SQUARING THE CIRCLE:
RECONCILING FEE-FOR-SERVICE
WITH GLOBAL EXPENDITURE CONTROL

HPRU 88:8
Health Policy Research Unit
Division of Health Services Research and Development
Office of the Coordinator of Health Sciences
The John F. McCreary Health Sciences Centre
The University of British Columbia
Vancouver, B.C.
V6T 1Z6

R.G. Evans
August 1988
Acknowledgements

Advice and encouragement is gratefully acknowledged from M.L. Barer, M.D. Low, the B.C. Medical Manpower Advisory Board, and the alumni of the UBC Health Services Planning Program. The author’s research is assisted by a Research Scientist Award from the National Health Research and Development Program.
Bargaining over Physicians' Fees

When Canada's public health insurance programs were extended in the 1960s to cover physicians' services, they adopted the process of reimbursement which had been developed by their private, physician-sponsored predecessors, the not-for-profit members of Trans-Canada Medical Plans (Shillington, 1972). Each provincial program pays physicians for their services according to a uniform fee schedule specific to that province. Under the public plans, however, the schedule is negotiated between an agency of the province - ministry of health or commission - and the provincial medical association. In earlier years the medical associations simply promulgated their "Schedules of Minimum Fees" independently.

As a general principle, the medical associations set the relative fees of items in the schedule. Negotiations focus primarily on the overall percentage increase in total outlays, the weighted average increase in fees, which is implicit in the pattern of increases in individual fee items. In the simplest description, medical associations determine the relative values of the different items, and bargain over the conversion factor for translating relative value units into dollars.

This description, however, neglects a critical aspect of the process. The overall increase in payments to physicians implicit in a fee schedule change is inevitably an estimate, based on the application of the new schedule to past data on the numbers of different fee items billed. Actual provincial outlays, which represent the gross receipts of the fee-
practice medical profession, are the product of both the level of fees and
the rate of billings or utilization of services during the period of
application of the schedule.

Thus even with a binding fee schedule, a provincial government's
financial liability remains open-ended. An estimate of outlays will be
prepared as part of the government budgeting process; but unlike the
allocation of funding for hospitals or for the other public health
activities of government, the total payments to be made to physicians are,
strictly speaking, beyond the control of provincial governments. The
bills must simply be paid as they come in. In this limited sense,
expenditures are indeed "out of control".

Total outlays are not, however, wholly arbitrary or unpredictable.
They are determined, in an accounting sense, by the numbers of active
practitioners in the province, and the activity level or volume of
billings of each. The latter, in turn, depends on the time and effort
which each physician puts into his/her practice. But it also depends on
the specific characteristics of the fee schedule, and the way in which
these interact with the patient flow and practice styles of practitioners
to yield an implicit rate of "billings per practitioner-hour".

Depending upon its structure and associated rules for payment, a fee
schedule may provide greater or lesser opportunities for practitioners to
expand their gross receipts by changes in practice style or billing
behaviour. Thus, while fee-for-service reimbursement represents, in principle, an open-ended commitment of public funds, the degree of "open-endedness" is itself influenced by the negotiation process. In general, provincial governments have attempted to ensure that fee schedules were structured so as to minimize the opportunities for practitioners to expand their billings without a corresponding increase in time and effort.

The success of these attempts has been variable. Over most of the history of the public plans the average billings per physician reimbursed have escalated at about one to two percent per year faster than average fees. But there have been dramatic examples of "utilization creep" through procedural relabelling and multiplication, most notably in Quebec between 1971 and 1975 (Contandriopoulos, 1986; Gabel and Rice, 1985). Physicians there were responding to a freeze on nominal fees during a period of rapid general inflation. Average gross receipts per physician rose 17 percent over this period, or 4 percent per year (Barer, Evans and Labelle, 1988).2

Until recently, however, most provincial governments have been able

---

1 Schedules which permit "à la carte" billing for minor diagnostic and therapeutic procedures performed in the course of an office visit, for example, allow physicians to increase their incomes by performing such procedures more frequently. Schedules which draw fine distinctions between different types of visits or procedures provide an opportunity for physicians to reclassify their services into higher-paying categories.

2 The increase in total expenditures on physicians' services in Quebec, divided by the reported number of "active civilian physicians" in Quebec, rose 30.5 percent over this period, or 6.9 percent per year. The number of physicians reported as being reimbursed by the RAMQ rose much more rapidly than the total Quebec physician stock.
to contain "utilization creep" within acceptable limits through the fee negotiation process. The open-ended nature of fee-for-service reimbursement has been a significant, but not an explosive, factor in overall cost escalation. But that process is a dynamic one, an on-going game rather than a once-for-all settlement. Changes in medical technology, in physician attitudes and practice styles, or in the general political climate, all have the potential to upset the balance.

In recent years, provincial governments have become increasingly concerned about the ever-present potential for large and unforeseen cost overruns in one of their largest budgetary items. Furthermore, the probability of such overruns may be perversely related to the general fiscal situation. When provincial budgets are tight, negotiations with physicians are tougher. But a very hard-nosed bargain is more likely to result in an unexpected bounce in utilization, as physicians look harder for ways to exploit the billing opportunities in a given schedule. As provincial governments give a higher priority to cost containment in health care, the open-ended nature of fee-for-service is becoming more problematic.

Even if large "utilization break-outs" do not occur, provincial governments are still exposed to the cumulative effects of annual increases in billings per physician which regularly exceed negotiated fee increases. Over a period of years, and on a very large expenditure item, relatively small slippages compound to large sums.
And finally, in addition to increases in fees and in (fee-adjusted) billings per physician, provincial outlays are driven upward by the increasing supply of physicians. While physicians argue over whether their increased numbers actually cause increases in utilization, and particularly bitterly over whether such increases may represent "unnecessary servicing", two facts are indisputable.

First, the supply of physicians is rising faster than the population (whether or not adjusted for changing age structure), by one to two percent per year, and is projected to do so for the indefinite future (Barer, Gafni and Lomas, 1988). And second, the average volume of utilization per physician - fee-adjusted billings - is continuing to rise. There is no sign of a saturated market; utilization and costs of medical services are rising in line with the numbers of physicians (Lomas, Barer and Stoddart, 1985; Barer and Evans, 1986; Lomas and Barer, 1986).

Bargaining over Medical Utilization

Accordingly, provincial governments are becoming increasingly concerned about the factors underlying the utilization of physicians' services, and the prospects for influencing or managing them (Anderson and Lomas, 1988; Rachlis and Fooks, 1988). But they are also trying to limit the open-endedness of the reimbursement system. Quebec adopted this strategy over a decade ago, in response to the experiences of the early 1970s, negotiating agreements limiting payments at both the aggregate and the individual physician level (Contandriopoulos, 1986). B.C. began
negotiating caps or "givebacks" in 1982 (Barer et al., 1987; Barer, Evans and Labelle, 1988).

In effect, provincial governments are shifting towards bargaining over total outlays, rather than just over fee level increases. As noted above, in earlier years bargaining over fee schedule structures and rules of payment was always carried out with a close eye on implicit (expected) total provincial outlays (or, from the physicians' point of view, implicit average gross incomes). But the global constraints are becoming more explicit and binding.

