which play an important role in the medical literature and in health policy discussions. These are defined in terms of the relation between health care and health status for particular consumer/patients. Use of amounts or types of care which an informed provider might reasonably expect to do a patient no good, or even harm, clearly represents over-utilization, from a medical standpoint. So, as we shall show below, is care with an expected payoff in terms of the user's health status, which is positive but too small (on some cut-off criterion which must be supplied externally) to justify its cost. Under-utilization, by contrast, implies that some amounts or types of care are not being supplied and used, which could increase the health status of some patient(s) by enough to justify their cost. Inappropriate care can involve simultaneous over- and under-utilization -- too much of some forms of care, not enough of others -- again in terms of expected effects on the user's health status.

These rather obvious, but very important, concepts cannot even be expressed in the standard economic framework which treats health care itself as the elementary object of choice for consumer/patients, and then goes further to impose the standard non-satiation postulate that "more is always better."³ There is a concept of "over-utilization" in this framework, but it is defined as a level such that the (assumed positive) benefits of care at the margin, defined in terms not of health increment but of consumer satisfaction, fall short of its resource cost. This is in general unrelated to "over-utilization" or "under-utilization" as discussed in the medical literature, except under very special circumstances. The semantic confusion between the two, particularly in the United States, has led not only to imperfect communication between

economists and other health-care analysts, but also to some rather serious confusion in health policy and the design of health insurance systems.

The relationship between health status and health care also plays a central role as an organizing concept in the analysis of the regulation of health care providers. Such regulation derives its principal *raison d'être* from the fact that the provider of health care, as individual or organization, is expected not merely, or even primarily, to satisfy her customers. She is required to seek specific outcomes, in terms of improvements in patient health status, and access to professionalized occupations or markets is conditional upon demonstrated ability and willingness to behave in ways believed consistent with these objectives.

On yet another level, the whole field of cost-benefit or cost-effectiveness analysis applied to public projects in health care is focussed on health status outcomes. The question which must be answered for any such program or intervention, before any economic analysis can be carried out, is does it work? Is it efficacious and effective for the purposes claimed? (Sackett 1980). If the program yields no improvement in anyone's health status, it is not worth buying at any (positive) price. The same is true, on our definition, for any other form of health care.

THE HEALTH CARE "INDUSTRY" IN CANADA: A BRIEF SKETCH

The collection of diverse organizations which assemble resources and produce and distribute health care is referred to as the "health care industry." The label often strikes practitioners or non-economist students of health care as odd, or even offensive; but in fact it carries no presumptions as to motivations or forms of organization, no necessary implication of belching chimneys, assembly lines, or concentration on "the bottom line." The health care industry does include privately owned, strictly for-profit corporations, of course, but it also includes professional practices, non-profit organizations such as hospitals and voluntary societies, and government departments and agencies. The range of motivations, forms of organization, and technologies (in a broad sense) is very wide; but all use up scarce resources of human time and skills, services of capital (buildings and machines), and raw material to produce or distribute goods and services whose sole or primary purpose is the improvement or maintenance of someone's health.

The structure and evolution of the Canadian health care industry is displayed at a very aggregate level in Table 1-1, which shows total expenditure on health care by category over the post-war period both in dollars and relative to Gross National Product. These expenditures are frequently referred to as "health care costs" and are alleged at various times to "spiral," to be "explosive," or to do other peculiar things. From another perspective, however, these data describe the sales of the health care industry to Canadians, by product line. They are total revenues, not costs of production. This dual nature, with each expenditure item being simultaneously a revenue item to someone else, has important implications for both interpretation and policy. Health care "expenditures" are costs to the rest of Canadian society, not to health care providers -- to the providers they are income.

TABLE 1-1
Expenditures on Health Care in Canada, 1946-82, by Major
Component (\$Mn) and as Percentage of Gross National Product
(Bracketed Figures)

	Hospitals	Physicians' Services	Dental Services	Prescribed Drugs	Personal Health Care¹	National Health Expenditures
1946	150.7 (1.27)	86.7 (0.73)	36.3 (0.31)	26.8 (0.23)	300.5 (2.53)	n.a. --
1951	326.4 (1.51)	153.0 (0.71)	51.0 (0.24)	42.9 (0.20)	573.3 (2.65)	n.a. --
1956	541.5 (1.69)	240.0 (0.75)	81.5 (0.25)	71.8 (0.22)	934.9 (2.92)	n.a. --
1961	949.0 (2.39)	388.3 (0.98)	116.7 (0.29)	135.8 (0.34)	1,589.9 (4.01)	2,375.50 (6.00)
1966	1,668.8 (2.70)	605.2 (0.98)	176.4 (0.29)	232.0 (0.38)	2,682.3 (4.34)	3,837.50 (6.20)
1971	3,152.8 (3.33)	1,250.4 (1.32)	311.5 (0.33)	402.5 (0.43)	5,117.2 (5.41)	7,122.30 (7.50)
1976	6,571.5 (3.42)	2,103.2 (1.10)	699.8 (0.37)	667.1 (0.35)	10,041.6 (5.24)	14,158.70 (7.40)
1977	6,928.3 (3.30)	2,309.0 (1.11)	827.6 (0.40)	746.0 (0.35)	10,810.9 (5.15)	15,532.60 (7.40)
1978	7,483.5 (3.23)	2,544.0 (1.10)	954.1 (0.42)	822.2 (0.35)	11,803.8 (5.10)	17,094.10 (7.40)
1979	8,239.5 (3.12)	2,843.5 (1.08)	1,106.0 (0.42)	918.2 (0.35)	13,107.2 (4.96)	19,067.20 (7.20)
1980	9,484.7 (3.20)	3,284.7 (1.11)	1,288.0 (0.43)	1,011.2 (0.34)	15,068.6 (5.08)	22,178.60 (7.50)
1981	10,724.4 (3.16)	3,741.0 (1.10)	1,482.9 (0.44)	1,205.0 (0.36)	17,153.3 (5.06)	25,769.30 (7.60)
1982	12,470.0 (3.50)	4,414.3 (1.24)	1,682.6 (0.47)	1,473.4 (0.41)	20,040.3 (5.62)	30,087.70 (8.40)

SOURCES: See Data Sources Appendix.

