Two-tier issues in low-income country health systems

Barbara McPake, Institute for International Health and Development, Queen Margaret University Edinburgh
Two-tier issues in low-income country health systems

Barbara McPake, Institute for International Health and Development, Queen Margaret University Edinburgh

Short biography:

Barbara McPake, BA, PhD is a health economist specialising in health policy and health systems research. She has more than 20 years experience in these areas based in three UK university departments. She is currently Director, Institute for International Development. Until 2006 she was Programme Director, Health Systems Development Knowledge Programme of the UK Department for International Development. She has extensive research degree supervision and other postgraduate teaching experience, wide experience in Eastern and Southern Africa and some experience of other developing regions and the UK National Health Service.

Correspondence address: Institute for International Health and Development, Queen Margaret University, Queen Margaret University Drive, Musselburgh, EH21 6UU

E-mail: bmcpake@qmu.ac.uk

Tel: +44 131 474 0000

Key messages:

Two-tier situations in health systems are commonplace in low income countries. They are defined as situations in which a single provider offers two services or price discriminates in selling one service.

Analysis of these situations and their outcomes for access, quality of care and equity for health services has under emphasised the importance of the supply side incentives created.

A model (previously published) designed to better understand the implications of the offer of more than one quality level in public hospitals suggests that low quality users may not benefit, especially where there are strong inter-relationships between the demands for the two services.

The model raises concerns for the implications of exemption policy, informal charging, insurance reform and pharmaceutical pricing that suggest a number of important areas of further research.

Word count: 6,053 (4,566 without front material, abstract and references)
Abstract

Two-tier situations, defined as those in which a single provider offers two services or price discriminates in selling one service are commonplace in low income countries' health systems. Examples include the provision of public and private wards in public hospitals, exemptions, and sliding scale charging systems and health facilities that negotiate separately with multiple third party payers.

Analysis of these situations and their outcomes for access, quality of care and equity for health services has under emphasised the importance of the supply side incentives created. These are complex because demands for the services involved are likely to be inter-related with respect to price and quality and a provider cannot alter price or quality levels of one service without considering the implications for the other.

A model (previously published) designed to better understand the implications of the offer of more than one quality level in public hospitals suggests that low quality users may not benefit, especially where there are strong inter-relationships between the demands for the two services. Only subsidy levels responding to the utilisation of the low quality service can protect low quality service users, to some extent.

The model raises concerns for the implications of exemption policy, informal charging, insurance reform and pharmaceutical pricing. The incentives for providers to reduce quality of services provided to exempted patients have generally not been considered. The role of market structure in ensuring that all demands are catered for in an environment of informal pricing has not been studied. Insurance reforms have taken little account of product differentiation incentives inherent in models designed to produce universal access. Strategies that determine international differences in pharmaceutical prices may be undermined by changes in the economic conditions in high income countries. In each area a number of important areas of further research are suggested.
1. Introduction

Two-tier provision exists where a single provider offers two services or price discriminates in selling one service. Although the term ‘two-tier’ will be used throughout the paper, in practice, there may be many tiers, but at least two. The phenomenon can arise wholly within the private sector, or can result from innovations in public provision that involve the marketing of public services.

An example on which much of the thinking that has informed this paper has been developed, is the phenomenon of ‘private’, ‘high-cost’, or ‘VIP’ wards and clinics in public hospitals. This is a case of a single provider (the public hospital) offering two services (a basic service and a superior service). Previous work (McPake et al., 2007) has explored the incentives inherent in this case. The purpose of this paper is to summarise the conclusions of that work for a policy audience unlikely to engage with the formal presentation in that paper, and to explore the further implications of those conclusions for a wider range of situations that are analytically similar.

Superficially at least, there are analytical similarities with the cases of exemptions (price discrimination for the same service, in principle), the offering of multiple health insurance packages by a single insurance agency (differentiated services at differentiated prices), hospital provision of services to clients covered by multiple insurers and under multiple policies (potentially differentiated services reflecting differentiated negotiation outcomes for each client group), the pricing behaviour of pharmaceutical companies in response to different procurers and procurement strategies (price discrimination) and informal charging situations (differentiated price by user; potentially differentiated services).

2. Incentives in two-tier provision

“It is not because of cost that some company or other has open third-class carriages with wooden benches … what the company is trying to do is prevent the passengers who can pay the second-class fare from travelling third-class: it hits the poor not because it wants to hurt them, but to frighten the rich.”

Dupuit, 19th century (cited in Phelps, date)

The basic insight underpinning this quotation is that demands for the two services or separately priced services may be inter-related. Dupuit recognised the potential for cross-quality elasticity: the demand for one service may respond to the quality level of the other. It suggests that providers may use this insight strategically, varying quality levels to maximise profit or revenue, for example and that one strategic option is to constrain basic service quality to encourage demand for a superior service. In principle, cross-price elasticity is also applicable: the demand for one service may respond to the price level of the other.