The logical limit of such a process is contemporaneous prorating. Under prorating, gross outlays for a province and time period are negotiated in advance as a fixed amount. If total claims for services provided during that period differ from the predetermined amount, the rate of reimbursement per claim is scaled up or down so as to equate total claims and agreed outlays. In effect, the conversion factor for translating billing dollars into reimbursement dollars becomes a floating value, and the fee schedule is explicitly a relative value scale. This process is rather different from what is currently going on in B.C. and Quebec, but the actual outcome is becoming increasingly similar.³

³ There is a difference in principle. Under current arrangements, if utilization increases are not large enough to raise total outlays to the "cap" amount, no adjustment is made in favour of physicians. The prorating scheme is symmetrical - physicians will receive the total amount bargained for, even if they bill for less. So long as the "caps" are binding, however, this distinction is without practical significance.
Physician reimbursement systems in the Canadian provinces thus appear to be moving toward an indirect form of prorating. This paper does not argue for or against that development, or attempt to describe in detail the present state of bargaining. Rather, it takes the trend as given, and assumes that the underlying forces are likely to persist. In this context, we suggest a more explicit and sophisticated form of prorating which may mitigate some of the inequities inherent in simpler schemes, direct or indirect. It also embodies some incentives for modifications in patterns of medical practice which may be considered desirable on other grounds, and provides practitioners, not individually but in local groups, with more scope for control over their own practice patterns.

The Basic Structure of a Regional Prorating System

Under simple prorating, the negotiations between provincial agency and medical association yield an explicit value, $B$, for budgetary outlays on physicians' services during a given year. They would also establish a fee schedule to govern the item-specific reimbursement process. In the course of the year, actual claims submitted would total $C$. The ratio $R = B/C$ is then the scaling factor to be applied to all fees, to determine actual amounts paid. In practice, of course, payments are being made throughout the year, so the scheme would require on-going monitoring of outlays against targets, and some form of end-of-year adjustment, up or down, in reimbursements per practitioner.

The potential problems with this approach are fairly well known. It
may be acceptable as a temporary measure, but if applied over a period of years it places great pressure on all practitioners to change toward procedure-intensive styles of practice. The reimbursement system is a process of "beggar-my-neighbour", or at least professional colleague, and the practitioner who tries to avoid exploiting to the full the billing opportunities in a given fee schedule loses out economically to those who are less inhibited. As this lesson is learned, total claims will escalate and the conversion ratio will fall; eventually the whole scheme may lose credibility.

Unfortunately, it seems logically unavoidable that any scheme to try to reconcile fee-for-service reimbursement and overall cost caps must have this beggar-my-neighbour aspect, although the process of negotiation and application may, as at present, hide the fact for a time from some of the participants. The proposal outlined here shares this characteristic, but it provides a few more degrees of freedom in the process.

The first change from global prorating is regionalization. Budgetary targets are set for each major region in the province. But they are not bargained separately, nor are they linked to the physician numbers in each region. Rather the overall negotiated amount for the province, $B, is allocated according to the population in each region, taking account of their age and sex mix. Thus regions are, in effect, capitated, and (subject to age and sex adjustment) the same amount of reimbursement is allocated on behalf of each resident of the province, regardless of where
he or she lives. This allocation then reimburses all the services provided to regional residents, by physicians located anywhere in the province (or out of it).

The population-based allocation, however, represents a significant policy decision. One could, instead, choose to allocate shares of the provincial total to regions on the basis of the numbers and billings of physicians currently located there. For example, one could designate each physician’s office address, and then allocate the total of $B for the current year in the proportions in which total outlays were split among regions last year, using the office location of the physician reimbursed as the indicator of the region to which the funds were sent.

The resident-based approach implies a judgement that the purpose of the financial allocations is to reimburse services on behalf of residents, rather than to provide support to particular physicians. It treats all provincial residents equally, regardless of where they are located.

The derivation of regional allocations of funds proceeds by first calculating age-sex specific relative rates of utilization of medical

---

4 There is no reason in principle why age and sex should be the only factors taken into account in determining the relative per capita allocations. If other descriptive data on regional populations are available which show a clear-cut correlation with utilization of physicians’ services, such other factors could easily be built into the relative values, without changing the total provincial allocation $B. (It would not, however, be appropriate to include the region-specific physician to population ratio as one such variable! The data must be descriptive of the population, not the health care system.)
services. The total provincial population is partitioned into age and sex groups, which might be ages 0-1, 1-4, 5-9, and so on by five year blocks up to 100. This partition yields 42 mutually exclusive and collectively exhaustive age-sex categories.

For each of these categories, we can compute a relative rate of utilization of (or at least expenditure on) medical services. This is equal to the average per capita expenditure on physicians' services for all persons in the province in that age-sex group, divided by the average per capita expenditure for the whole population. These relative rates are well below one for males aged 10-14, for example, and well above one for females (or males) aged 70-74. They can be derived from the most recent available billing data, and can be updated annually.6

The province is divided into regions, like the present regional hospital districts in B.C. Each region has, not only a different population, but a different age-sex distribution of population, such that each will have a different average per capita need for, or at least utilization of, services, associated with that age-sex mix.

---

5 A large number of categories make the allocation of funds quite sensitive to variations in utilization across age groups. Blocks such as 0-5, or 65 and over, hide a good deal of variation in utilization, as say, between 65 year olds and 90 year olds (Barer and Wong Fung, 1987). They can therefore yield unfair results across regions. On the other hand, too fine a grid may result in instabilities if the numbers in a particular age category are quite small.

6 The relative rates would presumably be computed for a period of time long enough to average out use patterns, and would be based on expenditures derived from a constant fee schedule.
Using the age-sex specific per capita utilization rates, however, we can define a weighted population for each region which will reflect these differences in needs. The numbers of people in each age-sex group in a particular region are simply multiplied by the corresponding relative rate of utilization of services, and then the products are totalled across age-sex groups. This yields a synthetic regional population which will be above or below the actual, according to whether the region has a population of above or below average users of services. The sum of synthetic regional populations aggregates to the actual provincial population.

The negotiated budgetary total for the province, $B$, is then split among the different regions in proportion to their synthetic populations, in equal amounts per weighted capita. This automatically gives relatively more to regions with elderly populations, and less to regions with younger populations, in exact proportion to their expected differences in medical services utilization. 7 If we refer to a representative region as region $j$, its budgetary allocation will be $B_j$.

The algebra is straightforward. The total provincial population, $P$, is subdivided into 42 age-sex groups. We can let $p_i$ be the population in a representative age-sex group $i$, where $i$ runs from 1 to 42. Let $E$ be total expenditure on physicians' services in the most recent available

---

7 These "expected" differences, however, are established by applying the provincial pattern of utilization by age-sex class to each region. There is no allowance for regional variations in this pattern.
period, and $e_i$ the expenditure on services for the $i$th age-sex group. Then the weight applied to persons in the $i$th group is $w_i = (e_i/p_i)/(E/P)$. If then $p_{ij}$ is the number of people in region $j$ who fall into the $i$th age-sex group, the synthetic population of region $j$ is $\Sigma w_i p_{ij} = q_j$. This may be greater or less than $\Sigma p_{ij} = p_j$, the actual population of region $j$, but the sum of the $q_j$ across all regions equals the sum of the $p_j$, i.e. total population $P$. Then the budgetary allocation for region $j$ is simply $(q_j/P)B = B_j$.

These $B_j$ of reimbursement are then available to pay the claims submitted by physicians for services rendered to the residents of region $j$. We can designate as $C_j$ the total of all claims submitted, on behalf of residents of region $j$, by all the physicians in B.C. (or out of it). Note that $C_j$ is not equal to the claims submitted by physicians resident in region $j$, although there will be considerable overlap, because physicians in region $j$ may also provide services to residents of regions $k$, $l$, etc., and at the same time residents of region $j$ may receive services from physicians in regions $k$, $l$, etc.