¹ "Personal Health Care" is the sum of the first four columns: -- hospitals, physicians' and dental services, and prescribed drugs. The more inclusive series of National Health Expenditures adds to these expenditures for other forms of (health-related) institutional care, services of other self-employed professionals, non-prescribed drugs, eyeglasses and other appliances, and costs of public health, research, capital investment, education, and administration of prepayment plans. The more comprehensive data were not compiled prior to 1960.

As the table indicates, these expenditures rose rapidly, both in dollar terms and as a percentage of total national expenditure, during the quarter century from 1946 to 1971. Expenditures on hospital care made up the largest share of the total, and also showed the most rapid growth, but physicians and prescription drug expenditures also made substantial gains. Overall, health care more than doubled its share of a national "pie" which was itself growing very rapidly during this period. Real national product (adjusted for price change) per capita rose 87.6 percent from 1946 to 1971, or 2.55 percent per year.

During the 1970s, rapid growth of health spending continued in dollar terms, but its share of national expenditure fluctuated in a range between 7 percent and 7 1/2 percent. The large numbers involved supported perceptions of "spiralling" costs, but like all economic measures, these were distorted by the inflation of the 1970s. It is clear that after 1971, the first year of complete nationwide public insurance coverage, the trend in health care expenditures/health industry sales changed from growth in share of national output to a roughly stable share. Hospital care and dental service costs increased their shares somewhat during the first half of the decade, while physician and drug expenditure shares fell back significantly. In the late seventies, hospital spending rose more slowly than GNP, while dentistry continued to expand its share, and the other two sectors were stable. But the share of health costs outside these areas began to grow, reflecting the expansion of public nursing homes and long-term care. This sector is likely to show continued growth, and substitutes to some extent for hospital services.

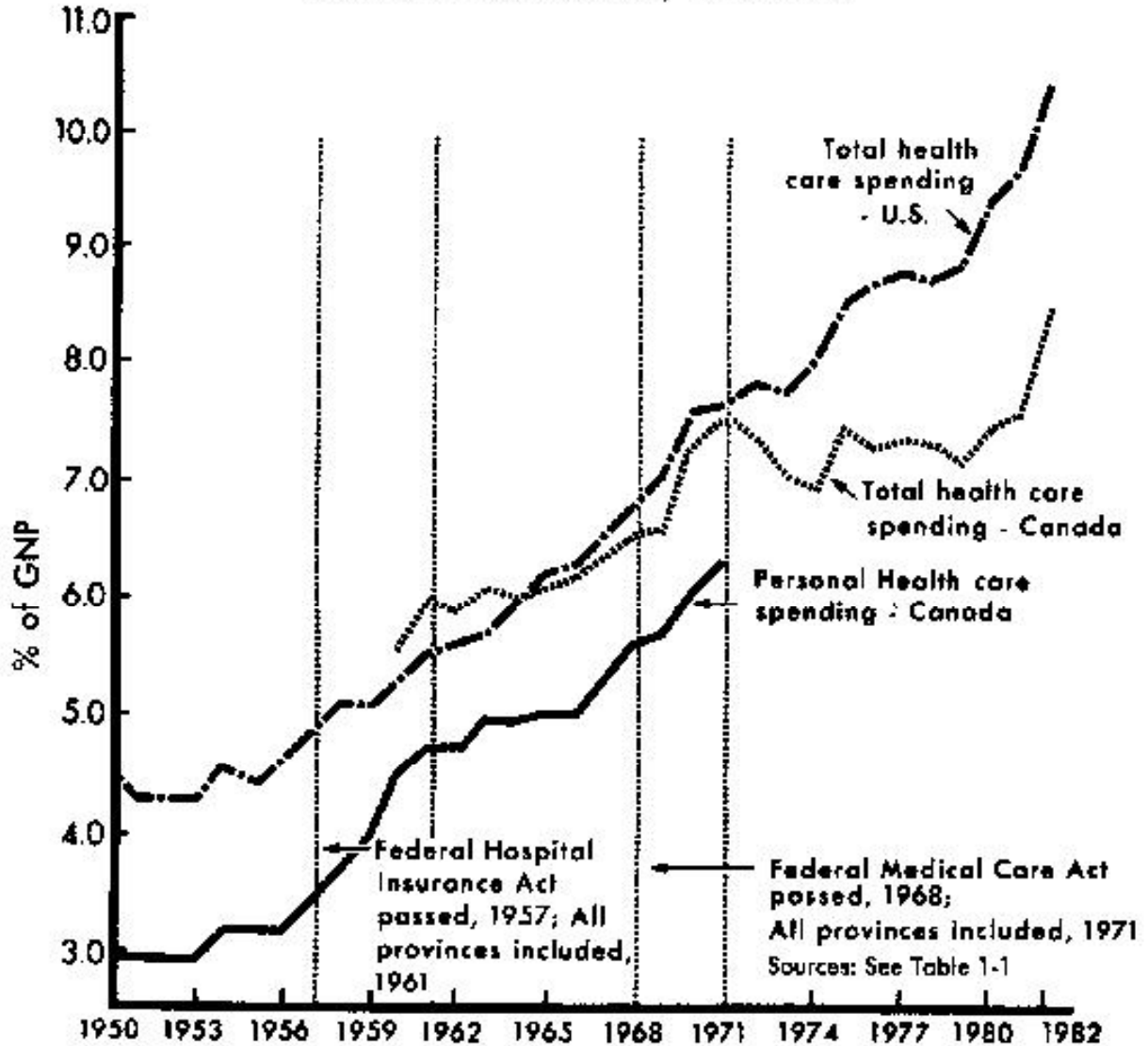
The significant effects of the public hospital and medical insurance programs on this cost experience are clearly shown in Figure 1-1, which displays health costs relative to GNP for both Canada and the United States from 1950 to 1982. It appears that costs escalated during the actual introduction of each program, but that the completion of universal public coverage in 1971 initiated a period of stability, in sharp contrast to previous Canadian or concurrent United States experience. Data for the early 1980s show a resurgence of growth in the health care share of GNP in both countries, though less pronounced in Canada. Deep recession in the general economy in 1982 has sharply accentuated this trend. The outlook for the 1980s, however, is a more appropriate topic for the end of this book than for the beginning.

