

# CHAPTER 13

## HEALTH MANPOWER POLICY-LEADING THE HORSES TO WATER

### PUBLIC INVESTMENT IN PRIVATE CAPITAL

Boulding (1968) finds "the whole manpower concept repulsive, disgusting, dangerous, fascist, communistic, incompatible with the ideals of liberal democracy, and unsuitable company for the minds of the young. (It) is basically ... an engineering concept, and one of the main problems of society is to keep engineers in a decently subordinate position." Without pretending to do justice to his subsequent argument, we may note that the concept of manpower<sup>1</sup> explicitly categorizes people, as well as their competence and intellectual development, as objects of analysis and inputs to production processes, means to some external ends, rather than as ends in themselves. It is in direct defiance of the categorical imperative.

Yet the manpower concept, like engineers, turns out to be unavoidable and often quite useful. The survival, let alone growth and development, of all organizations depends on the availability of human resources, time, effort, and skills, to carry out their activities. Governments may refer to manpower policies while private firms speak of personnel policy, but each must consider the nature, numbers, and source of the trained personnel needed for or implied by its other objectives. As in so many other areas, it is impossible *not* to have a policy, although the actual policy may be more or less explicit, consistent, or farsighted.

Nor can the manpower concept be blamed entirely on the engineers. Economists have quite happily developed the concept of "human capital." The time, energy, and other resources spent by a person in acquiring skills or abilities, or by others in inculcating those skills, are thought of as an investment in the creation of an intangible form of capital equipment which increases the productivity of the trainee, just as time and effort might have been devoted to creating a piece of machinery, physical capital, for her to work with.<sup>2</sup> And almost all economic analyses of labour markets or production processes explicitly treat labour as a commodity, an input to production or an object of exchange, differentiated, if at all, only by its admixture of "human capital" services.

Boulding is not the only one to object to the notion that this "commodity" can be manipulated separately from its human or social context. Unemployment is not simply an excess inventory of labour input, like a pile of coal, nor can human capital be "scrapped" on the basis of the same calculations as an obsolete blast furnace or a damaged automobile. But such concerns have troubled economists little, if at all.

They do, however, create an important distinction between the manpower policy of the state and the personnel policy of a private firm. Unless it is very large or highly specialized, the individual firm's decisions do not determine the market for labour. In acquiring skilled services, it can choose to hire or to train internally, and its "de-hiring" decisions, while potentially traumatic to an individual, do not in themselves foreclose other employment opportunities. The manpower policy of the state, however, is defined over the whole labour market (abstracting from immigration) and is thereby more constrained. If a public program or policy, such as the expansion of the health care system, requires more skilled personnel than are available, or different ones, educational programs must be established or expanded to train them, and the time

required may influence the timetable of program introduction. Moreover, educational programs in general have been accepted, in Canada as in most other countries, as a public responsibility separate from the particulars of health policy. Manpower issues thus cut across two conceptually separate areas of social policy whose co-ordination has been somewhat less than perfect.

On the other side of the coin, and perhaps more relevant to present Canadian conditions, the "de-hiring" decisions implicit in public programs may affect entire markets. If the expansion of the hospital sector is curtailed, while training programs continue to turn out nurses with ever more extensive and expensive training, some at least of the newly manufactured "human capital" will be idled. Similarly, medical schools, whose throughput now appears sufficient to keep the physician-to-population ratio climbing at least to the year 2031, create enormous stresses for provincial governments and medical associations trying respectively to limit the escalation of medical costs and to maintain their members' incomes. The state as reimbursing for, if not operator of, health programs has an implicit responsibility to provide employment for the "human capital" which it has encouraged and assisted individuals to acquire; surplus capital cannot be stockpiled or scrapped. Whether or not this responsibility is explicitly accepted, the political process will ensure that it cannot be entirely escaped. Thus the manpower decisions of the present set many of the parameters for delivery systems in the future, over as long as a generation. And the program constraints imposed by the skilled people who are not now available are much less serious and long-lasting than the constraints imposed by those who are.

## **A POLAR CASE: PATTERNS OF MANPOWER ALLOCATION IN PURE COMPETITION**

These constraints arise from the regulated and subsidized nature of the health care system itself; there are similar problems in other regulated occupations such as law. It is worth recapitulating how a hypothetical perfectly competitive and unregulated health care system would deal with manpower issues, not in order to suggest that such a system would be more appropriate, but only to indicate how the key issues of manpower policy derive from the peculiar nature of the commodity "health care" itself, and the peculiar regulatory and financing structures which this makes necessary. "Market failure," and public intervention in the production and distribution of the product, leads to "market failure" in the process of human capital production as well, and the need for further intervention at that level. Failure to co-ordinate such intervention can lead to dramatic and costly problems in the delivery system.

If the production and distribution of health care were not regulated, but governed by the laws of the free market as they exist in economics texts and editorials in the business press,<sup>3</sup> there would be no need for public "manpower policy." Privately owned firms, probably incorporated, would own and manage hospitals and medical clinics on a for-profit basis, competing for customers by some combination of (real or perceived) quality differentiation, promotional advertising, and competitive pricing. The "product lines" of such organizations, the range of services they chose to provide, would be determined by what they felt to be most profitable. Individuals could enter this market freely, as employees or self-employed practitioners, without the need for licensure, though they might well seek some form of public or private certification in order to communicate their skills to potential customer/patients in a convincing manner.

The manpower mix in this imaginary health care system would be governed by the hiring decisions of producer firms and by the ability of self-employed practitioners to find a clientele.

In particular, profit-oriented firms would hire the least-cost combination of personnel necessary to produce their products. The constraints on any resulting quality dilution would be, not licensure or regulation, but the refusal of customer/patients (assumed fully or adequately informed) to patronize low quality establishments. This might be backed up by regulations to prevent false advertising or other forms of consumer fraud, as in any other retail business, and perhaps also by some enforcement of minimum standards of product quality or safety, as in the food industry. Further, customers could use the tort law system -- malpractice claims -- to recover for bad outcomes traceable to producer fault. But *process* of production would be unregulated, and inter-provider competition would ensure that the least-cost mix of personnel, which might or might not be the least "human capital intensive" process of production, was used. Pharmacists would no longer find employment moving pills from large bottles to small, dentists in drilling teeth, or physicians in performing well-baby checks.<sup>4</sup>

In this world, the relative demand for and earnings of different types of personnel would be derived from consumer demands for the services they produced, relative to the numbers of people offering such services. The "net advantages" model of occupational choice predicts that informed persons choosing careers will treat training decisions as an investment in which current earnings are foregone (and perhaps extra effort required) in order to acquire skills which will command higher earnings later. These later earnings may be in money, or they may be in the form of desirable employment characteristics -- pleasant surroundings, flexible hours of work, "respect," self-determination. Net advantages refers to both. As in any other investment, the stream of future net advantages must be discounted back to the present, and compared with the training cost. Thus the *i*th career choice has a present value:

$$PV_i = \sum_{t=0}^{N^i} \left[ \frac{Y_t^i - \bar{Y}_t - E_t^i}{(1+R)^t} \right] \quad (13-1)$$

Here *t* refers to years from the present (*t*=0) to retirement after *t*=*N*<sup>*i*</sup> years -- self-employed occupations permit more latitude in choice of *N*. *Y*<sup>*i*</sup><sub>*t*</sub> is the earnings plus other advantages (which may be negative) in year *t* for occupation *i*, while  $\bar{Y}_t$  is net advantages of some reference occupation requiring no training. *E*<sup>*i*</sup><sub>*t*</sub> are direct training costs in year *t* for occupation *i*. The rate at which the individual discounts future benefits, *R*, should be equal to the real rate of interest set in the capital market, expected to hold over the period  $0 \leq t \leq N^i$ .<sup>5</sup>

The free market would apply also to training institutions and processes; thus *E*<sup>*i*</sup><sub>*t*</sub> would reflect the full costs of providing training for occupation *i* in year *t*, and *Y*<sup>*i*</sup><sub>*t*</sub> would be small or zero in the early years of a career with extended schooling.