The total of claims submitted for the care of residents of region $j$ will not in general equal the budgetary allocation for region $j$. Their ratio will be the prorating factor for region $j$; that is, $R_j = B_j/C_j$. Fee claims submitted for services rendered to residents of region $j$ are reimbursed at $R_j$ times the negotiated rate in the fee schedule; they carry a premium or a discount according to whether $R_j$ is greater or less than one. Looked at from the practitioner's point of view, services provided
to residents of (relatively) "underserviced" areas will automatically be reimbursed more highly than those provided to residents of (relatively) "overserviced" areas - no matter where the service was provided or the practitioner works or resides.

Of course, a practitioner choosing to practice in a relatively underserviced area can expect that services provided to most of his/her patients will carry a premium. Conversely, if s/he locates in a relatively overserviced area, fee claims will be more likely to be discounted. But it is the rate of servicing received by the regional residents which is decisive, not the number of physicians per capita reported as available or the services which they provide.

It may be viewed as a bit peculiar, that a physician seeing different patients for the same problem may receive different fees, depending on the patient's residence. But proposals to pay differential fees according to where the physician is located also amount to paying different fees for the same service. This approach merely extends the concept of regional fee differentials to reflect the fact that what is at issue is not, fundamentally, where physicians locate, but whether or not patients are served.

Under this approach, the incentives for physicians to relocate in underserviced regions are automatic, and require no explicit adjustment of relative fees, or "policing" of where the physician is actually located. The physician may locate wherever s/he likes, but the provincial servicing
patterns will determine how much s/he is paid (per service). From the provincial government's perspective, the "oversupply" of physicians in a particular region is no longer a financial problem, because an increase in physician numbers does not add to overall expenditures on behalf of residents of that region. The argument over whether increased numbers of physicians lead to increased rates of servicing then becomes moot. If they do, the $R_j$ value in the region of physician increase simply falls.\footnote{The provincial government, being thus protected from the financial consequences of increases in physician supply, might no longer wish to limit the issuance of geographically unrestricted billing numbers (Barer, 1988). The practitioner community might also look at that issue differently. Both government and practitioners should still however be concerned, on medical grounds, if particular populations are grossly over- or under-served. But these concerns would no longer be entangled with financial issues.}

The location of the physician is not, however, a matter of indifference. In remote regions there may be a need to provide "stand-by" capacity for emergencies. It may be appropriate to reimburse physicians simply for being available, over and above the actual services they provide. But this should and can be provided separately from the fee schedule itself, through modification of existing "allowance" programs.

Secondly, people in "under-doctored" areas may not, in fact, receive fewer services than people in comparable areas - they are not "under-served" (Horne, 1987) - but they may have to go to more trouble and expense to reach care. The premium for services supplied to persons in "under-doctored" regions may not be sufficient to induce "enough" (by some external criterion) physicians to locate there - additional inducements
may be required. But this approach has the virtue of an automatic incentive - it can always be further adjusted.

Finally, reimbursement of emergency services may require special treatment in a province like B.C. where a number of people have residences in (well-serviced) urban areas but travel to remote (and often dangerous) job sites. It would obviously be inequitable for the physician working in the remote interior, who provides emergency care to a logger or fisherman, to have his/her fee discounted because the patient's address of record is Vancouver or Victoria.\footnote{The problem may be broader than emergency care, for persons who are away from home for long periods of time. Their normal source of care may not coincide with their official residence. It seems unlikely, however, that care for this generally healthy sub-population will generate significant volumes of billings. If it does turn out to be problematic, the appropriate response seems to be to adjust the definition of residence.}

There is also a possible source of friction in the referral relationships between physicians in smaller communities, and in urban and metropolitan centres. If a general practitioner in a small community in region \( j \) refers a patient to Vancouver, and that patient receives extensive and expensive physician care, the claims for that care become part of the claims against the reimbursement allocation for region \( j \) - as indeed they should. But the process may be interpreted as "out-of-town" specialists tapping the pool of funds available for reimbursement of local physicians - as indeed they are. The billing activities of the metropolitan specialists, \textit{ceteris paribus}, reduce the value of \( R_j \) for the
region $j$ from which the patient was referred.

**Prorating by Region and Class of Service**

The scheme can be modified to limit this degree of financial conflict of interest, or at least make it a controllable parameter, in the following way:

The total budgetary allocation for physicians' services for the province, $B$, can be divided up among different types of physicians' services. These might, at a minimum, be general practitioners' services, and services of medical, surgical, and diagnostic specialists. Then:

$$B = BG + BM + BS + BD$$

where $BG$ is the budget allocation for general practitioners' services, and so on. The distribution of $B$ among these components is part of the negotiation process, although historical patterns would probably exercise a dominant influence. At this point, again, there is an important policy decision. Does one define these allocations in terms of particular types of services, i.e. fee schedule items, or in terms of payments to particular types of physicians? Again, consistent with the philosophy that the basic purpose of the health insurance program is to pay for services, not to provide incomes for physicians, we shall assume that the components of $B$ are allocated by type of service, rather than by type of physician providing the service.
We defined \( e_i \), above, as the average per capita expenditure of persons in the \( i \)th age-sex group in the provincial population. We can now partition \( e_i \) into \( e_{g_i} \), \( e_{m_i} \), \( e_{s_i} \), and \( e_{d_i} \), the average per capita expenditures of people in the \( i \)th age-sex group on general practice services, medical specialty services, etc. As above, these are used to derive weights for the age-sex components of the population, which in turn are combined with regional population patterns to yield region-specific and type-of-service specific allocations of reimbursement, \( B_G \), \( B_M \), \( B_S \), and \( B_D \).

The claims submitted on behalf of residents of region \( j \) would then be similarly categorized into groups \( CG_j \), \( CM_j \), \( CS_j \), and \( CD_j \). The matching of total claims by category against total reimbursements would yield a set of ratios \( RG_j \), \( RM_j \), \( RS_j \), and \( RD_j \), which would be the premiums or discounts to be applied to fees for services of the corresponding types to residents of region \( j \).

This would have the effect of insulating the receipts of practitioners of one type from the level of activity of those of another, within or across regions, except insofar as practitioners cross specialty boundaries. The general practitioner in a remote region who becomes a de facto general surgeon or anesthetist, because those services are not readily available to the population s/he serves, will also find that s/he

---

10 The synthetic populations in each region will be different for each of the types of services considered, insofar as the age-sex patterns of utilization differ across types of services. But this poses no conceptual or computational difficulties.
receives a premium fee for such services, because relatively few such claims are being submitted on behalf of that regional population.

On the other hand, the medical sub-specialist in an area in which the supply of general practitioners is rising rapidly, and services per capita likewise, will not find his/her reimbursement rate lowered in consequence. The value of $RG_j$ will fall, as the volume of general practitioner billings rises, but not that of $RM_j$.

Nor will a concentration of specialists in urban and metropolitan centres necessarily lead to their fees being discounted, insofar as they provide referral services to patients from a number of regions in the province. The medical sub-specialist above, if s/he receives referrals of complex clinical problems from all over the province, is effectively drawing on the reimbursement allocations in every region. On the other hand, the specialist who does not have a true referral practice, but is instead competing with general practitioners in a relatively over-supplied metropolitan area, will find that services to most or all of his/her patients are reimbursed from a "pot" that is stretched quite thin, and the fees are correspondingly discounted.

To the extent that residents in metropolitan regions receive more services from specialists than do residents in the rest of the province, the specialists who provide those extra services will be reimbursed at lower rates. The specialist who moves to a community where such services were previously unavailable, however, or more likely just hard to get to,
should do rather well on both volume of work, and fees.