Although the data in Table 1-1 are useful as background information about where we are and where we have come from, there is actually much less to them than meets the eye. The same data can be, and have been, used to support arguments by the Canadian Medical Association to the Parliamentary Task Force on Fiscal Arrangements (Canadian Medical Association 1981) (and elsewhere) that the health care system is "underfunded," and by various provincial Ministers of Health, that costs are "spiralling out of control." By themselves, they say nothing about whether the share of national resources devoted to health care is too high or too low, whether the mix of different types of health care services is appropriate, whether production is technically efficient or wasteful, or whether the services produced go to the right people -- whoever they are. In fact, in the absence of price information, the data of Table 1-1 do not necessarily tell us about quantities of health services at all.

As noted above, health care "costs" as usually defined are total industry sales, and total incomes earned from health services production. This identity relationship, displayed in Figure 1-2, is merely a sub-component of the overall national income accounting identity which equates Gross National Income, Expenditure, and Product. Sales or expenditures are in turn the product of the total number of items produced, goods and services, represented as a vector or a "shopping list" of quantities of different kinds of items, multiplied by their respective prices. And total

FIGURE 1-1

Health Care Spending as a Percentage of National Income, Canada and the U.S., 1950-1982



incomes are the product of the numbers of different types of persons drawing incomes from health services provision, multiplied by their average incomes (from this industry). Thus:

$$P_1Q_1 + \dots + P_iQ_i + \dots + P_nQ_n = N_1Y_1 + \dots + N_jY_j + \dots + N_mY_m$$

where Q_i is the quantity of health care of type i produced in a given time period, office visits, say, or appendectomies, MOD restorations or aspirin tablets, and P_i is its average price. N_j is the number of people of a particular type drawing incomes from health care, and Y_j is the average income earned.⁴

These incomes are earned by supplying resources -- labour, capital, raw materials -- to health care production, and it is these resources which are the true "costs of production" of health care. They also give rise to "opportunity costs," in that resources (by definition) have alternative uses in the production of other commodities or services of value, and are scarce, insufficient to serve all the alternative valued uses to which they might be put. Thus the opportunity cost of health care, or of anything else, is the foregone opportunities, the other things which might have been, but were not, done with the same resources of human time and skill, bricks and mortar, equipment and raw materials.

Seen from this perspective, an increase in health care expenditures which was the result of, say, a 50 percent increase in all prices of all health care, goods and services, which in turn raised all incomes in that sector by 50 percent, but which left resource flows into (principally people working in) the industry and commodity production from it unchanged, would *not* raise the cost of health care to Canadian society as a whole. No more other things -- shoes or ships or sealingwax -- would be given up to produce more health care. Of course, the distribution of the total social product would change, with a larger share going to health care workers or other resource suppliers, and a smaller going to everyone else. A variety of secondary effects might follow from that, as more people tried to enter health care jobs, governments raised taxes, borrowed, or cut expenditures to meet the new costs, private insurers raised premiums, or patients reacted to higher out-of-pocket costs. But in the first instance, if one assumes a time before any such reactions had occurred, what would be recorded as a 50 percent rise in "health costs" would be merely a transfer of wealth from one group of Canadians to another (with some foreign participation through corporations).

The economist's focus on real resources is, of course, paralleled by the health care analyst's interest in real outputs. The increase in "health costs" from the 50 percent price/income hike has no effect on the volume and mix of health care produced either. Unless one is prepared to speculate on a placebo effect from high-priced care, it follows that the increased expenditures will have no effects on anyone's health status. Again, of course, a variety of secondary effects (patients unable to pay increased prices, new suppliers attracted into the industry) could lead to changes in the pattern of the Q_i and N_j which did have health effects. But there is no automatic link between changes in health "costs" as currently measured, and health care. Accordingly, an "underfunding" argument, unless merely a statement about relative incomes, must focus on the flow of new resources into health care and resulting larger volumes of output.

That this is not merely an abstract or theoretical issue is demonstrated by Table 1-2, which assembles some data on real outputs of the different components of health care over the post-war period. Ideally, we would like consistent, long-term sector-specific price indices to deflate the figures in Table 1-1 to constant-dollar values. But with the available data, only rather imperfect indices can be developed, particularly for earlier years, and work in this area is still incomplete (Barer and Evans 1983). But the Table 1-2 data do indicate some sharp divergences between trends in health care "costs," and in service flows.

TABLE 1-2
Hospital Utilization and Wages; Physician Availability and Incomes,
Canada, 1946-1982
(Bracketed figures are average annual percent growth since last figure.)

	Hospital Inpatient Days per 1000 pop.	Index of Hospital Services per Patient Day	Hospital Relative Wages Index	Physicians per 1000 Population	Physician Relative Incomes
1946	1,334.7	48.8	n.a.	1.044	3.26
1951	1,418.9 (1.23)	53.4 (1.82)	63.1 --	1.023 (0.41)	3.14 (0.93)
1956	1,578.1 (2.15)	58.9 (1.98)	70.6 (2.27)	1.111 (1.66)	3.54 (2.43)
1961	1,639.5 (0.77)	69.2 (3.28)	85.2 (3.83)	1.167 (0.99)	3.94 (2.16)
1966	1,793.9 (1.82)	85.3 (4.27)	93.2 (1.81)	1.325 (2.57)	4.83 (4.16)
1971	1,896.6 (1.12)	100.0 (3.23)	100.0 (1.42)	1.517 (2.74)	5.57 (2.89)
1976	1,954.7 (0.61)	112.0 (2.29)	114.9 (2.82)	1.733 (2.70)	3.83 (7.22)
1977	1,976.6 (1.12)	111.8 (0.18)	114.6 (0.26)	1.767 (1.96)	3.68 (3.92)
1978	1,996.0 (0.98)	n.a. --	n.a. --	1.786 (1.08)	3.43 (6.79)
1979	1,986.9 (0.46)	117.6 (2.56)	112.4 (0.96)	1.805 (1.06)	3.43 -
1980	1,986.8 -	119.7 (1.79)	116.9 (4.00)	1.828 (1.27)	3.36 (2.04)
1981	1,972.4 (0.72)	122.5 (2.34)	119.1 (1.88)	1.859 (1.70)	3.28 (2.38)
1982	1,957.6 (0.75)	123.0 (0.41)	123.7 (3.86)	1.912 (2.85)	n.a. --
Average Annual Growth Rates					
1946-56	1.69%	1.90%	--	0.62%	0.92% *
1956-66	1.29%	3.77%	2.82%	1.78%	3.16%
1966-76	0.86%	2.63%	2.12%	2.72%	-2.29%
1976-82	0.02%	1.57%	1.24%	1.65%	-3.05% **

SOURCES: See Data Sources Appendix.