(*Y*<sup>*i*</sup><sub>*t*</sub> -  $\bar{Y}_t$ ) represents, abstracting from its net advantages aspect, the foregone earnings associated with training in career *i*.<sup>6</sup>

In long-run equilibrium *PV*<sub>*i*</sub> will be equalized at zero across all *i*; no one occupation will command any advantage over any other in life-time, discounted net advantages. If it did, informed entrants would differentially select that occupation, leading to a fall in its members' incomes. Similarly, underpaid occupations would fail to recruit entrants, supply would fall, and earnings rise.<sup>7</sup>

Of course *annual* earnings are not equalized; occupations with long training periods will yield higher annual earnings during the (shorter) working period, as they should and indeed must to attract recruits. But inter-occupational earning patterns will (abstracting from other advantages) bear a determinate relationship to each other -- high earning occupations over a

lifetime yield a surplus just large enough to repay (with interest) their training cost (direct and foregone income).

In this environment, "manpower policy" is unnecessary. A shortage or surplus of a particular class of personnel, resulting from shifts in consumer tastes or producer technology, will be reflected initially in the bidding up or down of the incomes of those types of personnel. The income changes will result from rising or falling prices for the types of goods or services which these personnel participate in producing. The particular mix of skills embodied in a given class of personnel will be determined by producing firms -- large corporations or self-employed practitioners, according to the dictates of least-cost production. As incomes in different occupations shift, the *PV*'s of career decisions rise or fall, and recruitment patterns follow. Eventually, supply responses (or shifts of personnel from careers with low *PV* to those with high) will reestablish equilibrium and will equalize *PV*'s once more across all occupations.<sup>8</sup>

## **THE PRINCIPAL CHANNELS OF PUBLIC INFLUENCE OR CONTROL**

Of course, the above scenario does not describe health care delivery systems or the processes of personnel recruitment anywhere in the world, and has not for at least a century -- if ever. Public intervention, to encourage or restrain, is pervasive and, in principle, almost universally accepted.<sup>9</sup>

This intervention takes place in three primary ways:

1. The use of state authority to regulate occupations, either directly or more commonly by delegation to private professional bodies;
2. The provision of subsidies for educational costs, and the determination of how much and what kind of training capacity will be made available; and
3. The structuring of public delivery or reimbursement systems to determine how many of what classes of people shall be hired to provide, or reimbursed for providing, which kinds of services.

There exist, of course, a variety of other public policies which also affect manpower availability or use -- legislation governing malpractice liability, for example -- but the above three are the big levers.

The process of occupational regulation is of fundamental importance insofar as it determines which combinations of skills and capacities can be assembled in particular individuals. To pursue the "human capital" metaphor, the productivity-enhancing "machines" which are embodied in trained individuals can be assembled in a wide variety of different ways, with a broad or narrow scope or range of functions, and with deep or superficial competence. A very extensive training program can turn out an elaborate and costly piece of "human capital" capable of performing many functions and dealing with problems of great complexity or sophistication; a less extensive and costly program will produce "capital equipment" with a more limited range of functions, whose possessors must refer complex problems to another provider. This is, of course, merely another way of looking at the issue of alternative types of health care providers, addressed in chapters 6 and 7 above. The point here is that the public regulatory structure defines the boundaries between occupations, determining who may do what, and thus

regulates which bundles of capacities shall be assembled together and which kept separate. It further determines what bundles of capacities shall be required to perform which functions -- as emphasized above, self-regulating groups tend to require that licensed providers possess all capacities before exercising any, thus "overcapitalizing" beyond what is technically necessary. Finally, the regulatory structure determines the "process of production" of human capital, the educational steps necessary to acquire and use (legally) the capacity to provide services, as well as prescribing (and proscribing) particular activities of its possessors.

From the point of view of manpower policy, then, the state at the most basic level determines by regulation what the categories of manpower shall or shall not be, and how the bundles of capacities which they represent shall be acquired and used. It is of central importance to realize that these are regulatory decisions; they are not pre-determined by technical or market considerations. The occupational structure, and its educational requirements, are matters of policy choice, and the consequences of those choices are extremely far-reaching.

Yet these issues are frequently neglected in discussions of manpower policy, which usually take existing occupational definitions and divisions, and thus educational requirements, as given data. This is probably because governments tend to take their decisions in these matters with closed eyes, pretending to themselves and others that no decision is being made. The delegation of state authority to private provider associations, which then perform the regulatory functions and delineate occupational roles, permits both parties to disclaim responsibility. Provincial governments can claim that, having once delegated regulatory powers to "expert" external bodies, they are no longer responsible for the consequences, while private regulators disregard (or even deny) the fact that their authority is entirely derivative from the state. Neither claim has a shred of legal validity, being radically inconsistent with the fundamental constitutional principles of parliamentary sovereignty. But politically, the "separation" may be very convenient.

Moreover it can be reinforced by the assumption, often implicit in defenses of occupational regulation, that the existing allocations of functions and required training patterns are in some way predetermined by underlying technical necessity. Regulatory requirements are treated as if they merely reflect the only way people *can* be trained or services provided. The variety of regulation and experience across jurisdictions, however, and the experience with alternative practitioners of various types -- different ways of designing and assembling "human capital" -- suffice to refute this assumption when it is made explicit.

Any analysis of public manpower policy, therefore, must start with the regulatory structure, or else the largest part of policy is left unexamined. Within that structure, however, the state also plays a fundamental role, through its funding and subsidy policies in determining how many people, and of what types, will be trained. Almost all funding for post-secondary education flows through provincial governments; and while universities exercise independent control over their curricula and programs, a public decision to expand or contract facilities cannot be ignored. The offer of funding for a new professional school, or for a major expansion, represents more temptation than a university, however highly principled, is likely to be able to resist. A denial of funding, of course, is decisive. Moreover, although direct charges for education are relatively small in Canada, especially in the health occupations,<sup>10</sup> public subsidies may be paid directly to students in particular programs in the form of grants or low interest (perhaps forgivable) loans. At a later stage, the direct funding of training positions such as hospital residencies may be used to influence numbers and types of specialists. And tax legislation or fee negotiations with providers, or public service salary and benefit policy, may be used to influence the costs or economic benefits of continuing professional education.

Of course the educational policies of individual provinces may be undercut by interprovincial migration, either in or out, emphasizing the need for cross-provincial

coordination.<sup>11</sup> At the national level, migration considerations have at various times been an important part of physician manpower policy, but not of policy toward the other health occupations.

The rate of physician in-migration seems to be very sensitive to the "point" value of the occupation for immigration purposes. When it was classified as a shortage occupation in the late 1960s and early 1970s, immigration nearly matched domestic production of physicians. When this status was withdrawn in 1975, immigration fell by about three-quarters. Later in the decade, physician out-migration to the United States became more rapid, and was argued by some to impose constraints on public policy -- especially fee-setting -- in that if physicians were dissatisfied, they would all leave. Increasing saturation of the United States market, combined with the observation of continually increasing physician-to-population ratios in Canada, has muted this point, but international migration remains a potentially significant aspect of physician manpower policy. In particular, public (federal) policy can alleviate a perceived shortage of physicians quite swiftly (and cheaply) by re-opening immigration. A surplus, however, is quite another matter -- it seems unlikely that a policy of inducing out-migration would be politically viable.

The payment mechanism is also a critical part of manpower policy: "Who supplies what?" is closely linked with "Who is paid for what?" Most obviously, the right to bill the provincial insurance program for services rendered to patients is of great value to the various self-employed health professions. The decision by a particular province to issue billing numbers to members of a particular occupational group plays a critical role in encouraging the expansion of that occupation. On the other hand, as noted in chapters 6 and 7, the refusal to reimburse services of particular intermediate-level personnel, either in self-employment or as employees of other "approved" professions, will result in professional "still-birth" regardless of the availability or technical competence of such personnel.

More subtle problems arise in those areas where two different occupations share overlapping competences -- should the reimbursing pay each the same? In some cases, ophthalmologists and optometrists providing refractions, for example, the answer has been "yes." On the other hand, general practitioners and specialists are paid different fees for the same services in all provinces but Quebec. "Cost of production" and "equal pay for equal work" fee setting are difficult to reconcile -- economists instinctively reject the former. But if two different professions -- optometrists and ophthalmologists, for example -- are reimbursed at the same rate, and the patient perceives the latter as representing "more" -- at least a broader range -- of human capital than the former, will this not give the latter a marketing advantage, even if the former are less costly to train and equally effective? Yet differential reimbursement for equivalent services is difficult to justify on equity or efficiency grounds. In designing reimbursement systems, it appears advisable to think through their implicit manpower policy content, and be sure that one is happy with the results -- accidental policy is not necessarily optimal.