Creating Controllable Financial Interdependence Among Specialties

This "multiple pot" approach implies complete separation of reimbursement rates for different specialties, or at least different types of services. The global allocation of $B_j$ for all services provided to residents of region $j$, on the other hand, implies complete integration. Every physician's activity, of whatever type, lowers the potential reimbursement of every other. An intermediate approach, with a "controllable" parameter, links the different specialty "pots" or allocations for each region. Such linkage recognizes explicitly the "gate-keeper" role of the physician, and particularly the general practitioner, in shaping the pattern of medical care.

To illustrate this intermediate approach, suppose that in region $j$ the allocation for diagnostic services $BD_j$ is insufficient to cover all the claims for diagnostic services for the population of region $j$, $CD_j$. Instead of adjusting the value of $RD_j$, the ratio of reimbursements to claims, so as to equate total payments to total allocations, part of the difference is made up by drawing on the allocations for the other services. As an example, when claims exceed the budget allocation, let half of the difference be paid from the budget allocations for reimbursement of the other services.

This will still leave total claims greater than the amount allocated,
and so diagnostic services will still be reimbursed at a discount from the fee schedule. But this discount, \( R'D'_j \), will be only half as great as if no funds had been withdrawn from the allocations for other services: \( R'D'_j = (B'D_j + C'D_j)/2C'D_j \), as compared with \( R'D_j = B'D_j/C'D_j \).

The corresponding amount, \( (C'D_j - B'D_j)/2 \), is subtracted from the reimbursement allocations, \( B'G_j \), \( B'M_j \), and \( B'S_j \), available for other services. The distribution process could be simple, lowering each of the budgetary allocations by the same percentage, or sophisticated, trying to allocate responsibility for different diagnostic tests to different specialties.

The basic idea, however, remains simple. To the extent that the actual billings for diagnostic services over- or under-run the amounts budgeted for their reimbursement, some proportion (in this example 50%) of the discrepancy, positive or negative, is shifted to the reimbursement for other clinical services (and thereby for other practitioners who ordered the tests). If practitioners in region \( j \) make less use of tests (per capita, age-sex adjusted) than those in region \( k \), then (adjusted for population) \( C'D_j < C'D_k \). But \( B'D_j \) and \( B'D_k \) will not be affected. Ceteris paribus, then, \( R'G_j > R'G_k \), and the same for medical and surgical specialists. There is a direct monetary incentive for more restrained test ordering behaviour.

Of course there is an even stronger direct incentive when all the claims and reimbursements in a given region are pooled in one pot, since
all the over- and under-runs in each service area have a direct impact on the total $R_j$ for the region. But the relationship is much less apparent for the practitioners of a particular type. Furthermore, the "linked pot" approach permits the size of the sharing ratio to be varied to "fine-tune" the incentive.

Practitioners may feel that 50 percent represents too much of an incentive, and would lead to skimping on appropriate tests. Perhaps the simple existence of a link is sufficient to attract attention; perhaps 25 percent sharing would be sufficient. The relationship is obviously open to further discussion. Indeed, it can be made part of the negotiation process; and it may be that physicians in different regions would wish to choose different sharing ratios.

Financially, the provincial government should be indifferent to the ratios chosen, although like physicians themselves, the provincial government must be concerned if the financial incentives threaten to lead to inappropriate patterns of practice. In this respect, however, province and physicians have a shared objective, at least in principle. The capping of the overall outlays at $8B removes from consideration those financial questions over which payer and payee must inevitably disagree.

The practitioners in each community, however, will have a direct and identifiable financial stake in the patterns of practice in that community. Insofar as they economize on the use of other resources in providing care, their own gross receipts are increased. And conversely.
The principle can be extended. The over- or under-runs in the specialty service allocations can likewise be split in half, or in some other proportion, and part allocated back to the general practitioner pool. Then the total amount allocated over the claims for general practitioner services would be the amount originally allocated on behalf of residents of the region, \( BG_j \), plus one half of any under-runs, or less one half of any over-runs, in the specialty service pools, after adjusting for the allocation of over- or under-runs in the diagnostic service pool.

To be specific, the adjusted reimbursement ratios for the different service areas are computed as follows:

1. \[ RD'_j = \frac{BD_j - 0.5(BD_j - CD_j)}{CD_j} \]
2. \[ RM'_j = \frac{BM_j + m(BD_j - CD_j) + CM_j}{2CM_j} \]
3. \[ RS'_j = \frac{BS_j + s(BD_j - CD_j) + CM_j}{2CS_j} \]
4. \[ RG'_j = \frac{BG_j + g(BD_j - CD_j) + 0.5(BM_j - CM_j + m(BD_j - CD_j) + BS_j - CS_j + s(BD_j - CD_j))}{CG_j} \]

where as before, the \( RG'_j \), etc., are the ratios by which the fees for each service category are marked up or down; the \( BG_j \), etc., are the total allocations of reimbursement by service type and region before adjustment for over- and under-runs in other service categories; the \( CG_j \), etc. are total claims for reimbursement, by region and service type, valued at the fee schedule currently in effect before prorating; and the \( m, s, g \), are the proportions of the diagnostic services under- or over-run which are allocated to each of the other three service categories - medical
specialties, surgical specialties, and general practice, respectively.

Since in these equations under-runs and over-runs in each category are assumed to be divided in half, with half borne within the category and half spread to another category, it follows that:

\[(5) \quad m + s + g = 1/2\]

But this need not be so. One could define the sharing ratios differently. As one moves away from one half, the algebra gets a little more involved. But it would be a straightforward exercise to set up equations (1) to (5) in terms of general parameters for the sharing ratios.

To recapitulate, the allocation described in these equations first compares total billings for diagnostic services provided to residents of region j with the total amount budgeted for such services. Half of any over- (under-) run is then debited (credited) to the budgetary allocations for medical, surgical, and general practice services in the proportions m, s, and g (which may be set ad lib but must sum to one half). The reimbursement rate for diagnostic services provided to residents of region j is then marked down (up) so as to equate the total claims to the original budgetary allocation augmented (reduced) by the amounts charged (credited) to the other service categories. The effect of the transfer will be to cut in half the the size of the mark-up or mark-down on fees for diagnostic services.
The budgetary allocations for each of the classes of specialty services, reduced or augmented by their share of the over- or under-run in diagnostic services, are then compared with the total claims for these services in the region. Once again, half of any over- (under-) runs are debited (credited) to another allocation, in this case to the allocation for general practitioner services. The fees paid for medical or surgical specialty services for residents of region \( j \) are then marked down (up) by ratios \( RM'_{j} \) and \( RS'_{j} \), sufficient to equate total claims of each type to the budgetary allocations for each, reduced or augmented by amounts charged or credited to either diagnostic or general practitioner services.

The amounts transferred to or from diagnostic services, however, depend on the over- or under-runs in that budget, while the amounts transferred to or from general practitioner services are determined by the over- or under-runs in the specialty sectors themselves.

Finally, the allocation for general practitioners, reduced or augmented by the amounts transferred to or from each of the other three service budgets, is compared with the total claims for general practitioner services, to determine the value of \( RG'_{j} \), the amount by which fees for general practitioner services are to be marked up or down from the overall fee schedule in region \( j \).

The "residual status" of general practitioner services in the financial computation corresponds to a view of the general practitioner as the basic gate-keeper in the system. Billings for other types of services
depend ultimately on the referral decisions of the general practitioner. Accordingly some proportion of the financial implications of each of those decisions, for good or ill, comes back to the original gate-keeper.

But the general practitioner is not the only gate-keeper; s/he shares this role with the specialists who also order diagnostic tests and initiate and/or carry out diagnostic and therapeutic manoeuvres. The pattern of transfers of funds among the different service allocations is intended to reflect the performance of this role.