* 1947 data

** 1976-1981 only

NOTE: Hospital days are public general and allied special; the index of hospital services per patient day is expenditures deflated by an index of input prices; hospital relative wages are an index divided by an index of average weekly wages and salaries. Physician incomes are from Taxation Statistics, relative to average taxpayers, all taxable returns, and may be downward-biased after 1976; see Data Sources Appendix. Relative wage data prior to 1961 are particularly shaky, being unadjusted for changes in skill mix or hours of work, and should be treated as impressionistic only.

The rapid increase in hospital costs in the 1950s, for example, was associated with increasing utilization rates (patient days per thousand population) and intensity of servicing (principally increases in person-hours per patient day, but also more highly trained personnel, and other inputs). In the 1960s, utilization increased at a slower rate, intensity of servicing per day increased rapidly, and relative earnings of hospital workers increased at a more rapid rate. In the 1970s, utilization is almost static, intensity is growing very slowly, and relative earnings, after a burst early in the decade, are almost flat after 1976.

For physicians' services, the changes are even more dramatic. In the period 1947-1971, the income of the average physician rose from 3.26 times that of the general taxpayer to 5.57 times. To some extent this could result from increasing productivity, but an improvement in relative income status requires productivity gains faster than the general average, which, as noted, was itself quite rapid in that period. A large part of the increase in expenditure on physicians' services from 1946 to 1971 was, in fact, a price/income phenomenon, which did not increase the "opportunity cost" of medical care or have any effect on anyone's health status, but merely transferred wealth from Canadians generally to physicians.⁵ The actual increase in physicians per capita, which might be expected (apart from variations in working hours or numbers of other associated personnel) to translate into increased medical care use, is relatively slow until the 1960s, and then becomes very rapid,

In the early 1970s, numbers of physicians and medical care utilization per capita continued to increase rapidly. The federal government's *Medical Care Annual Report* (1981) indicates that from fiscal year 1971-72 to 1978-79, the average annual increases in physicians per capita and utilization per capita were 2.8 percent and 4.6 percent, respectively. "Costs" stayed stable over that period, however, because service prices and physician incomes (adjusted for inflation) were falling. Annual rates of medical fee increase are estimated at 4.9 percent, compared with the Consumer Price Index increase of 8.3 percent. The 1.7 percent annual increase of utilization per physician still left revenues per physician falling in real terms, and falling even more sharply relative to other workers. Table 1-2, drawing on taxation data, shows the same process in a longer-term context. The rapid increase in relative incomes of physicians started well before Medicare and, indeed, was most pronounced in the early 1960s. And the very sharp reversal of the early 1970s had moderated, but not ended, by 1981.⁶ The stability of physician expenditures as a share of GNP since 1971 thus masks an increase in their real resource cost, more skilled manpower employed, as well as an increase in the real volume of services supplied. In the previous two decades, by contrast, the increase in opportunity cost and service volumes is greatly overstated by the expenditure data.

FROM DATA TO EVALUATION: HOW MUCH IS ENOUGH?

Even if our health expenditure data were "perfectly" adjusted for price change, however (sidestepping some important ambiguities in that process), they would still fail to answer questions about underfunding or over-utilization, about the appropriateness of care patterns, or about technical efficiency versus waste. The fundamental problems of resource allocation remain: How much of what sorts of resources to allocate, to the production of which forms of health care, how, and for whom? A common response by health care practitioners is that services

should be provided to meet the "needs" of people in the community, either as expressed by people themselves, or as interpreted by practitioners.

Economists, on the other hand, conceive of allocation questions in terms of a trade-off or balance between the values people attach to the production from resources in any particular use, and the opportunity cost of such production. The "right" allocation to health care depends both on the perceived payoff to more health care and on the values attached to the other outputs foregone -- roads, steak dinners, sewage plants, housing, fighter planes, symphony orchestras -- because resources which could have been used to produce these things are instead used to produce health care.⁷ So long as a society's preferences, however defined, are such that additional health care of some sort is valued more than its cost in terms of foregone opportunities, then the output of that form of care should be increased. If, on the other hand, it is believed that at the current level of provision a reduction of care would reduce well-being by less than the gains from redeployment of the released resources in some other area of production, then the health sector (or part of it) is too large. The "right" allocation is that at which the value of additional resources is perceived to be equal, whether used in health care production or in the next best opportunity.