## **INFORMATIONAL AND INTERACTION PROBLEMS IN PUBLIC MANPOWER POLICY**

Working through these three main channels, government establishes health manpower policy either consciously or by default. It is simply impossible to have no policy, though fragmented, inconsistent, ill-considered, and excessively costly policy is all too possible.

Of course, deliberate policy, even if coherent and well thought out, is always vulnerable to unforeseen shifts in demography, technology, and social "tastes" in the broadest sense. Canadian policy toward physician manpower in the 1960s is a classic example. The population forecasts of the Hall Commission missed completely the "great obstetrical contraction" of 1965 and its aftermath (as did everyone else), so that by 1981 they were too high by about five million people, or 20 percent. The effects of such an error on manpower requirements estimates are, of course, massive; the resulting overestimate led to the putting in place of the present excess physician-training capacity.

The problem, however, is not merely that the future is inherently unknowable and that all forecasts are erroneous -- barring blind good luck. The error in the demographic forecasts was observable as early as 1966, and the new trend was clearly established by 1970. But instead of rethinking domestic capacity, which would have threatened jobs and careers in universities particularly, policy makers and forecasters made efforts first to rationalize away the new demographic reality as a short-term aberration, and then to justify the resulting dramatic increases in physicians per capita. Even in the mid 1980s, the nettle of manpower limitation has yet to be firmly grasped. It is much easier, politically, to gear up than to gear down.

The lesson, then, is not so much the difficulty of forecasting and planning, but that of acting on certain kinds of forecasts. The policy response is biased according to the type of information generated in the planning process -- for good, sound economic reasons. One must therefore strive to improve, not merely the quality of planning information, but the incentives on decision-makers to use it -- to confront the obvious. In any case, failures of rationality do not establish a positive case in favour of confronting the future in ignorance or absence of mind.

The public manpower planning problem can be thought of, without doing too much violence to reality, as a very large cost-effectiveness problem. From this viewpoint it can be fitted into the conceptual framework of chapter 11. Alternative types of manpower can be thought of as different investment programs in the sense that they represent the commitment of resources in the present to yield future benefits consisting of their contribution to the health and well-being of the community. At a very abstract level, human capital, physical capital, knowledge or research capital, or "capital" in the form of enhanced resistance to disease or injury resulting from preventive programs, all represent alternative ways of trading present resources of human time and effort, and physical inputs, for a stream of future health benefits.

Investments in manpower, however, add significant complexities to this analysis. Apart from the very large number of alternative "investment programs" available, and the difficulty of identifying and measuring the future benefits, two additional problems stand out.

First, the investment is a joint one -- the public at large and the individual choosing a career are both involved. There is no point in setting up training programs, however cost effective, in which no one wishes to enrol. For most of the health professions, over most of recent history, this has not been a problem. The relative financial rewards of the health occupations have been such as to lead to a substantial unsatisfied demand for entry, so that manpower policy can operate by direct rationing of access. But this need not be true in general, especially in those occupations such as nursing in which large numbers of qualified personnel choose not to practice. A policy of subsidizing entry to an occupation in which conditions of work are insufficiently attractive to hold the labour force is equivalent to keeping a sieve filled by adding ever more water -- possible, but peculiar.

Second, and very important, the payoffs to the various "investment projects" are non-additive. For simplicity, we assumed in chapter 11 that the streams of costs and benefits associated with particular projects were independent of the type and scale of other projects underway, so that the measurement of cost-benefit or cost-effectiveness balances for any one

project, or decisions to include it in a public agency's "portfolio" of projects, need not depend on what else is going on.

In general, however, this assumption will not be valid. The payoff to research on respiratory disease will be influenced by the effort and success level of anti-smoking campaigns. But it is dramatically falsified in the manpower field, where many of the principal issues revolve around who does what. Particular types of personnel are substitutes for, or complements to, others. We have emphasized the substitution relations above, because the policy-induced distortions in those relationships seem so severe and costly. But obviously surgeons and anaesthetists, or surgeons and hospital ward staff, represent complementary inputs whose numbers require some sort of balance. Manpower planning for one occupation cannot be done in isolation from the rest.<sup>12</sup>

### AN "ACTIVITY ANALYSIS" PLANNING FRAMEWORK

Manpower planning frameworks can be appallingly complex in the abstract, reducing to depressing simplicity, if not triviality, in application. One way of setting the problem up, however, which is based on Leontief's Input-Output models of the general economy, has the advantage of bringing out a number of the awkward assumptions involved in conventional methods.

First, hypothesize a population,  $P$ , which represents the target group -- people of Canada, people of a particular province -- for whose benefit the planning exercise and its *sequelae* are being carried through. We can then subdivide  $P$  into categories  $P_i$ , in which  $i$  indexes the amount and pattern of health services needs of different members of the population. Thus we have a vector,  $[P_1 \dots P_i \dots P_n]$ , such that summing over all  $P_i$  yields

$$P, \sum_{i=1}^n P_i = P$$

and within each group  $P_i$ , people are believed to be more or less homogeneous with respect to their expected service needs. A very common way of doing this is to let  $i$  index age-sex categories, but depending on the data and research effort available one could also break out separately particular ethnic, geographic, or employment sub-groups, or (at any point in time) people with particular chronic conditions.<sup>13</sup>

Given the population structure, one can attempt to determine, at least on a probabilistic basis, its needs for future services.

$$\begin{bmatrix} S_1 \\ \vdots \\ S_j \\ \vdots \\ S_m \end{bmatrix} = \begin{bmatrix} e_{11} & \dots & \dots & \dots & e_{1n} \\ \vdots & & \vdots & & \vdots \\ \vdots & \dots & e_{ji} & \dots & \vdots \\ \vdots & & \vdots & & \vdots \\ e_{m1} & \dots & \dots & \dots & e_{mn} \end{bmatrix} \times \begin{bmatrix} P_1 \\ \vdots \\ P_i \\ \vdots \\ P_n \end{bmatrix} \tag{13-2}$$

where the range of  $m$  different health services to be provided is represented by the vector  $S_1 \dots S_m$  and the  $m \times n$  matrix  $E$ , whose elements are  $e_{ji}$ , expresses the amount of services of type  $j$  expected to be needed by the average individual of type  $i$ . Put another way,

$$S_j = e_{j1}P_1 + e_{j2}P_2 \dots + e_{jn}P_n \tag{13-3}$$

and so on for all  $j$ . Each person of type 1 requires  $e_{j1}$  units of service of type  $j$  -- tooth restorations? office visits? measles shots? -- per time period.<sup>14</sup> And total need for services of type  $j$  is the sum of the needs for each of the  $i$  different population classes, which in turn is the product of the number of people in that class,  $P_i$ , times their average service needs,  $e_{ji}$ . The vector  $S$  then represents the "shopping list" of services required for population  $P$ .

The structure of the  $E$  matrix will (or should) embody detailed information about the effectiveness of particular services of the sort expressed at a very aggregate level in the health status curves of Figure 1-3. The appropriate servicing pattern for people of category  $i$ , represented by the vector  $(e_{1i} \dots e_{mi})$ , depends on the illness pattern of people in group  $i$  and on the capacity of services of each of the  $m$  different types to respond to those needs. The level of  $e_{ji}$  should be set such that further services of type  $j$  supplied to group  $i$  people have a zero, or positive, payoff depending on whether the efficacy curve for services  $j$  to people  $i$  has a well-defined kink (Figure 1-3a) or approaches the horizontal asymptotically (1-3b). In any case, it is the point of application of social priorities for care.

The specification of  $[e_{ji}]$  may be quite straightforward for fields like dentistry, where the range of different services is small and the population can be split fairly neatly into different "need" groups. The optimal values of the  $e_{ji}$  -- optimal maintenance schedules, choice of restoration technique -- may be subject to some debate, but the categorizations  $j$  and  $i$  are fairly easy. Other health care fields, with a very wide range of different services or fuzzy boundaries between services, may present substantially greater problems of definition. Nevertheless, as we shall show below, if the task is not attempted explicitly, it will be done implicitly, and difficult problems are not necessarily better solved in ignorance.