Accordingly the m, s, and g values which determine the impact of diagnostic services over- or under-runs on the different referring practitioners could be adjusted in a more detailed and test- (or procedure-) specific manner to reflect the extent to which the ordering of certain types of tests is linked with particular practice specialties. If sub-sets of tests can be identified which are clearly ordered almost exclusively by providers of medical specialty services, for example, over- or under-runs for those tests should be charged or credited back to the specific allocation for those services. This extra element of sophistication might become particularly important if, as discussed below, allocations for hospital care are also linked to the practitioner allocations.

One could further strengthen the financial incentives associated with this "gate-keeper" role, by changing slightly the way in which the short-falls or savings in diagnostic services are flowed through to
practitioners. The process is expressed in the following modifications to equations (1) to (4).

\[(1A) \quad RD'_j = (BD_j - 0.5(BD_j - CD_j))/CD_j \]
\[(2A) \quad RM'_j = (BM_j + CM_j + 2m(BD_j - CD_j))/2CM_j \]
\[(3A) \quad RS'_j = (BS_j + CM_j + 2s(BD_j - CD_j))/2CS_j \]
\[(4A) \quad RG'_j = (BG_j + g(BD_j - CD_j) + 0.5((BM_j - CM_j) + (BS_j - CS_j)))/CG_j \]

These equations embody a different treatment of the over- and under-runs in the diagnostic services budget. It is still the case that a share of these discrepancies (for the sake of example, one half) is transferred to the other service allocations, and divided among them in the proportions \(m\), \(s\), and \(g\). But amounts transferred from (to) the diagnostic services allocation are no longer treated equivalently with the initial allocations \(BM_j\) and \(BS_j\). They are not pooled in with those allocations for the purpose of computing the specialty service over- (under-) runs to be shared with the general practice allocation. This implies that such discrepancies are not flowed through to the general practice services budget, but have a dollar for dollar impact on the payments to medical and surgical specialists.

Looked at another way, any over- or underrun in the reimbursement for specialty services is shared half and half with providers of general practitioners' services (although of course the sharing ratio need not be chosen at that level). This is intended to reflect the role of the general practitioner in making referrals. But sums allocated to or from
the diagnostic services budget are not shared, they affect the payments to specialists in their entirety. Hence the consequences of test-ordering behaviour by specialists (as expressed in the ratios of the parameters m and s) fall directly on specialists.

Example: The Allocation Scheme in Hypothetical Operation

For concreteness, we provide a numerical example of how the reimbursement program would operate. (In this example we use the sharing pattern embodied in equations (1) to (4), without the variant just outlined above.) Consider a region within B.C. which has a population of 100,000 people. If the total allocation to reimburse physicians in B.C. through the Medical Services Plan were, say, $900 million, for a total population of three million, this would imply a per capita allocation of $300. Let us assume that the population of this region has the same age distribution as the provincial population, so the total amount available to reimburse services provided to this group of people is $30 million.

Let us suppose that, on the basis of last year's billing patterns and negotiations, 40 percent of this is allocated for reimbursement of general practitioners' services, 20 percent for diagnostic services, and 20 percent for each of medical and surgical specialists' services, or:

\[ BG_j = 12 \text{ million, } BM_j = 6 \text{ million, } BS_j = 6 \text{ million, and } BD_j = 6 \text{ million.} \]

Now suppose the claims for medical services provided to this
population - by physicians anywhere in B.C. - come in at $7 million for diagnostic services (CD_j), $5 million for surgical services (CS_j), $7 million for medical specialists' services (CM_j), and $11 million for general practitioners' services (CG_j) (defined in all cases by the type of service provided, not the specialty of the provider). In total, this region is right on its allocation, and under a simple prorating system, all claims for services to its residents would be paid at 100 cents on the dollar.

But under this proposal, the amounts actually paid will reflect the discrepancies between the budgeted amounts, and the claims made. On average, the total of claims made will be fully reimbursed, but some will carry premiums, and others discounts. The million dollar over-run in diagnostic services will be cut in half, and only half borne by the suppliers of diagnostic services. Their claims will be prorated by \((6.5/7) = .92857\), and reimbursed at 92.857 cents for each dollar - a discount of 7.1 percent in response to an overrun of total claims by 16.7 percent. The remaining $500,000 of over-run in diagnostic services will be allocated equally (for simplicity) among the other three practitioner groups, $166,667 being withdrawn from each.\(^{11}\)

\(^{11}\) This implicitly assigns providers of each service group equal responsibility for the over-run. One might argue for equiproportionate assignment, which given the budgetary amounts in the example would have led to a deduction of $250,000 from the GP services and $125,000 from each of the specialty groups. The numbers in the example would have been changed, with higher rates of reimbursement for each of the specialty services, and lower for the GP services.
The medical and surgical specialists' services would each be left with an allocation of $5,833,333, to cover claims of $7 million and $5 million respectively. But again, the discrepancies would be cut in half, and half of each transferred to the general practice allocation. The surgical specialties would have a surplus of $833,333, of which $416,667 would be credited to GP services. The remainder would support a premium reimbursement of $1.08333 reimbursement per dollar of claims. From the medical specialties, half of the shortfall of $1,166,667, or $583,333, would be debited to the allocation for general practice services, and the remaining shortfall would result in their claims being discounted to 91.667 cents on the dollar.

The budgetary allocation for general practice services would be reduced by $166,667 as their share of the diagnostic services over-run, then increased by $416,667 as half of the surgical services under-run, then reduced by $583,333 as a share of the medical services over-run, for a net adjustment (reduction) of $333,333. This would still, however, leave them with an excess of reimbursement allocation over claims; $11,666,667 compared with $11 million, so they would be reimbursed at $1.06061 per dollar of claims. (The extent of rounding applied in actual practice will depend on the precision of the computers.)
The outcome of the process is summarized as follows:

<table>
<thead>
<tr>
<th>Type of Service</th>
<th>Budgetary Allocation</th>
<th>Claims Received</th>
<th>Amount Paid</th>
<th>Prorating Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Practice</td>
<td>$12 million</td>
<td>$11 mn.</td>
<td>$11,666,667</td>
<td>1.06061</td>
</tr>
<tr>
<td>Medical Specialty</td>
<td>$6 million</td>
<td>$7 mn.</td>
<td>$6,416,667</td>
<td>.91667</td>
</tr>
<tr>
<td>Surgical Specialty</td>
<td>$6 million</td>
<td>$5 mn.</td>
<td>$5,416,667</td>
<td>1.08333</td>
</tr>
<tr>
<td>Diagnostic Procedure</td>
<td>$6 million</td>
<td>$7 mn.</td>
<td>$6,500,000</td>
<td>.92857</td>
</tr>
<tr>
<td></td>
<td>$30 million</td>
<td>$30 mn.</td>
<td>$30,000,001</td>
<td>1.0 (avg.)</td>
</tr>
</tbody>
</table>

The full amount of the negotiated budget is paid out, and types of services which exceed (fall short of) their initial allocation are provided with additional resources. But they are not fully reimbursed for their over-runs, or "penalized" for their under-runs, as would happen in an across-the-board prorating system - or for that matter, in this case, by a simple reimbursement of all claims. Restraint is rewarded, and over-runs are discounted, whether it be in billings for one's own services, or in referrals to other specialists or for diagnostic services. And of course the flexibility exists for regional groups of practitioners to negotiate higher or lower sharing ratios.
Extensions and Analogies

Nor need the process stop at this point. Since the categories G, M, S, and D are defined as types of services, not as types of practitioners, there is no reason why particular sub-sets of services, such as obstetrical services, should not be isolated for different treatment or given their own "pots". The finer the differentiation of specific services, however, the more important it will become to adjust capitation rates across regions to take account of characteristics other than age and sex which may affect needs for and utilization of medical services. Furthermore, the interdependencies among different classes of services may make the appropriate sharing formulae for over- and under-runs more complex. But there is no reason why the system cannot be fine-tuned from year to year while it is in operation.