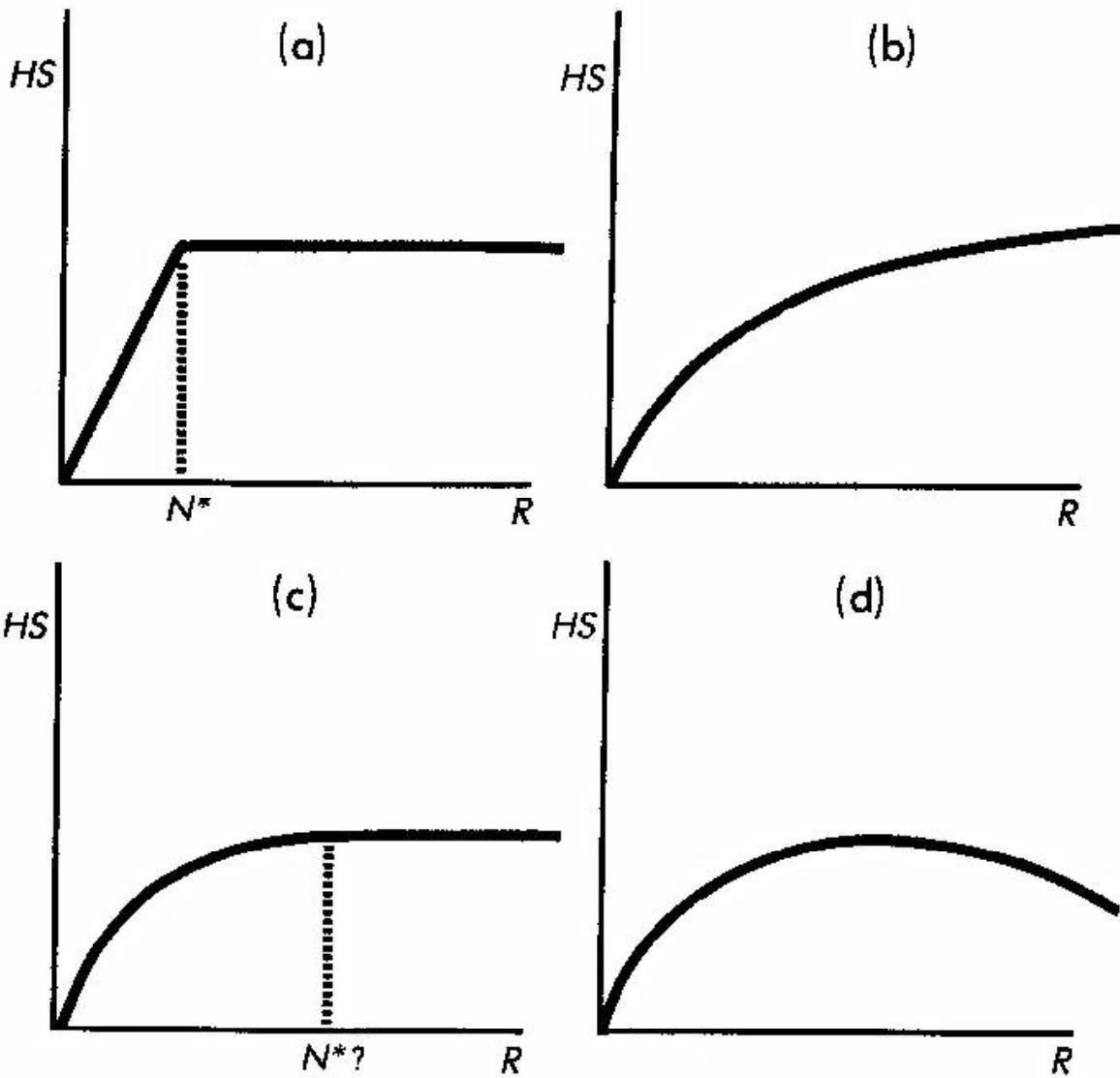
This rather abstract approach can be given more concrete form by postulating a relationship between resource inputs and health status, as shown in Figure 1-3. Panel (a) of that figure depicts a relationship consistent with the "meeting the needs" approach to resource allocation. More resources devoted to health care by a particular society yield substantial increases in health status for some members of the community, up to the point that all "needs" are met, N^* , after which further resources used yield no health benefits to anyone.⁸

Need is thus a technical concept, a statement about the capacity of particular resource inputs to influence the health of a particular person or group. The technical expert can legitimately claim special expertise in judging the needs of others; and, indeed, given the special circumstances of physical disability and emotional stress which frequently attend the use of health care, the external observer may have an advantage quite apart from expertise. Surgeons do not, in general, operate on members of their families, just as lawyers are advised not to handle their own cases.

It is clear that if panel (a) of Figure 1-3 accurately represented the technical relationship between health care and health status, resource inputs beyond N^* would represent overuse. But it does not follow that N^* is optimal, Health status per se, at least on its narrow definition, is not the only object of human activity or source of satisfaction. In panel (a), equal resource inputs up to N^* yield equal gains in health status, but these may well represent declining (though always positive) increments in utility or well-being. The "law" of diminishing marginal utility postulates that the benefits that a consumer/patient derives from further increments of any valued good or service fall, as the total amount possessed/used rises. Moreover, as health care production

FIGURE 1-3

Alternative Representations of the Relationship Between Community Health Status, HS , and the Quantity of Real Resources of Manpower, Capital, and Raw Materials, R , Devoted to Health Care Production



expands, the drawing of successive increments of resources from other sectors causes them to shrink, and thus their marginal value (the opportunity cost of health care) to rise. The balancing of health status payoff and opportunity cost could occur before the health sector expanded to N^* . Some level of unmet need would then be optimal, in a broader sense.

As panel (a) is drawn, however, N^* may well be optimal. If the relation has a steep slope up to N^* -- health care is highly efficacious until needs are met -- and if health status is highly

valued relative to other things until "perfect health" (or at least the limit of the capacity of health care to improve health) is reached, then the kink at N^* will ensure that the marginal value of health care exceeds its opportunity cost everywhere below N^* . The practitioner's focus on meeting needs is thus justified as economically efficient.

Panel (b), on the other hand, presents a more troublesome case. If the curve never flattens out, then "needs" can never be met. There is always something more which could be done. This need not imply an unbounded concept of health status; the curve could be asymptotic to a horizontal "perfect health" line. But if its slope is everywhere positive, then in a world of finite resources, unmet needs are inevitable. Every society will have to develop some institutional mechanism for deciding where and how to limit the resources devoted to health care. And this decision is not a technical judgement (Williams 1978). The "expert," the practitioner or researcher, may legitimately claim superior competence in identifying and describing the shape of the resource-health relationship. But the ultimate choice of where to stop is a collective and social one. The institutions evolved in different societies may give practitioners a special role in this decision process, but the decision itself cannot be based on technical considerations. Societies may also, through market mechanisms, give more or less weight to the choices of individual patients, but since the resources used are almost entirely collective resources, whether raised through public programs or private insurance, again the decision is ultimately a collective one.

Panels (c) and (d) represent additional possibilities for the resource-health relationship. In (c), "flat-of-the-curve medicine" (Enthoven 1980) is the collection of activities which have, in fact, no health payoff. Unlike panel (a), however, the flat-of-the-curve is not reached discontinuously at a well-defined "need" point. There is some point at which the slope of the curve goes to zero, but this may be very difficult to identify in practice. And the problem does not present itself to individual decision-makers, clinicians, or patients in the form of Figure 1-3, but rather as a very specific decision in an individual case. The pressures to "do something" may be intense, while information about expected payoffs may be imperfect or simply unavailable. Hence the possibility that the health industry in aggregate, or specific sub-components of it, may extend resource use well beyond the point at which interventions cease to be effective. Yet each participant may act, under imperfect information, with the best of intentions, and thus be highly resistant to the suggestion that the curve is flat in her vicinity. The aggregate curves of Figure 1-3 represent probabilistic abstractions in a world of uncertainty -- health care is not an exact science.

In panel (d) is depicted the relationship described by Illich (1975): health care is hazardous to health. There is a range of output over which health care yields positive benefit; but past some point actual harm develops. It is important to stress this possibility because, for particular forms of diagnosis or therapy, it is well established that overuse leads to harm. More is most emphatically *not* better. Indiscriminate use of antibiotics, surgery on healthy organs, excessive diagnostic radiography, are obvious examples. In other areas -- Caesarian section for example, or tonsillectomy -- there is considerable expert concern that current levels of provision are in the downward-sloping range. Even screening can lead to interventions which endanger health. In general, for every particular form of health care, there is some point at which excess does harm. For some forms, in some places, there is evidence suggesting that point has been reached. Whether it is being approached on average or in aggregate is not clear.