Given the "shopping list"  $S$ , there are a variety of forms of resources -- manpower and other inputs -- which can be used in their production. We can then specify a technology matrix  $T$ , whose elements  $t_{kj}$  represent the number of units of resources of type  $k$  needed to produce one unit of services  $j$ . They are the "input-output" coefficients. We can then write:

$$\begin{bmatrix} R_1 \\ \vdots \\ R_k \\ \vdots \\ R_q \end{bmatrix} = \begin{bmatrix} t_{11} & \dots & \dots & \dots & t_{1m} \\ \vdots & & \vdots & & \vdots \\ \vdots & \dots & t_{kj} & \dots & \vdots \\ \vdots & & \vdots & & \vdots \\ t_{q1} & \dots & \dots & \dots & t_{qm} \end{bmatrix} \times \begin{bmatrix} S_1 \\ \vdots \\ S_j \\ \vdots \\ S_m \end{bmatrix} \tag{13-4}$$

defining the amount of each of  $q$  types of resources which must be used up in producing the shopping list of services represented by vector  $S$  -- if technology  $T$  is employed. Each row of the matrix  $T$  yields an equation:

$$R_k = t_{k1}S_1 + t_{k2}S_2 + \dots + t_{km}S_m \tag{13-5}$$

implying that if  $R_k$  represents minutes of time of a particular class of personnel, say registered nurses, services of type 1 require  $t_{k1}$  minutes each of the time of persons of type  $k$ . But  $R_k$  can also be non-human inputs, either time of capital equipment, or physical units used up of such things as drugs or film.

Putting all these together, we can write the matrix equation:

$$R = T \bullet E \bullet P \tag{13-6}$$

where  $P$  is the  $nx1$  vector population groups  $P_i$  distinguished by their differing characteristics related to health care needs;  $E$  is the  $m \times n$  matrix of elements  $e_{ji}$  which convert population numbers to quantities of different services needed; and  $T$  is the  $q \times m$  matrix of elements  $t_{kj}$  which express the amounts of each different type of resource input needed to produce each unit of service.  $E$  may be thought of as the Epidemiology matrix, and  $T$  as the Technology matrix, though as we have noted above, the internal structure of the  $e_{ji}$  must embody policy choices as well as epidemiological information. So, as we shall see below, do the  $t_{kj}$ .

To derive manpower requirements, of course, we need yet an additional intermediate stage, in which numbers of personnel or other measures of total resources available are converted into working time. Thus a given number of people in occupation  $q$ , say  $M_q$ , do not necessarily each provide the same amount of physical input  $R_q$ . The ratio  $R_q/M_q$  may vary with provider age, sex, skill level, training location, geographic site, or form of economic organization. This variation is quite distinct from the variation in *service output* per class of manpower, which may depend on the machinery, assistants, etc. available; rather it indicates the time, effort, and skill *input* to production represented by a particular unit of "human capital" and its attached human being.

If provider personnel of type  $q$  are divided into  $W$  different categories, *e.g.*, by age and sex, such that the average time, effort, and skill per person differs across categories, then total inputs of services  $q$  will be:

$$R_q = r_{q1}M_{q1} + r_{q2}M_{q2} + \dots + r_{qw}M_{qw} \tag{13-7}$$

where  $r_{q1}$ ,  $r_{q2}$ , etc. represent the relative input levels from different classes of  $M_q$ . The number of *people* of type  $q$  will be measured by  $M_q = \sum_{h=1}^w M_{qh}$  but unless the *proportions*  $M_{qh}/M_q$  stay constant, the number of people required to generate a given volume of personnel services will fluctuate -- quite apart from the effects of shifts in the  $r_{qh}$  themselves.

### "MAGIC RATIOS": THE LIMITING CASE OF ACTIVITY ANALYSIS

Going back to (13-6), it is important to keep in mind that it is a matrix equation, resulting from substituting (13-2) into (13-4) and embodying all the detail of each. Neglect of the internal structure of  $T$ ,  $E$ , and  $P$  yields the traditional and much more limited manpower planning approach of "magic ratios." The search for "optimal" physician (dentist, pharmacist, podiatrist -- not yet veterinarian!)-to-population ratios continues to be a popular way of trying to sidestep the awkward and demanding  $T$  and  $E$  matrices. In the magic ratio approach, the  $P$ -vector is collapsed to a scalar -- total population  $P$ .  $T$  and  $E$  are likewise collapsed to a single number, and the relationship between  $M$  and  $R$  is assumed to be constant. Combining all these constants:

$$M_q = \Theta_q P \tag{13-8}$$

where  $\Theta_q$  is "the" optimal ratio of providers of type  $q$  per capita. Clearly, the task of manpower planning is greatly facilitated.

By contrast with (13-6) and (13-7), however, it is also clear that (13-8) has imposed two different types of drastic simplification, on the  $M_q$  and  $P$ , and on the  $T$  and  $E$ .

The assumptions that provider and user populations are homogeneous are serious, but not too difficult to remedy. Obviously "optimal" magic ratios should depend on population structure, the relation of obstetricians to the birth rate, and paediatricians and extended care nurses to the age structure, being fairly clear examples. But extending the "magic ratio" approach to take account of variations in age-sex structure, at least, is not very demanding. One merely chooses age-sex specific values of  $\Theta_q$ , measures the structure of  $P$ , and repeats the calculation.

## **MANPOWER, POPULATION AGING, AND UTILIZATION: CHICKEN OR EGG?**

Alternatively, one can abandon the perspective of "optimality," and merely measure age-sex specific utilization ratios. One can then project changes over time in average per capita and total utilization if these specific rates are held constant while the population evolves. This can be a rather useful exercise. Several such studies have been done for Canada which showed on the basis of mid-1970s population projections that the "great greying" of the Canadian population would not, in fact, have much impact on medical care utilization (an increase of less than 10 percent per capita from 1976 to 2001, and of 15-20 percent from 1976 to 2031 (Boulet and Grenier 1978; Denton and Spencer 1983)). The effects on hospital and other institutional use would be much greater, but even so would represent relatively low annual rates of increase compared with the large changes of the 1950s and 1960s.

When these utilization projections are compared with projections of physician manpower availability based on class sizes, immigration rates, and attrition patterns in the late 1970s and early 1980s, obvious discrepancies emerge. Projected increases in physician supply outrun population growth by 25 percent from 1976 to 2031, yielding population-to-physician ratios around 430. Weighting population growth by changing age structure reduces this to about 18 percent, but leaves the basic conclusion intact. If current rates of physician manpower production are maintained, either per capita utilization rates (age-sex adjusted) must rise, or physician workloads must fall. In either case, it is also true that either physician average incomes must fall (in real terms) or per capita medical expenditure must rise, with some combination of increasing fees (faster than the general price level) or increasing workload per physician. And these adjustments must continue for fifty years! The glacial effects of manpower policy -- slow but devastating -- are thus graphically displayed.

The 1976 population projections have been overtaken by an acceleration of declines in mortality rates in the late 1970s which is continuing into the 1980s. This has the effect of increasing both the total projected population and its average age. Thus the oversupply of physicians may turn out to be less severe -- relative to current age-sex specific use rates<sup>15</sup> -- though the implication of either long-term continuing increases in per capita medical care costs or falling physician real incomes remains.

All such projections, however, refer to potential *future* developments. Changes in mortality rates, even large ones in terms of past experience, affect population structure with long lags. (Becoming is not being.) And the aging of the Canadian population, whether faster or slower than anticipated, cannot by any stretch of the imagination explain or justify the conflicts over resource allocation in the late 1970s and early 1980s; it is on an altogether different time scale. Allegations that Canadian health care is currently either "underfunded" or suffering from "cost explosions" because of population aging, *i.e.*, demographic shifts combined with a pattern of

more or less constant age-sex specific utilization rates, are simply unsupportable from the demographic data. They form part of the rhetoric surrounding a struggle over political priorities - attempts to provide an apparently objective basis for one or other set of interests.