A substantial extension which is quite within the general spirit of the approach, would be to include expenditures on hospital services as an additional category. In terms of the algebra above, one would introduce budgetary allocations and claims $BH and $CH, and again link over- and under-runs to the amounts allocated for reimbursement of different types of physicians' services. This linkage would create incentives similar to those in capitated plans in the United States, in which changes in patterns of hospital use have direct consequences for the reimbursement of physicians.

The process would be an order of magnitude more complex for hospital
services in Canada, however, because present accounting and budgeting systems do not permit the attribution of expenditures to the care of particular patients. Thus one can readily establish whether utilization rates for a particular regional population, measured in age-sex adjusted days or separations per capita, are above or below the averages for the province as a whole. But one can only very crudely translate these into corresponding variations in per capita expenditures. Per diem costs by hospital are notoriously inadequate for this purpose.

Nevertheless, one could go part way by comparing regional age-sex adjusted hospital utilization rates with the provincial averages. Such relative utilization rates could then be linked to the budgetary allocations for physicians' services, augmenting these allocations for residents of regions with low hospital utilization rates and conversely. In fact the Ontario Ministry of Health has done something similar for years in reimbursing Health Service Organizations, providing them with an additional budgetary allocation based on estimates of their "saved" hospital patient days.

The translation of hospital utilization patterns into regional allocations for physician reimbursement would open up some very contentious areas for negotiation, over both the relative costs of different institutions, and the regional patterns of "need" insofar as they may not be reflected in age-sex distributions. Nor is such linkage necessary, if the objective is simply to reconcile fee-for-service reimbursement of physicians with predictability and capping of overall
outlays on physicians' services.

But if one looks forward to a broader objective, that of providing physicians with opportunities to benefit from finding ways to manage the care of patients more efficiently, rather than simply more intensively, as at present, then at some point the integration between the physician and the hospital must be addressed explicitly. It is, after all, a standard finding that the principal economies in the American capitation-based Health Maintenance Organizations have in the past come from more careful use of hospitals and hospital-based services (Luft, 1981; Manning et al., 1984). To date, Canadian physicians have had no way to participate, at least not directly, in any savings which their behaviour might generate in the hospital sector.

It is this latter objective, of generating incentives for providers to adopt more cost-effective styles of care and for patients to seek out such providers, which motivates the current policy of promoting capitated care in the United States. Under the U.S. Medicare system for the elderly population, Health Maintenance Organizations (HMOs) which undertake to provide (or contract for) "all necessary care" for a defined group of people are reimbursed at 95 percent of the Adjusted Average Per Capita Cost (AAPCC), an estimate of what the cost of providing physician and hospital services to that group would have been, in that local area, under fee-for-service reimbursement (Anderson et al., 1986).

The AAPCC is adjusted for the age, sex, welfare and institutional
status of the covered group, as well as for historical patterns of cost difference in the local county area. (The system is applied across the entire U.S.) A principal difference between the U.S. system and the proposal advanced here, however, is that patients self-select into and out of HMOs. It is observed that past utilization patterns are a powerful predictor of future costs, for individual patients, resulting in strong incentives for HMOs to seek out previous low users. Further, it appears that more healthy patients tend to choose HMOs. In consequence, setting capitation rates at 95 percent of regional AAPCC may actually increase overall costs, and in a dynamically unstable way (Ellis and McGuire, 1987; Muldoon and Stoddart, 1987). This has led to increasing interest in adjustment of the AAPCC for the past utilization pattern of the enrolled group.

But the counter-argument is equally clear. Any "capitation" system which adjusts the reimbursement rate to present or past utilization rates, generates incentives for increased servicing which undermine the whole point of capitation as a method of encouraging more conservative styles of practice. The dilemma seems unavoidable in capitated systems with self-selected membership. An alternative would adjust the AAPCC, but only for past "non-discretionary" utilization (Anderson et al., 1986), if this could be defined unambiguously.

A more radical suggestion is a geographically-based capitation system - so-called "carrier at risk" - in which insurance carriers would bid on geographically defined populations. The successful bidder would undertake
to pay all the (covered) expenses for that population, in return for a
global payment from the Medicare administration, and would then undertake
to manage the care process within that region (U.S. Congress, Office of
Technology Assessment, 1986; U.S. Congress, Congressional Budget Office,
1986).

Under our proposal, the medical community of the province becomes the
"carrier at risk". But for the province as a whole, that risk is not
financial, because the total allocation $B$, predetermined by negotiation,
is the payment both to the "carrier" and to the providers. Its
distribution among regions is also determined; the only financial risk
relates to the distribution of those totals among different practitioners.
Unlike the U.S. proposals, however, the local groups of practitioners at
risk also have predominant control over styles of practice, and a
professional responsibility for quality of care.

The U.S. geographic carriers, as purely financial agencies, would be
involved in a cost-minimizing struggle with the practitioners in their
regions. They would find themselves in some form of continuous
negotiation with providers - rather like small scale provincial
governments, but without the bargaining leverage which comes with covering
the whole population (and controlling the legislature!).

The principal "risk" in the present proposal, apart from possible
perverse patterns of intra-profession behaviour, arises if the regional
allocations of reimbursement do not adequately reflect the "needs" of the
populations covered. If some regions have populations which are "sicker" than others, the practitioners who serve those populations will be inadequately reimbursed relative to their colleagues in healthier areas. They will have to work harder, for less. Should the regional allocations therefore respond to factors other than age and sex?

This is essentially the problem addressed by the Resource Allocation Working Party (RAWP) process of regional allocation in the U.K. National Health Service (Carr-Hill, 1987, n.d.[1988]), which likewise attempts to define region-specific allocations of resources on the basis of the "needs" of residents of each region. Unlike the U.S. discussion of adjustments to the AAPCC, the RAWP allocations have in general deliberately avoided linking resources to historical patterns of utilization and expenditure, for reasons which Carr-Hill spells out clearly.

At the regional level, and without the problems of patient and carrier selection which bedevil the U.S. approach to capitation, historic patterns of use and cost reflect the location and activities of providers rather than the needs of patients - highly unequal patterns which were exactly what the RAWP program was intended to correct. From this point of view - which is implicitly accepted in proposals like that of Anderson et al. (1986) to relate the AAPCC only to past "non-discretionary" utilization - even the current adjustment of the American AAPCC for relative per-beneficiary cost patterns in the local county is
On the other hand, if the facilities and personnel are centralized - in London or Vancouver - and the resources are attached to the individual persons wherever they may be, it is essential that the reimbursement system provide for transfer of resources from one region to another whenever people cross regional boundaries to obtain services. Under the proposal in this paper, a transfer occurs automatically. In the U.K., proposals to create "internal markets" within the National Health Service address this sort of implicit or explicit inter-regional contracting for services (Enthoven, 1985; Institute of Health Services Management, 1988). But in the absence of a fee-for-service system, the structuring of such "internal markets" is a more complex task.