Panels (c) and (d) enable us to represent overuse in health care, as well as economic, terms. Operation on the flat-of-the-curve, or *a fortiori* on the negative slope, is clearly medical

overuse.⁹ In panel (b), on the other hand, more is always better from a health standpoint and global over- or underuse can only be defined economically. (Particular forms of care, however, can be overprovided in health terms; in panel (b) this would imply production below the curve). Panel (a) represents the possibility of a happy coincidence of health and economic criteria which may hold in particular settings, but seems implausible in general.

In each panel the issue of technical efficiency is side-stepped by the assumption that a society's health system is operating on the curve, not below it. The curve represents the maximum health status attainable with a given resource commitment, under the constraints of present knowledge. It may of course shift, as knowledge changes. But it is always possible to be inefficient, and attain less than the maximum possible health status with given resources. Maintenance of efficient production is a problem logically separate from that of deciding how large the health sector should be, although both are "solved" by the particular set of institutions which a society establishes to operate its health care system. Moreover, one's judgements as to whether to expand or contract the resources available to health care may well depend on how efficiently one believes the system is using resources at present, or would use more.

Economic analysis tends to focus on questions of resource allocation, rather than technical efficiency, on the ground that in the private sector, where production is carried on by for-profit firms, the incentive of profit maximization will ensure that costs of production are minimized and resources are not wasted. But most health care production, including as we shall see that in professional practices, is not carried on with the objective of profit maximization.¹⁰ Accordingly, issues of technical efficiency and appropriate institutional design assume great importance in the economics of health care.

MODELLING THE HEALTH CARE UTILIZATION PROCESS

Despite the central importance of health status to the organization and delivery of health care, its investigation on a societal or system-wide basis receives remarkably little attention. The *Canadian Sickness Survey* of 1950-51 (Canada, Department of National Health and Welfare and Dominion Bureau of Statistics 1960) provided a baseline for the initial development of public hospital and medical insurance programs, but this initiative was not followed up. Almost thirty years later, the Canada Health Survey was developed for inclusion with the monthly Labour Force Survey, but what was originally intended as a continuing source of new information was converted, for budgetary reasons, to a one-time effort. We are left with a "snap-shot" at a single point in time. (Abelson et al. 1983; Canada, Health and Welfare Canada and Statistics Canada 1981). Thus Canadians now spend over \$30 billion annually (1982) to improve their health, but make no systematic effort to measure the results.

Instead, the gap between health care and health status is bridged by inference and assumption, a process with significant parallels in economic methodology. To illustrate these parallels, we outline two alternative ways of conceptualizing or modelling the process of health care delivery, each with markedly different assumptions and policy implications, but very similar structure. These we label the Naive Medico-Technical and the Naive Economic models. In their naive forms they are poles apart, but more sophisticated forms converge towards each other.

The Naive Medico-Technical model begins with utilization, which may be either observed directly or inferred from capacity measures such as the number of physicians per capita. On this is superimposed assumptions about the roles and objectives of providers and users of care.

Providers are assumed to supply care in response to "needs" in a technical sense, and their perceptions of "needs" are assumed to be as accurate as possible with present knowledge. They thus control the mix and volume of care supplied, subject only to "barriers to care," economic or informational, which may prevent consumer/patients from seeking needed care or complying with a recommended regimen, or to limits on the capacity of providers as a whole to meet all the "needs." The proper role of public policy, in this framework, is to remove economic barriers to care (by providing public or mandating private insurance), to ensure that sufficient personnel and facilities are trained and established, and perhaps to launch educational campaigns to encourage care-seeking and compliance.

The behavioural concept which an economist would call "supply," that is, the amount providers wish to provide at current costs of production and rates of reimbursement, is, in the Naive Medico-Technical model, equated with the technical concept of "need," and its dependence on prices is suppressed. Utilization is in turn determined by "supply," except insofar as it is constrained by barriers or shortages.

Demand, that is the amount consumer/patients themselves might wish to consume at current levels of prices and incomes, plays no important role in this process. Consumers may not demand as much care as they need, but this is presumably due to ignorance or inability to pay. Both should be corrected by appropriate policy. Consumers might also make initial visits which, in the judgement of providers, are unnecessary or "frivolous," although frivolity at the care initiation stage presumably depends on whether a particular patient could reasonably be expected to know that the contact was unnecessary. The model is a bit fuzzy on this point; if a patient thought a visit was necessary but should have known better, is that overuse? Presumably expert providers terminate unnecessary episodes of care after one visit, and attempt to educate patients as to when to seek care, so that this qualification is unlikely to be quantitatively significant. The implicit expectation that patients will rarely seek care they know to be unnecessary is consistent with our formulation above that health *status*, not health *care*, is valued by consumer/patients. On the important issues of how to ensure technical efficiency in health care production, or how the earnings of providers will or should be determined, the model is silent.

While over-utilization, utilization exceeding needs, could in principle occur in this model, it is ruled out in practice by the assumptions that providers' decisions control use and that providers respond only to need. Health planning, and particularly manpower planning, in the context of this model is constantly responding to shortages, and unmet needs, because whatever level of provision is occurring in any part of the planning jurisdiction is, by definition, no more than enough.

The Naive Medico-Technical model is easy to criticize on a number of grounds, being in particular vulnerable to the empirical observations that apparently equally competent practitioners show wide variations in their perceptions of need and in their responses and that collective perceptions of need rise steadily over time, so as to lie always at or just beyond current levels of provision. Moreover, these perceptions of need, as reflected in utilization patterns, often bear only a tenuous relation, if any, to the available scientific evidence as to the efficacy and effectiveness of particular interventions. But the Naive Economic model is equally vulnerable, differing only in identifying observed utilization with the behaviour of a different set of actors.