What *is* happening is that age-sex specific utilization rates are themselves changing, in such a way as to increase substantially the *relative* utilization of the elderly. Extensions of technological possibilities and increases in available manpower and facilities translate into increased intensity of health care servicing. And these increases occur to a greater extent among the elderly, as they are on average "sicker." It is easier to justify interventions as the organism slowly deteriorates -- there is always something wrong (Lubitz and Deacon 1982; Evans 1984). In the end, the process of dying provides a very extensive field for potential intervention. Rather than health care utilization being driven by "needs" associated with an aging population, it may be that developments on the supply side, technology and manpower pressures, are driving up use among the elderly. The usual (hypothetical) manpower planning process, as expressed thus far in this chapter, may be standing on its head.

### **PLANNING WITH MAGIC RATIOS: IMPLICIT ASSUMPTIONS AS TO EPIDEMIOLOGY AND TECHNOLOGY**

Age-sex breakdowns of the  $P$ -vector, such as underlie the discussion of population aging and manpower "needs," could be supplemented by information on other sub-populations of special need status -- high or low -- in order to refine the planning process. Birth rates, for example, have shown very large shifts independently of population age structure. With enough data one could also pinpoint and project high-use chronically ill or handicapped groups. Furthermore, cross-regional planning could make much more use than at present of population structure data. It should be obvious that a city like Victoria, B.C., with a large elderly population, needs a much higher volume of hospital and medical services per capita than does a young rural region.<sup>16</sup> Planning processes which try to compare manpower availability (and adjust funding) across regions solely on the basis of undifferentiated *capitas* will build in obvious inequities.

At the other end of the equation, the volume of resource input represented by a particular person or piece of equipment is not a given constant either. Obviously not all trained personnel are in the labour force at all times; some drop out temporarily or permanently and others work part-time. In the self-employed sector, effort and time levels per practitioner are variable. Age and sex explain some of this variation; insofar as they do it should be possible to define a stock of effective, or standardized Full Time Equivalent (FTE) personnel which would correspond to the  $R_k$  values of (13-4). Data for this purpose are less well developed than for population and utilization adjustment, but the conceptual problems are equivalent, and such work is underway.<sup>17</sup>

Much more difficult are the problems involved in determining the internal structure of the  $E$  and  $T$  matrices. The  $E$  matrix summarizes "what is to be done" -- how population "needs" are to be defined -- and the  $T$  matrix summarizes "how to do it." The magic ratio approach compresses the relevant rows of each into the  $\Theta_k$  parameter, the optimal  $X$ /Population ratio. The trick is, then, to identify at this level the optimum or target  $\Theta_k$ .

In practice it has turned out to be quite easy, as evidenced by the large numbers of such optimal ratios available for physicians -- all different. The traditional method was to identify a range of  $\Theta$  values across different geographical regions for a particular class of personnel  $k$ . The highest such  $\Theta$  was then selected as a standard, and manpower "needs" were defined as the number of personnel needed to bring all other regions up to the  $\Theta_k$  in the highest region.

Obviously, such needs estimates could be increased by making comparisons across smaller regions -- counties, say, instead of provinces.

Apart from its rather implausible assumptions about where new personnel would choose to locate -- why would they not distribute themselves in the same proportion as present personnel, preserving the measured "shortage" at a higher overall level of  $\Theta$ ? -- this approach rests on two further and fundamental assumptions. First, whatever services are being provided in the target or standard region are all "needed" and appropriate -- overservicing or an inappropriate mix of care is ruled out by assumption. And second, the process of provision in that region is accepted as an appropriate standard of technical efficiency. Since *all* the personnel in the most highly endowed region are "needed," they must all be working as efficiently as can reasonably be expected, and must be providing only care which is needed in the sense of being sufficiently effective relative to its cost to justify its provision. In terms of Figure 1-3, the slope of the curve even in the most highly endowed region (often referred to as most favoured, to drive the point home) is assumed steep enough to justify that level of provision, and providers are assumed to be on (or near) the curve, not below it.

Such faith is touching, and has been strong enough to survive dramatic increases in personnel availability. But it is buttressed by the fact that since the entire  $E$  and  $T$  structure is compressed into  $\Theta$ , these assumptions are wholly untestable. Only in the context of a much more detailed analysis of specific services, their patterns of production and provision, can the magic ratio be checked. And although more than enough work has been done on *individual* services to call both assumptions into serious question, if not refute them altogether, such work has rarely been assembled to the level of comprehensiveness that it could guide overall manpower policy.

Indeed, detailed service analysis can be carried out in such a way that it preserves the twin assumptions of the magic ratio. The Requirements Committee of the National Committee on Physician Manpower, reporting in the mid-1970s on physician requirements by specialty (Canada, NCMP 1975), used detailed billing data from the public medical care insurance plans to explore medical care utilization patterns. But it imposed as an assumption that whatever was being done was necessary. Its specialty working parties expressed some opinions (not always consistent) about who ought to do what, and some assumptions were made about technological and demographic factors which might increase future needs, but at root the approach was fully in the spirit of the traditional "magic ratio" -- whatever is, is right, or at least is not overdone.

On the technological efficiency side, the committee avoided any serious consideration of alternative forms of personnel as substitutes for physicians. There was some discussion of physician time requirements per procedure, but since these too were based on currently observed patterns, and since it was felt that physicians in general worked longer hours than they should, the result was predetermined that more physicians were needed to produce the same number of services. Magic ratio assumptions were thus preserved over an apparently more sophisticated analysis, to ensure that shortages would again emerge.<sup>18</sup>

## **DEFINING THE TECHNOLOGY: DIRECT MEASUREMENT OR THE "ENGINEERING" APPROACH**

It is much easier to criticize present or past efforts at manpower planning, which are always constrained by available data, time, resources, and conceptual apparatus, than to do the job "right." It is also much less dangerous, since attempts to provide a detailed structure will inevitably lead to many and detailed errors and points of criticism. "Magic ratios" bundle all the

individual errors into one large fallacy, which is harder to critique because it is harder to come to grips with.<sup>19</sup> There are two general classes of approach to measuring the structure of the  $T$  matrix at least, the "engineering" and "econometric," which arise out of the economic theory of production and which have yielded some useful results (Hadley 1974; Reinhardt 1973).

First, in the "engineering" approach one can attempt to measure the  $T$  matrix directly, by observing actual practice data, consulting with experts, conducting experiments, etc. For this purpose, however, the matrix of (13-4) is too restrictive and must be augmented. It embodies only *one* technology, one way of producing each service. Further, that one possible technique is assumed to display constant returns to scale; proportionate increases in all inputs yield an equi-proportionate increase in output.<sup>20</sup>

To modify (13-4) we must add additional columns, new technologies, to  $T$ , reflecting the fact that there may be several ways of producing  $S$  with different mixes of inputs. Some of these alternatives may use inputs -- nurse practitioners, denturists -- which are not used at all in the other approaches, in which case  $q$  is increased to span a wider range of input types, and rows as well as columns are added to  $T$ . The result is represented as:

$$\begin{bmatrix} R_1 \\ \cdot \\ \cdot \\ \cdot \\ R_k \\ \cdot \\ \cdot \\ \cdot \\ R_q \end{bmatrix} = \begin{bmatrix} t_{11}^1, t_{11}^2, t_{11}^3 & \text{---} & \text{---} & \text{---} & t_{1m}^1, t_{1m}^2 \\ & & & & \\ & & & & \\ & & & & \\ & & & t_{kj}^p & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ t_{q1}^1, t_{q1}^2, t_{q1}^3 & \text{---} & \text{---} & \text{---} & t_{qm}^1, t_{qm}^2 \end{bmatrix} \times \begin{bmatrix} S_1^1 \\ S_1^2 \\ S_1^3 \\ \cdot \\ S_j^p \\ \cdot \\ \cdot \\ S_m^1 \\ S_m^1 \end{bmatrix} \tag{13-10}$$

where as shown there are three alternative techniques available for producing service  $S_1$ , using different combinations of inputs  $R$ , and the total amount of  $S_1$  produced will be  $S_1 = S_1^1 + S_1^2 + S_1^3$ , the sum of the amounts produced using each technique. Similarly there are two alternative ways of producing  $S_m$ . The total number of columns of  $T$ , and rows of augmented  $S$ , will depend on the number of alternatives available for each service, which will vary by service.