The RAMP allocation formulae in the U.K. allocate resources by region on the basis of age, sex, and socio-economic status, recognizing the generally accepted fact that poorer people have greater health needs. (The American AAPCC does likewise, in that it includes welfare status as an adjustment factor.) Unfortunately that general agreement does not extend either to how socio-economic status should be measured, for the purposes of this allocation, or, equally important, how resource

12 It is a fortiori inappropriate to adjust for the institutional status of the covered population, since on a geographic basis institutionalization rates are so closely associated with the availability of facilities. But to fail to do so in a selected population would be to invite disaster, because of the very large differentials in expected health care costs between institutionalized persons and those living in the community.
distributions should be adjusted in response to whatever differences are measured (Carr-Hill, n.d.[1988]). In the circumstances, it seems prudent to leave this as an area for future research.\(^{13}\)

Indeed, a significant positive feature of this proposal may be precisely that it could create a demand for better information on the health status and needs of people in different regions, as part of the process of regional competition for funds. It has frequently been noted that Canadians spend a great deal on health care, but devote even less attention than most other countries either to measuring its impact, or to studying the underlying needs of the population. Under this reimbursement system, a region (or its practitioners) who felt that they were being under-reimbursed, would have a strong incentive to try to develop reliable data on population health status which could be inserted into the financial formula.\(^{14}\)

Some Practical Considerations

One need not, however, solve all the problems at once. The proposal outlined here is relatively straight-forward conceptually. Its principal

\(^{13}\) It would not be difficult, however, to calculate differences in health care utilization and costs for (status) native populations, and adjust regional allocations for these. If there were significant regional variations in identified welfare populations, this too could be included in the formula; but one would not expect this factor to be very important in a geographically pooled population.

\(^{14}\) Carr-Hill's (1987) very useful review of the need for and problems with SES adjustments is a response to just such a request from a region in the NHS.
disadvantages, at least in the short run, seem to be the requirement for substantially more data on the populations served by provincial medical insurance systems, and the potential difficulty of communicating its essential features to the practitioner and the general population. The shift in reimbursement focus from practitioners to beneficiaries requires a complete and up to date computer file of beneficiary data, including age, sex, address, and any other information to be used in determining population weightings for regional reimbursement allocation. This is not now available in all provinces, but is spreading.

Understanding and acceptance by the practitioner community may be more problematic, but will depend on their recognition that regional and specialty prorating actually provides more incentives and opportunities for practitioners collectively to manage the pattern of medical care provision. By separating the determination of total provincial outlays from the negotiation of fee schedule structure, this system reduces the range of issues in conflict between reimburer and practitioners, and expands the areas of possible common interest and co-operation.

The main question is whether the simultaneous maintenance of fee-for-service reimbursement for physicians, and global capping of expenditures, is in fact a plausible and acceptable objective. If it is, and

---

15 A more complex system has been in operation for over a decade in the Federal Republic of Germany, involving independent prorating by each of the numerous sickness funds (Brenner, 1988). The prorating mechanism applies only to ambulatory care; physicians' services to hospital inpatients, and drugs, are not included. In this system, physicians earn different fees for the same service according to the sickness fund a patient happens to be enrolled with, which in turn depends on occupation,
particularly if we are in the process of moving to it in a back-handed manner anyway, the allocation process outlined above seems to have certain desirable features.

i) From the provincial government's point of view, it has the major advantage of making outlays predictable in advance, and eliminating the combination of utilization slippage, and contentious bargaining over "give-backs" which now occurs periodically in B.C.

ii) The regional, resident-based system mimics the behaviour of a "market" in that relative fees automatically rise (fall) in relatively under- (over-) serviced areas. Practitioners can choose their locations freely, but accept the consequences for their fees and incomes.

iii) Allocation of reimbursement by type of service provides a similar automatic fee adjustment, encouraging the more equal distribution of specialty services.

iv) Of particular importance, groups of physicians in each region have a strong incentive to examine and manage their own patterns of practice. Physicians serving patients from regions with relatively high rates of referral for diagnostic work or specialty consultation employer, or area of residence. Funds with "sicker" enrollee populations, however, must collect a higher proportion of the payrolls of their covered population, and/or discount the fees paid on their behalf. The equity of such an approach is obscure.
(or hospital use) may wish to consider whether these patterns are appropriate; in this consideration the provincial government will be able to co-operate rather than acting as an adversary.

v) Last, but by no means least, the linkage of regional financial allocations to (estimates of) beneficiary needs, rather than practitioner activity levels, may finally stimulate some serious interest in the differences in regional patterns of medical need, and their correlates within and outside the health care system.

Until now the Canadian health insurance plans, while comparatively successful in world terms, have tended to focus the attention of both providers and reimbursers almost exclusively on financial issues, and almost always in conflict. The system of reimbursement outlined here might not only move several important policy issues out of the adversarial arena, but also create "clients" for better information on both the effectiveness of medical interventions, and the needs of the populations served. There would also be an obvious channel for the application of that information "at the coalface" - in the determination of actual budgetary allocations and patterns of medical care utilization.
APPENDIX

Some Issues of Implementation
SOME ISSUES OF IMPLEMENTATION

The process of physician reimbursement proposed in this paper - a combination of capitation and prorating on a regional and service-specific basis - requires information on the region of residence of the enrolled population. At present this information is not available in all provinces. In B.C., for example, many people are enrolled in the Medical Services Plan through employee groups, and the plan records contain only the address of the employer.

Other provinces do maintain address information for each enrollee. For that matter, address information is also kept current for holders of drivers' licences. Thus the information requirements of this reimbursement system are by no means infeasible or unreasonable. But its full application is clearly restricted to jurisdictions in which such computerized address information is part of the operational data base.

A somewhat less sophisticated form of the proposal, however, can be put into operation, pending the availability of enrollee address data. This more limited system has many of the same features as the full scheme, but the incentives which it embodies are not as precisely targeted. It does, however, preserve the main characteristics of combining fee-for-service with overall budgetary control, and of providing physicians with some incentives for more conservative practice styles.

As in the more detailed system, the reimbursing agency and the
physicians' association bargain over a total allocation of funds for reimbursement, $B$, and its allocation into sub-components $BG$, $BM$, $BS$, and $BD$, which are budgetary allocations for general practice services, medical specialty services, surgical specialty services, and diagnostic services respectively. As in the original proposal, these allocations are to reimburse types of services, not types of practitioners.

The total $BG$ is then broken up into regional allocations $BG_1$, ..., $BG_j$, ..., $BG_n$, as before, over the $n$ regions of the province, on the basis of the age-sex adjusted population of each region. The population adjustment factors are based on the relative rates of utilization of general practice services, by age and sex class, province-wide. For specialty services, however, no sub-provincial allocation of funds is made. For purposes of reimbursing such services, the province is treated as one large service area. Thus total claims for reimbursement, $CM$, $CS$, $CD$, are aggregated at the total province level. A single prorating factor will apply to all specialty services of a given type, provided anywhere in the province.

These prorating ratios can be adjusted to reflect the interconnectedness of medical practice, just as in the more detailed proposal above. Over- or under-runs in the diagnostic services allocation can be met in part (or, if desired, whole) by transfers from or to the allocations for the other services, and similarly specialty over- (under-) runs can be reflected in the prorating factor for general practice services. The difference is that net allocations for general practice
services in each region are affected by over- or under-runs in specialty services for the whole province, rather than responding only to levels of services provided to the residents of that region.

For general practice services, the (adjusted) regional allocations $B_{G,j}$ will reimburse all such services provided by practitioners in region $j$. Thus total claims for general practice services, $C_{G,j}'$, will be the claims submitted by the practitioners in the area, regardless of the residence of the patient. In contrast, $C_{G,j}$ in the original proposal was total claims for such services submitted on behalf of the residents of region $j$, by practitioners anywhere in the province.