In the Naive Economic model, demand plays a central role. This is the amount and mix of care consumer/patients choose to accept, in response to their own preferences and incomes, and the costs to them of care. This behavioural construct is then identified, by assumption, with utilization. Informed consumers choose which forms of care to use, in what quantities. Of course,

use may fall short of demand if supplies are constrained, but cannot exceed it. Nor will a shortage persist, if prices are flexible, since the quantity consumers demand depends on the prices they pay. If prices rise in response to a shortage, demand will fall, and balance is restored. Moreover supply, what providers wish to produce, is also assumed price-dependent. If prices rise, more will be offered for sale. Thus price plays the key role of equilibrating supply and demand, and so long as prices adjust smoothly, utilization will always be equal to both demand and supply. Need, on the other hand, drops out of this framework. Consumers' demands will presumably depend inter alia on their perceptions of their own needs. Insofar as external experts' judgements of needs correspond to consumers' judgements, they are thus embodied in demand; insofar as they differ, they are, by assumption, irrelevant.

The models could hardly be more different in appearance, yet, in fact, both share a common structure. Both start with an observed or observable statistical datum -- utilization -- and then assume its correspondence to a hypothesized behavioural concept -- consumers' demands or providers' preferred supply patterns. In each case observed utilization can fall short of, but not exceed, the hypothetical concepts, and such shortfalls represent some form of institutional failure (barriers to care, or sticky prices). Both make specific assumptions about the processes of care provision and consumption -- informed consumers or professional providers make the utilization decisions. And both have an implicit criterion for how care ought to be allocated -- to meet technically determined needs, or to respond to the preferences of consumers as expressed in willingness to pay. Finally, neither model is particularly realistic.

The Medico-Technical model, however, dominated health care planning and policy making in Canada to the end of the 1960s, and still underlies much thinking on this subject. Scratch a health care provider to any depth, and she will usually turn out to conceive of health care delivery in some such terms. Yet as noted, "meeting all the needs" is not possible if the health status curve in Figure 1-3 does not flatten, nor desirable if it flattens very slowly. Nor does the empirical evidence support the hypothesized chain from unambiguous, externally determined needs, to supply, to use.

On the other hand, critiques of the organization of the health care industry have been launched for at least a generation on the basis of the Naive Economic model -- usually calling for freer entry, more competition, more reliance on prices, and less on subsidies or insurance. The adoption of such a framework as a basis for analysis necessarily imposes the strong, though implicit, assumption that the extensive and detailed structure of public and self-regulation and control which governs health care in every developed society is all a terrible mistake, the result of a nefarious conspiracy by suppliers imposed on an ignorant and gullible public, which should be dismantled as soon as possible. Though all who use the unmodified free market economic model are in fact making this assumption (or are just intellectually inconsistent), few have had the courage to follow Friedman (1962) in declaring it openly.

These general critiques have had little impact in Canada (although rather more in the United States) because they are confronted with a perception among health care providers and general population alike that health care is somehow "different," that it is a commodity to which the normal rules of the marketplace do not apply. Or put another way, the patterns of resource allocation in health care which would result from reliance on private, arm's-length market-type institutions to control its production and distribution are demonstrably inferior to more regulated alternatives. Insofar as the Naive Economic model ignores the whole issue of "differentness," and begs the question of the commodity status of health care, analyses based on it have in turn been largely ignored -- and probably deservedly so.

But the market model exercises a powerful attraction, not least because of the inherent weaknesses of the Medico-Technical approach. It is the standard taken for the "normal" organization of production, distribution, and exchange, and the predominant form in industrialized "mixed capitalist" economies, because under certain conditions market mechanisms for resource allocation do have powerful advantages. Moreover, if totally unregulated free markets are inappropriate institutions to organize all resource allocation in health care, it does not follow that market mechanisms of some form have no role to play. Nor can one assume that whatever system of regulated organization may have developed in a society is necessarily optimal. The arguments for licensure of physicians do not necessarily extend to pharmacists or veterinarians, nor does a case for some form of regulation necessarily imply the forms of licensure we observe. And the justifications for public insurance of hospital costs will not all survive transplantation to dental care or eyeglasses.

Thus it is of central importance to the study of the economics of health care to understand in what ways this commodity fails to meet the conditions under which private market institutions would govern its production and allocation satisfactorily. Analysis and policy can then be firmly based on a clear linkage between nature and sources of market failure and appropriate institutional response. And the question as to why unregulated markets "fail" in health care provision is another version of the question, Why is there a separate economics of health care at all?

An early response was to try to list the features of health care which distinguish it from the standard "commodity" of the economics textbooks (e.g., Klarman 1965). These lists could be extensive: uncertainty of incidence of illness, and hence variability of "demand" for care; impact of the health or health care of one person on another via, e.g. contagion; prevalence of non-profit and non-profit-maximizing forms of organizations supplying care; barriers to entry in the form of licensure; extensive public regulation, both direct and delegated, of providers and consumers; a mix of consumption and investment services in health care; numerous and detailed job definitions with rigid barriers between jobs; large element of personal service; health care as a "need"; illness, and to some extent, care as assault on personal integrity, physical and emotional; high level of consumer ignorance about effects of care and capability of providers; mix of care and educational elements in much of provision; many consumers (children, elderly and infirm, severely injured) incapable of directing their own consumption; and so on.

An immediate response is that every one of these "peculiarities" is characteristic of some other commodities as well, to which the rebuttal is, yes but not all of them at once. But so what? The listing approach, however accurate and complete, can never tell us which peculiarities are critical, and why, and what their implications are for appropriate institutional design.

Arrow (1963) noted the important distinction between inherent and derivative characteristics, of features intrinsic to health care itself as distinct from particular institutional responses. Thus uncertainty of illness impact is inherent; but licensure or non-profit organization are derivative, social responses to other perceived peculiarities. This enables us to pare down the list considerably.

He also identified clearly the source of the difficulties created by consumer ignorance, noting that while ignorance about production technology is pervasive -- how many people know how an electronic calculator is put together? -- the critical issue is knowledge of what the product will do for the user.¹¹ The health care consumer is generally ignorant not only of how the commodity is produced. but of what it will do for her. The buyer of calculators, or commodities

in general, is not. Moreover, knowledge is asymmetric; the provider is generally perceived, by all concerned, to have a substantially greater degree of knowledge than the user.

Arrow referred to both the unpredictability of illness and the lack of consumer information about appropriate use (and as part of this, provider quality) as "uncertainty," but the concepts are quite different. Uncertainty is now used to refer to unpredictability, by anyone, of future illness events: Will I break my leg skiing this winter? One may (or may not) know the probabilities in advance, but cannot know of the event. Asymmetry of information refers to the present situation -- is my leg now broken, and what should I do about it? -- and the professional provider is usually assumed to have much better information than I on this subject.