This approach enables one to represent opportunities for input substitution within the activity analysis framework, though it continues to impose constant returns to scale.<sup>21</sup> It does, however, mean that resource requirements cannot be determined solely on the basis of technical and epidemiological information, which is as it should be. The optimal allocation of production of each  $S_j$  among the alternative technologies available is essentially an economic decision, and should respond to the relative costs of the inputs  $R_q$ . The cost per unit of  $S_j$  will be equal to the cost of all the inputs used in its production, and these will in general differ across techniques.

It does not follow, however, that one "best" technique will necessarily dominate all the others, such that only one will ever be used. Since resources come in lumpy units (such as people) which are not freely divisible, and have alternative uses, and since geographic considerations (at least) will constrain the size of production units -- the Canadian health care system cannot all be put into one huge building just outside Winnipeg -- it may turn out to be optimal to use mixes of techniques, or different techniques in different settings. The point to

emphasize, however, is that the analytic framework must embody the *possibility* of substitution, and that economic (relative cost) criteria must enter this process in a central way.

Studies of this sort have been carried out in considerable detail for some sectors of health care. A computerized model of dental practice, for example, has been developed which not only identifies all the specific services of a dental practice but breaks them into sub-functions which may (but need not always) be included, and assigns time requirements for each (Kilpatrick *et al.* 1972). Personnel can be "custom-designed" for the model, in terms of the bundle or package of capacities they embody, and realistic problems of queuing, personnel co-ordination, and transfer from patient to patient can be built in. It is thus a much more sophisticated and realistic version of the above model framework, and can be used to determine global manpower requirements under different assumptions as to practice size, structure, and location.<sup>22</sup> In other branches of health care, the application of this framework has been less detailed, but it has been applied to primary medical care (Smith *et al.* 1972) and at a much more aggregated level, to pharmacy (Evans and Williamson 1978).

The strengths of this approach are also its weaknesses. Precisely because the framework provides opportunities to exploit extensive and detailed information, it also *requires* such information. Measurement of the  $T$  matrix has usually been done in one or a few specific sites, whose representativeness may be questionable. Practices or clinics which offer themselves for study are unusual by definition. Yet the extent of "engineering" information required makes it unlikely that one could ever apply the technique over a large number of practices "in the field." Measured or estimated  $T$  matrices demonstrate what is possible, but not necessarily for everyone or under all circumstances. And one is inevitably left with the concern that the large discrepancies which show up between such studies and actual practice may reflect constraints in the field that the experimental setting has neglected.<sup>23</sup>

## **STATISTICAL OBSERVATION OF PERFORMANCE IN THE FIELD: THE "ECONOMETRIC" APPROACH**

The alternative approach takes as its basic units of observation individual production units, not specific services. Variations in patterns of manpower or other inputs used across medical practices, dental practices, hospitals, pharmacies, etc. are identified and linked to variations in output use. This amounts to direct estimation of the production function of (13-9), relating quantity of output in each practice to the amounts of different inputs which it employs.

Immediately, however, one is faced with the problem of defining output. Equation 13-9 referred to a single type of service,  $S_j$ , such as an item in a fee schedule, for which it makes sense to think of one or several alternative, but specific, production processes.  $S_j$  measures a quantity of (more or less) homogeneous units. But what is the "output" of a medical or dental practice, or a hospital? Obviously, it is a vector, a range of different services or commodities in different amounts. If each practice, or unit of observation, produced the same range of different services in the same proportions, then we could measure the "output" of each by focussing on its production of any one service -- say office visits. But they do not, and such portmanteau concepts as an office visit or a hospital day or stay may conceal a great deal of inter-practice or inter-institution variety.

Given an extremely large number of observations, as well as detailed output information on each, one could probably take direct account of this variation in output pattern.<sup>24</sup> Instead, however, analysts use a single scalar measure of practice output,

$$Q^v = W_1 S_1^v + \dots + W_m S_m^v \quad (13-11)$$

where now  $Q^v$  and  $S^v$  refer to the output index and the services of different types produced by the  $v$ th practice or production unit. The weights  $W_j$  must be the same for each practice. If each practice faces the same set of service prices -- or a common fee schedule -- as do medical practices in each of the Canadian provinces, then we may use these fees as weights:

$$B^v = Q^v = f_1 S_1^v + \dots + f_m S_m^v \quad (13-12)$$

Then total practice billings are a proxy for total output, whose usefulness depends on the accuracy with which the  $f_j$ , specific service fees, reflect the relative resource costs of the different services -- including self-employed practitioner time. In an environment where fees are variable across practices, one must deflate practice billings by some index of the relative fees charged in that practice, if available. Alternatively, one may abandon billings entirely and use one or a handful of specific classes of services (with weights assigned a priori by the analyst) as representative of the practice as a whole, in the hopes that other service outputs will either move more or less in proportion, or be "small."

However the output definition problem is addressed (and like the problem of representativeness of detailed practice data, it can only be addressed, not solved), the analyst assembles data:

$$\frac{Q^1, R^1_1, R^1_2, \dots, R^1_q}{Q^v, R^v_1, R^v_2, \dots, R^v_q}$$

on each of  $v$  practices, indicating their levels of use of each of the classes of input, and resulting index of output. She then seeks some mathematical formula which will combine the  $R^v_q$  in each practice to produce an estimated level of output  $\hat{Q}^v$  for that practice. A "good" formula is one which leads to small discrepancies between actual observed  $Q^v$  and the  $\hat{Q}^v$  given by the formula.

A common approach is to try to minimize  $\sum(Q^v - \hat{Q}^v)^2$  or the sum of the squared deviations between actual and "predicted" output. For any particular formula the analyst chooses, statistical techniques can be used to select parameters to achieve the "least squares" fit, but the choice of the formula itself, the functional form, is part of the black art of econometrics.<sup>25</sup>

However it be done, the end product is some relationship, say, for example:

$$Q^v = \hat{Q}^v + e^v = A(R_1^v)^\alpha (R_2^v)^\beta (R_3^v)^\gamma + e^v \quad (13-13)$$

where  $A, \alpha, \beta, \gamma$ , are the same across all observed and analyzed practices, and  $e^v$  is a measure of the extent to which the formula fails to match the actual output of the practice. Small values of  $e^v$  relative to  $Q^v$  suggest that the formula is a good representation of the way in which the various types of inputs, personnel, capital equipment and/or supplies, are combined to yield service Output.<sup>26</sup>

Such a relationship, once found, is merely one way of expressing the "production function," the relationship between productive resources and outputs of goods and services. The technology matrix  $T$  is similarly a production function, or set of them, as is a model of a practice designed

for computer simulation. The single equation analytic representation of this relationship is popular among economists because of its compactness and analytic tractability, but it has the vices as well as the virtues of simplicity.<sup>27</sup>

For the purposes of manpower planning, however, an analytic representation such as equation (13-13) can be helpful. It permits us to measure the productivity, in actual practice experience, of different mixes of inputs. In this way, it parallels the type of evidence which arises when specified practices are augmented by additional and/or more highly trained personnel. Letting  $R_k$  be some form of intermediate-level personnel, for example, equation (13-13) enables us to read off the increase in practice output which results from increasing by one the number (FTEs) of such people in a practice for given numbers of other inputs.

The analysis generalizes to multiple inputs, divisible into fractional FTE values. It provides a mathematical representation of the extent to which different inputs -- such as professionals and auxiliaries -- can substitute for each other. The relationship (in more sophisticated forms than 13-13) can embody increasing or decreasing returns to scale over different levels of output. It will usually indicate declining marginal productivity for a particular input resource; if, for example, professional and auxiliary time were the only inputs in production, then as the auxiliary to professional ratio rose, the contribution to output of additional auxiliaries would tend to fall.<sup>28</sup> But in general a given volume of services may be produced with very different mixes of inputs, particularly manpower inputs.