The fees reimbursed to practitioners in each region $j$ will be scaled up or down according to whether this value $C_{G,j}'$ falls short of or exceeds the original budgetary allocation for the region, $B_{G,j}$, as augmented or reduced by transfers from or to the other specialty services. But the adjustment to these prorating factors to take account of these transfers will have to be in the same proportion in each region. This follows from the fact that, in the absence of patient residence data, one cannot attribute over- or underruns in specialty service use to any particular group of referring practitioners.

Any services falling into the specialty categories, provided by practitioners in region $j$, will be scaled up or down for reimbursement purposes by the ratios defined at the provincial level.
This process of reimbursement does not depend on any information about the residence of particular patients; it requires only estimates of the total population in each of the regions, and its age-sex distribution. Implicit in the process is the assumption that general practice services are provided within relatively compact service areas, and that patients do not travel long distances, or cross regional boundaries, for such services. (Or rather, such effects are assumed to be small.)

To the extent that this assumption does not hold, however, regions which are "net exporters" of general practice (their practitioners provide services to patients from other regions) are placed at a disadvantage relative to "net importers". The funds allocated to reimburse practitioners in each region are based only on the persons resident in that region, and are not adjusted for border-crossing.

The inability to adjust for such border-crossing, however, is simply a consequence of the lack of data on patient residence - in the absence of such data one cannot even reliably identify "exporting" and "importing" regions. (Note that it is the status of the region which matters - the individual practitioner who sees a patient from another region is not thereby disadvantaged so long as there is some corresponding patient from his/her region travelling to a physician in another region. The prorating factor which applies in region j depends on the net flows.)

To the extent that such inter-regional flows are perceived as a problem, the obvious answer is to speed up the generation of patient
residence data, and the shift to the process of reimbursement in the main proposal. That fully residence-based system automatically adjusts for any border crossing which may occur. The justification for moving to an interim, practitioner-based system is simply that in the case of general practice services, border-crossing by patients is unlikely to be a major problem.

The interim proposal does, however, preserve the financial incentive for general practitioners to locate in relatively under-serviced areas. The total amount of reimbursement allocated to region \( j \) is dependent on the population size and structure in that region, independently of the number of practitioners in it, so prorating factors will tend to be above unity in regions with relatively few providers of such services, and conversely.

As emphasized in the main proposal, however, a patient-based system creates an incentive to provide services to residents of under-serviced areas. This interim system creates an incentive to locate in underserved areas. For general practice services, the distinction may not be very significant. But it could yield some perverse results, if practitioners begin moving their office addresses of record. Could practitioners spend one or two days a week in a shared office in northern B.C., report that as an office address, and then come back to Vancouver to work for the rest of the time? Under this scheme, such a manoeuvre might enable one to have the northern B.C. prorating factor applied to all of ones' Vancouver billings.
The answer is probably yes; if a scheme has a loophole, someone will take advantage of it. But the problem is unlikely to be large, or to develop overnight. If the system of reimbursement based on practitioner locations is clearly identified as an intermediate stage, pending the development of a residence data base for patients, any such distortions will be corrected automatically by the shift to a patient-based system.

With respect to specialty services, the interim proposal has the weakness that it treats the whole province as one service area. Thus it provides no incentive for specialists to locate in one area than in another - if regional concentration of specialists is seen as a problem, this proposal will not address it.

Furthermore, the incentive effects of the main proposal are diluted. Over- and under-runs of allocations for diagnostic and specialty services can be identified only at the level of the entire province, so the linkage between "pots" of funds is also at the provincial level. If practitioners in one region become more conservative in their diagnostic testing behaviour, for example, this tends to raise the prorating factor across the whole province, but obviously by a relatively small amount.

On the other hand, it is possible that the mere creation of this explicit financial linkage will have the effect of drawing physicians' attention to the questions of practice style and utilization patterns. In any case, with a shift to the residence-based system of reimbursement, the incentives would become regionally targetted within a few years.
To increase the incentive effects of the proposal, it would be possible (and indeed probably essential, for political acceptability) to report to physicians individually the components of their gross receipts, on an annual basis. Such a report should show their gross billings by service category, the adjustments for transfers among different service allocations, the prorating factors applied, and the adjusted gross receipts. This would enable each physician to see at a glance the way in which billings for different classes of services interacted with each other, and the financial consequences of different practice patterns.

Such an annual report implies an annual payment period, with budgetary allocations and prorating factors set over that time. It would be equally possible in principle to establish the scheme over a quarter, or even a month, depending on the administrative processes involved. One should, however, distinguish between the contract or budgetary period, and the period(s) over which rates of reimbursement could be adjusted. Several of the latter could be nested in the former.

For example, suppose the negotiation period was a year, and the various budgetary allocations were determined over that period. Clearly one cannot know what all the relevant inter-service transfers and regional prorating factors are until all the claims are in and settled for that year. One could simply pay claims as they came in, but then there will be

---

16 A combination of explanation in text, algebra, and examples might also help - for a program of some complexity, the more explanation the better.
a closing adjustment in which additional payments will have to be made to
some practitioners, and amounts recouped from others. The latter is, of
course, problematic. The delay for end-of-period settlement will of
course be greatly reduced as the province moves to on-line bulk billing,
but it cannot be eliminated entirely.

One possibility is simply to pay all claims as they come in at, say,
80 percent of the negotiated fee schedule, and then make end of year
supplementary payments (with interest) as appropriate. This assumes that
no prorating factors will fall as low as 80 percent, which seems
plausible. Alternatively, one could pay claims at 100 percent, or close
to it, and then recoup or augment funds as necessary by a discount or
supplement to next period’s claims (starting at the beginning of the next
payment period). This would be closer to the after-the-fact adjustments
which now take place, but the difference is that the amounts recouped or
paid would have to calculated separately for each practitioner. There
would obviously be some problems if practitioners left the province, or
retired, problems which do not arise in the 80 percent approach.

With the increased computerization of the payment process, however,
it may be possible now or soon to monitor outflows on a monthly or bi-
weekly basis. This would permit reimbursement in each budgetary period to
begin at 100 percent of negotiated fees. Then, within the period,
observed outflows could be compared with budgeted targets, by region and
specialty. Since there is a distinct seasonal pattern to medical service
billings, and even (within short periods) a response to the pattern of
statutory holidays, these targets would have to be set to reflect such normal patterns.

Adjustments could be made to the proportion of claims being reimbursed, on a monthly or bi-weekly basis, according to whether these targets were being over- or under-run. This would not give an advantage to practitioners who billed early (or late) in the year, because there would still be an end-of-period settlement to ensure that each practitioner's reimbursement corresponded to the formulas above. But the size of this adjustment would be significantly reduced, insofar as adjustments were taking place all along the way.

Practitioners would not be required to accept either discounts on their fees from the beginning of the budgetary period (which assumes overruns until proven otherwise) or placed at risk of being required to pay back substantial sums (also rather unpopular). Some subsequent adjustments are inevitable - and require interest compensation - but they can be minimized by continuous monitoring and adjustment. On the other hand the payment process becomes much more complex administratively than the simple approach of paying at 80 percent or 90 percent and then paying a one-time end-of-year adjustment. The latter approach may also have the non-trivial advantage of making it easier for the individual practitioner to understand how the system is working, and how his/her reimbursement is being determined.
References


Barer, M.L. and P. Wong Fung (1987), Fee Practice Medical Service Expenditures Per Capita, and Full-Time-Equivalent Physicians in British Columbia, 1983-86, HMRU 87:1, Vancouver: Division of Health Services Research and Development, University of British Columbia