Culyer (1971) explored in more detail the issue of interpersonal effects, demonstrating that these go far beyond contagion. The concept of "need" implies not only a technical judgement about the impact of health care on health, but also an obligation on someone, or on society generally, to respond. Similarly, the identification of illness as an assault on integrity or of ill people as unable to make appropriate consumption responses, implies a special collective response to protect consumer/patients in situations in which what is at issue is not the satisfaction the consumer may derive from bundles of goods consumed, but possible radical modifications or dissolution of the consumer herself. Thus, the concept of external effects, or of interactions between one person's consumption pattern and another person's well-being, is much broader and deeper in health care than for most other commodities.

These three intrinsic characteristics, uncertainty of illness incidence, external effects in consumption, and asymmetry of information between provider and user form the fundamental triad from which all other listed characteristics can be derived, either as variants or as social responses to them. Licensure responds to asymmetry, public subsidy or supply to externalities, insurance to uncertainty. Such things as high degree of personal service or mix of consumption and investment elements are relevant only insofar as they reinforce asymmetry of information; if consumers are fully informed, then private markets handle such problems without difficulty. "Need" and "loss of personal integrity" as sources of market failure come through external effects and informational asymmetry respectively. They are the answer, at the most basic level, to the question, Why is health care different? or equivalently, Why do we see such extensive intervention in, and supersessions of, private markets in health care?

Each of these three characteristics, however, creates different problems and has different implications for institutional design; moreover, different types of health care share these characteristics to very different degrees. Accordingly, we now proceed in the next three chapters to a more detailed examination of each.

NOTES

¹ In technical terms, we define the consumer's utility function U as extending over a set of commodities X_i , including health care (HC), and also over health status HS. But health care enters positively via its presumed positive effect on HS, its direct effects are negative. Thus:

$$U = U[X_1, \dots, X_n, HC, HS (HC)]$$

where $\delta U/\delta HS > 0$, $dHS/dHC \geq 0$ in the relevant range (although $dHS/dHC < 0$ is technically possible), and $\delta U/\delta HC < 0$. This formulation differs from the characteristics approach to consumer theory in that HS is a characteristic of the consumer, not the commodity; and the contribution of care to health, dHS/dHC , depends on the status of the user and is not intrinsic to the commodity (Evans and Wolfson 1980).

² These harmful effects are presumably outweighed by the direct satisfactions of consumption, though in the case of addictive substances like tobacco this presumption may be invalid.

³ The non-satiation postulate is preserved in our framework, but it is attached to health status, not health care. More health is always better. Nor need this imply unbounded health; "perfect" health may be approached asymptotically. But the relationship between health and health care is not necessarily monotonic, Analyses of the demand for "health" as opposed to health care (e.g., Grossman 1970) which assume that the marginal health productivity of health care is always positive (even constant returns to scale production functions!) are cut off by this counterfactual from the consideration of over-utilization in the sense above. Moreover, the concept of "health status" which is implicit in such analyses is only loosely related to the term as conventionally used.

⁴ It should be kept in mind that "incomes" are more than salaries, wages, and professional fees, and "people earning incomes" is broader than health care workers per se. Since most branches of health care are labour-intensive and employ specialized and distinctive personnel, much of expenditure is identifiable as incomes of doctors, nurses, dentists, technicians, etc. But expenditures on commodities, drugs, equipment, or buildings, say, generate wages, profits, "rent interest and dividends," in the associated manufacturing sectors. Not only drug- or equipment-company shareholders, but in the United States, owners of hospital bonds, form part of the set M of income earners, and analyses (or policies) which ignore their participation will be incomplete (will probably fail).

⁵ Price indices for physicians' services based on list or reported fees rose more rapidly than the CPI over the period 1947-71, at 3.6 percent per year on average, compared with 3.0 percent for the CPI (Barer and Evans 1983). But such indices of list fees do not capture the increase in prices received due to improving collections ratios, which appears to have been substantial during the period of spread of private and public medical insurance. Deflation of physician incomes by listed fees leads to implausibly high "productivity" gains. Efforts to estimate changes in fees actually received over the period 1947-71 suggest that fees collected rose between one and two percent per year faster than list fees, averaging over the whole period.

⁶ Though as noted in Table 1-2, the income data are somewhat suspect after the mid-1970s. Alternative estimates prepared by Health and Welfare Canada show physician relative incomes rising after 1977, and up sharply in 1981 and 1982.

⁷ The time period is a critical feature of these implicit substitution processes. Obviously, one cannot instantaneously convert paediatricians to fighter pilots or first violinists, or vice-versa, and sewage plants do not make good hospitals. But over a longer time horizon, say a decade or more, training programs for different forms of skilled manpower can be expanded or contracted to change significantly the types of people available, and the construction industry can turn out hospitals, roads, factories, or runways with the same human and material resources.

⁸ The definition of health status for individuals, and its aggregation across individuals, are both highly problematic, and the aggregation of resource inputs to a single measure also raises numerous difficulties. Moreover, health status is a trajectory through time, and a static representation implies some way of representing and adding up expected future states. Culyer (1978) provides a discussion of the conceptual problems involved, and an application. Despite severe problems in its operationalization, however, some such concept of a resources-health relationship underlies all health policy discussion. (Nor should economists be overly critical -- who ever estimated a Walrasian general equilibrium system?)

⁹ It is not *necessarily* economic overuse. Economists who treat health care as an argument in the utility function, or a valued object in and of itself, sometimes argue that "unnecessary" operations, e.g., are justified if (informed) patients are willing to accept them, even if they confer no health benefit. This sounds very much like Münchhausen's Syndrome again. By defining health status, not health care, as the relevant object of value, and assigning health care per se negative weight, we close off this rather peculiar argument and gain in realism.

¹⁰ Self-employed practitioners may or may not maximize *incomes*, but incomes are not profits. Failure to observe this basic theoretical distinction has muddled many discussions of practitioner behaviour.

¹¹ In general, the consumer with a utility function U defined over commodities X_i is assumed to know, at least as well as anyone else, the marginal utilities $\delta U/\delta X_i$ and hence to make informed consumption decisions. In the case of health care, since it is an input to health, $U[X_i, HS (HC)]$ implies that the consumer needs to know not only $\delta U/\delta HS$, but also the technical production function relation, dHS/dHC .