## **FOR WHOM IS EFFICIENT MANPOWER USE OPTIMAL?**

What is possible, however, is not necessarily optimal. At the level of the individual practice, optimality is achieved when the cost of adding one more unit of an input just balances its contribution to revenue -- marginal outlay equals revenue product. (This makes the strong assumption that the practice objective is profit-maximization). If auxiliaries are available in a reasonably free market at a constant wage, marginal outlay equals that wage, plus perhaps some additional administrative overhead. Underutilization of auxiliaries (or any other input) would be indicated if the amount of additional practice revenue which an additional auxiliary would generate exceeds the wage plus overhead cost of that auxiliary.

A number of such studies have indicated underuse in this sense. Using an estimated relationship similar to (13-13), plus information on the relative wage rates of professionals and auxiliaries, one can demonstrate what the optimal mix of inputs would be -- optimal in the sense of maximizing practice net income per professional -- and thus what the physical output of services per professional could be. The results generally indicate substantial potential for gains in output per professional. Any given level of aggregate output, system-wide, could be produced with fewer professionals and more auxiliaries, and at lower overall cost.

This leaves an important question as to why the existing system leads to underutilization. This issue has been dealt with from a different perspective in chapters 6 and 7; here our focus will be on the implications of this observation for manpower policy.

One suggestion is that the wage rate understates the cost of using auxiliaries in individual practices because professionals do not like to work with, or supervise, large teams. If so, the underuse represents a form of non-money income or satisfaction for professionals, which their strategic position in the management of provider firms enables them to indulge at the cost of the public generally.

Alternatively, however, one should note that equation (13-13) does not in fact describe the relationship of inputs to practice *revenues*, but only to physical outputs. If one can assume that additional output will be reimbursed at the same price (fee) as current output, then it is legitimate to compare physical productivity, multiplied by those fees, to wage costs. But this assumes that each practice faces infinitely elastic demand at the going price. It may be, however, that increased output must be "marketed." This could be, in the rather implausible neoclassical model, by price-cutting and competitive behaviour. More plausibly, particularly in the Canadian setting where each practice faces the same fixed fee schedule, the practitioner must adjust practice style, recall patients more frequently, and generally invest in expanding utilization to match capacity. In either case, however, costs are incurred which may be monetary, or more likely, direct disutility. "Marketing" and entrepreneurial behaviour generally are considered unprofessional. The net effect may be that for an individual practice, expansion may be "unprofitable" in a total utility sense, even if in purely technical terms, less costly outputs could be substituted for more.<sup>29</sup>

These considerations greatly complicate the manpower planners' problem. The cross-practice studies tend to confirm the non-optimality -- excessive costliness -- of current manpower mixes which emerges from direct analyses of health care production. Their evidence on this score is particularly important because the econometric methodology is subject to severe conservative biases. In particular, it cannot show the potential for manpower substitution by types of personnel not now in use because of legal, habitual, or other restrictions. Equation (13-13) reflects only what is, not what could be. Further, it builds in all the arbitrary behavioural patterns of individual practitioners and treats them as technological constraints. A common finding, for example, is that differentiation among different types of auxiliary personnel -- high and low skill -- does not improve the fit of equations like (13-13), so all are lumped together. Yet we know that the productivity of these different people is potentially very different; the problem arises in the way they are being used. Different patterns of use in the United States of females (nurses) and males (physician assistants) with similar training suggest the same problem. Statistical studies of actual practices thus indicate the bare minimum of possibility.<sup>30</sup> The activity analysis models, based on identifying patterns of service needs and the technical capabilities of different types of personnel, suggest that one can go much farther.

### **THE ONLY THINGS WE LEARN ...?**

But the behavioural and "marketing" limits raise the very awkward question: How? Suppose one had all the data necessary to fill in the matrices of equations (13-6) and (13-10), and had made the political decisions involved. Educational programs were then re-ordered to produce the manpower needed in the optimal mixes and amounts ( $M_1$  ---  $M_q$ ), allowing for the differential productivities of the different constituents of each stock  $M$ . What then? If individual provider units make use of non-optimal mixes of manpower, for (economic or non-economic) reasons which seem to them good, the results will be under- or unemployed personnel of some types, and shortages of others. The motivations of individual decision-making providers may lead to patterns of macro-behaviour which are globally inefficient, but they cannot be ignored.

The sad history of the nurse-practitioner in Canada illustrates this problem: there is no point in training people, however cost-effective, whom practitioners will not wish to, or cannot afford to, hire. Of course the picture is clouded by the excess supply of physicians; a serious policy of manpower substitution would have cut back on physician supply even as new types of personnel

were being produced. If that had been done, the innovation might have had more chance. But one cannot be sure that an attempt to induce a "shortage" of medical services under present production patterns would have stimulated a shift to the more efficient model. A number of other responses, less satisfactory and politically more uncomfortable, could easily be imagined.

One is left, then, with the fairly obvious conclusion that manpower policy, however sophisticated, cannot be developed independently of the structure of the delivery system. If changes in type, quantity, and mix of personnel are desired, then they must somehow be fitted in, either to the present or to some modified form of care provision. Hence the monumental stability of health care delivery through decades of apparent change. Present manpower training decisions determine future production choices over as long as a generation, by the "human capital" they create. Yet present patterns of delivery severely constrain manpower choices by determining who will and will not be permitted to provide services, in what settings. Manpower policies, other than simply quantity adjustments, require simultaneous planning for and intervention in delivery, if they are to be effective.

Accordingly manpower issues make up the most difficult, far-reaching, and critical of the public investment decisions made in health care. It is no wonder they are so frequently made unconsciously, by delegation, or by default. What is perhaps remarkable is that the results are not worse than they are.

## NOTES

<sup>1</sup> Or worse, ill-conceived neologisms such as personpower.

<sup>2</sup> Nor need the "human capital" represent specific skills. Health and general education spending have often been advocated as "investments" which increase the subsequent productivity of the labour force. Healthy educated people represent more "capital" in the sense of ability to do work, than ill or ignorant ones. The heavy concentration of health spending on the elderly or chronically ill, and the significant consumption aspects of education, weaken this argument somewhat. It can be restored by expanding the concept of "human capital" to include sources of direct satisfaction to the holder -- the stock of "health capital" as both an object of investment and a source of future benefits, for example. But despite its theoretical elegance, this approach appears to have generated more confusion than clarity.

<sup>3</sup> Though perhaps nowhere else.

<sup>4</sup> Actually, this is an overstatement. Some consumers may derive direct satisfaction from being cared for by the "high priced" help; others may treat excess or underutilized human capital as a signal for superior quality, despite the absence of supporting evidence (and the presence of contrary evidence). If so, some providers will find it profitable to offer services using more expensive personnel -- at a higher price. Consumers will be able to select from different levels of input capital intensity, but therapeutic equivalence. The market provides Cadillacs and Chevrolets, Toyotas and Mercedes, but all will take you where you are going.

<sup>5</sup> Obviously the model should allow for uncertainty with respect to earnings, interest rates, maybe even success of entry attempt. Furthermore, individuals have inherent differences in physical and intellectual endowments which affect their expected payoffs. These complicate the conceptual model, without changing its implications. They do, however, severely inhibit empirical testing. But then, most testing is in practice carried out in terms of explicit earnings, which amounts to the assumption that "net advantages" are always proportionate to money income. The implausibility of this assumption means that empirical rate-of-return calculations across occupations can only be indicative at best.

<sup>6</sup> The U.S. educational system displays some of the characteristics of a market, but relatively few. A number of professional training schools are moving toward, or have established, full-cost tuition charges, and U.S.

governments, federal or state, do not have the same degree of influence as Canadian provincial governments over number and capacity of training programs. But the *supply* side of professional education bears no resemblance to a competitive market, with free entry of (possibly for-profit) training institutions designing their programs to provide whatever the market is willing to buy. So long as public or delegated regulation controls the qualifications required to enter a profession, it simultaneously controls "approved" institutions and curricula. To observe a genuinely competitive educational system at work, one must go back to pre-Flexnerian days. Starr (1982) provides a first-class description and analysis.

<sup>7</sup> The equality of present values holds *at the margin*. People whose inherent abilities are matched to a particular occupation will find that they can earn a positive, perhaps very large, *PV* in that occupation relative to the next best opportunity -- the Wayne Gretzky effect -- because these advantages, being unique, cannot be competed away. The equalization of *PVs* occurs after adjustment for "rents to abilities" -- which further inhibits empirical testing. Equalization also depends on a continuous distribution of abilities; "holes" in the distribution may lead to unequal *PVs*.

<sup>8</sup> Eventually. But as is well known, the fact that a static equilibrium may exist is no guarantee that dynamic processes will get you there quickly -- or at all (Nelson 1981). Working off a surplus may take half a generation -- it depends on the relative sizes of stocks and flows of personnel and the magnitude of necessary adjustments. Furthermore, new entrants must be fully informed, not only about current and future demands for personnel, but also about supplies, which will depend on the decisions of other potential entrants. They need to know what everyone else is going to do in order to make their own decisions optimally. This interdependence can lead to dynamic over- or undershoot, depending on how entrants assess the behaviour of other potential entrants. And the existence of an eventual equilibrium a generation or two hence is small comfort to those who guess wrong -- in the long run we are all dead.

<sup>9</sup> There are a few economists still out growling in caves in the woods somewhere that the whole system would function better if the state withdrew entirely, but they are not taken seriously even by themselves. The real policy issues, as recent U.S. experience has made clear, are: "Who shall control the state's interventions, and for what ends?" For tactical, political reasons, some forms of intervention are sometimes referred to as "de-regulation" or "free enterprise"; the deception is probably deliberate.

<sup>10</sup> In Canadian universities, program fees charged to students vary much less than program costs, so the proportion of program costs which are subsidized, though very high on average, varies greatly over programs. In general, the subsidy rate is highest on the most expensive programs, which also lead to the highest paying careers, and are lowest on the least costly and least economically rewarding (to the individual). Justifications for this pattern are not immediately apparent.

<sup>11</sup> Alternatively, provincial regulations may, as is all too common among professionals generally, limit the transferability or reciprocity of licensure. This seems far too high a price to pay for any effect it might have in improving the effectiveness of provincial manpower policies. The advantages of professional "Balkanization" from the point of view of professionals in migrant-attractive provinces are obvious. But it is offensive from every other perspective.

<sup>12</sup> This statement is obviously false. Of course it can, and usually is. The results, however, tend to be rather unfortunate.

<sup>13</sup> The vector of  $P_i$  is defined at a single point in time, but obviously manpower planning requires, explicitly or implicitly, assumptions about how  $P_i$  unfolds through time. Total population  $P$  will (usually) rise, but its proportions will also change, most obviously as its age and sex structure evolves. They may also be endogenous, as levels of service provision at one point affect -- positively or negatively -- later needs. Neonatal intensive care may increase the proportion of the population with high future needs; particular forms of prevention may reduce it.

<sup>14</sup> Clearly, the  $S$  and  $e$  must refer to some time period, say a year. Then  $S$  has dimension services/year, and  $e$ , services/year per person.

<sup>15</sup> These should not, however, be assumed to be optimal.

<sup>16</sup> Though not necessarily as many as it has.

<sup>17</sup> The Health Manpower Research Unit at the University of British Columbia, for example, is preparing detailed FTE stock data by health occupation, and related analyses for physicians have been carried out at the Department of National Health and Welfare and in other provincial ministries.

<sup>18</sup> The finding of a surplus would of course, be threatening to practitioners as well as to medical school staff -- it raises awkward questions about how practitioners, who are obviously not unemployed, spend their time.

<sup>19</sup> The optimal ratio of people per physician is 612. No it isn't ... and so on.

<sup>20</sup> The more general economic approach is to express the production relationship as a production function:

$$S_j = h^j(R_{1j} \dots R_{qj}) \quad (13-9)$$

$$\text{subject to } \partial h^j / \partial R_k \geq 0$$

saying merely that the volume of possible output of  $S_j$  depends in an unspecified way on the amounts of the different inputs used for  $j$  production, and more inputs yield more, or at least no less, output. (13-4) is then a special case of (13-9), since holding all other service levels constant we can solve for  $S_j$  as a function of the  $R_k$  available (beyond that needed for the pre-set levels of other  $S$ ).

<sup>21</sup> The isoquants are faceted, not smoothly curved, but if several techniques exist and linear combinations are permissible, plenty of scope for substitution is allowed. Constant returns to scale is probably not too serious a restriction either; apparent scale effects in health care production often turn out to result from product mis-specification.

<sup>22</sup> It was used for this purpose, by the B.C. Children's Dental Health Research Project (British Columbia, CDHRP 1975) and demonstrated that a given pattern of "needed" services could be provided in radically different ways depending on how practices and personnel mixes were organized. Total, system-wide costs of most efficient combinations were as much as 40 percent below those of standard practice organizations, and implications for training programs were massive. Not surprisingly, the results were shelved.

<sup>23</sup> That this may be *possible* does not imply that it can be assumed. Pangloss-style circular reasoning can lead one into the trap of believing that if more efficient manpower use were possible, rational, self-interested practitioners in the field would have discovered and adopted such techniques. Therefore, any discrepancy between the field and the "engineering" analysis indicates flaws in the latter. As chapters 6 and 7 demonstrate, this would be theoretically untenable even if managers in the field were fully rational, fully self-interested, and unregulated -- which they are not. There are good economic reasons to believe the models may be right.

<sup>24</sup> Perhaps instead of estimating  $S = h(R_1 \dots R_q)$  by multiple correlation techniques, one could estimate  $g(S_1 \dots S_m) = h(R_1 \dots R_q)$  by canonical correlation. I do not know, and do not believe anyone has.

<sup>25</sup> This is the same process as is carried out in fitting hospital cost functions; see chapter 9.

<sup>26</sup> There is, of course, a good deal more to the story than this. Equation 13-13 is not necessarily the best functional form available, only one convenient for representation. Three classes of inputs involve a disturbing degree of aggregation -- physician, non-physician, and square feet of office space, *e.g.*, or dentists, non-dentists, and drills -- of inputs which may have widely differing characteristics and capacities. But the data available on each practice, system-wide, is often insufficient for greater detail.

<sup>27</sup> Indeed there is a serious methodological question as to whether such relationships exist in any causal or structural sense, particularly when estimated at the aggregate, economy-wide level, or whether their "estimation" is merely picking up parallel trends in economic aggregation. At the level of more homogeneous activities, primary medical practice, *e.g.*, or dental practice, distortions imposed by aggregation are much less severe. But considerable caution is warranted -- a single-equation production function fitted across acute care hospitals, for example, may be more misleading than informative.

<sup>28</sup> Returns *to scale* refers to the pattern of output response when all inputs are increased or decreased in proportion; increasing, decreasing, and constant returns to scale describing greater, lesser, or equi-proportionate output change. Diminishing returns to a single input or factor refers to a tendency for increases in output to become progressively smaller as units of a particular input are added to production, holding all others fixed in supply. A combination of constant returns to scale but diminishing returns to each factor is built into the *T* matrix, being most pronounced in the expanded form of equation (13-10). Its derivation, however, is a bit more involved.

<sup>29</sup> In the jargon, practice profits are maximized, with respect to any input, when marginal outlay (*MO*) = marginal revenue product (*MRP*). (We assume that as utilization of any input increases, *MO* is either constant or rising, *MRP* is constant or falling.) If input markets are competitive, *MO* = Wage (*W*) -- inputs can be hired at a constant unit price each, however many are hired. If output markets are competitive (or fixed fee), *MRP* = value of marginal physical product (*VMP*), and *VMP* = Price of output (*P*) times marginal physical product (*MPP*). Equation 13-13 enables one to calculate *MPP*, and optimal input is chosen where  $MPP = (W/P)$ . (As input use increases, *MPP* starts high and falls). But if *P* falls as output increases (the practice faces a downward-sloping demand curve for its own output) then  $MRP < VMP = P \times MPP$ , because increasing output requires price cuts. So  $MRP = W < P \times MPP$ , and  $MPP > (W/P)$  -- less input will be hired. The same effect will arise if *P* is nominally constant -- fixed fees -- but costly marketing effort (in money or pride) must be undertaken to expand the practice.

<sup>30</sup> Of course if all practices were profit-maximizers operating in a fully informed, fully competitive (which includes free entry) market environment, such behavioural distortions would have been squeezed out by competition. But if that dream-world model applied, one would not be concerned about manpower issues in the first place.