In our model published in 2007, we constructed a model of the incentives of the situation of private wards in public hospitals in two stages. Utilisation levels of each service, basic and superior, were modelled as dependent on the prices and qualities of both services. The hospital’s ‘reaction function’ was defined as the set of quality levels it will offer given the price levels for the two services in order to maximise its objective (the objectives of profit and revenue maximisation were compared). An authority, the Ministry of Health, set prices in knowledge of the hospital’s reaction function in order to maximise its objective (the cases
below assume the objective is to maximise total utilisation: this assumption is varied in the
original paper) and provided a fixed sum subsidy to the hospital. The situations in which the
Ministry provided a flat rate subsidy and a subsidy based on the level of basic service
provision were also compared. This set up was loosely based on arrangements in Zambia in
the late 1990s, the situation which had first inspired the enquiry.

The model generates predictions about the following variables:

1. The crude quality ratio (the quality level of the premium service relative to the quality
level of the basic service. Quality is modelled as the level of resources allocated to
the service.
2. The relative price (premium service relative to basic service)
3. The ‘value for money index’ which is the relative price per unit of quality of the
premium service compared to the basic service
4. The basic service quality level.

It therefore enables insight into the factors influencing equity (variables 1 to 3) and quality
(variable 4).

3. Model results

One approach used to understand the results of the model was to simulate the effects of
changes in key variables on the outcomes of interest. These simulations used specific
values for the variables in the model, which were loosely based on estimates of these for
Zambia, sometimes in the absence of data. They showed that as the demand functions for
the two services became more strongly inter-related through quality effects (as the cross-
quality elasticity increased), the quality level of the basic service fell. The crude quality ratio,
relative price and value for money indices all also fell, indicating that price and quality levels
tended to converge, while ‘value for money’ increasingly favoured premium service users
who paid a lower price per unit of quality throughout the simulation.

Increasing the cross-price elasticity of demand had effects on the value for money index and
on basic service quality almost identical to those for cross-quality elasticity. However, the
directions of the effect on the relative price and crude quality ratio were opposed. Rather
than converging qualities and prices, the two variables increasingly diverged and the two
services therefore became increasingly differentiated.

To some degree, these results confirmed Dupuit’s insight: the more strongly inter-related
were demands for the two services through cross price and quality effects, the more the
quality of the basic service was negatively affected – hence mechanisms to protect basic
service quality would appear a priority under two-tier provision.

The model also indicated that the manipulation of prices may be a weak mechanism to
achieve utilisation, equity and quality goals. Under the specific conditions of the simulation,
no level of demand inter-dependence produced an equitable result as measured by the
value for money index which measured the relative price paid in the basic and superior
service per unit of quality (or unit of resource allocated to the service).

In a further simulation, we considered whether increasing the extent to which the Ministry
weighted basic service utilisation more highly than superior service utilisation in making its
price setting decision could alter the outcome. By weighting basic service utilisation more highly, variables were influenced in the expected directions – but even as the weight on basic service provision approached 1, and the weight on superior service provision approached 0, those using the basic service paid more per unit of quality (or unit of resource allocated to them) than those using the premium service (the value for money index did not reach 1).

Finally, we considered whether transforming the flat rate subsidy to a per unit subsidy (while constraining its total to the same level), could improve outcomes. This strategy proved the most effective at securing the improved relative position of basic service users. Resulting in higher basic service quality, lower relative price, hence lower price per unit of quality and in the sole case of all the simulations, a value for money index indicating a lower price paid per unit of service quality for basic service users than for superior service users. The effect on the premium service was to render it more exclusive: higher price, higher quality and lower utilisation. The intuitive explanation is that paying a per unit subsidy for the basic service increased the effective price received per basic user by the hospital. To secure privileges, the superior service users have to outbid this higher price and this is only feasible given higher quality levels.

Policy implications of the model for two-tier public hospital provision

For the case of a private ward or clinic in a public hospital, operating under conditions similar to those modelled, several implications are clear. First, the creation of a second, superior tier of provision does not automatically generate a cross subsidy for the basic service, as often seems to be assumed. Whether or not a cross subsidy that favours the basic service user will be feasible depends on the demand conditions: the levels of own and cross price and quality elasticities, or the responsiveness of the demand for the two services to the price and quality conditions in their own and the alternative service and the cost of increasing demand by investing in quality improvement. The extent to which a feasible cross-subsidy will be achieved depends on supporting policy variables – whether price is set consistent to such a cross-subsidy and whether the payment mechanism supports this outcome.

It has sometimes been suggested that the challenge under two-tier provision is to ensure that the full cost of superior service provision is covered by its price, so ensuring that the public subsidy is dedicated to the group assumed to be poorer, who choose basic services. However, under the assumptions of this model, increasing the price of the superior service increases the hospital’s incentives to allocate resources to that service – a price increase for the superior service may not improve the situation and may even worsen it.

The relevance of the model to a wider range of issues in low-income country health systems

The primary purpose of this paper, however, is to consider the implications of the model for a wider set of situations which on the surface at least have two-tier characteristics.

Low-income country\(^1\) health systems are almost all characterised by inadequate public health spending in relation to what is required for universal coverage with even a basic

---

1 ‘Low-income countries’ are defined here in line with the definition used by World Bank statistical annexes and distinct from ‘middle-income countries’ in which universal coverage for basic services may be largely achieved and financed and insurance provision may be widespread.
health package. In consequence, a public-private mix arises in which a private sector fills gaps, largely on the basis of willingness-to-pay via an out-of-pocket payment system. The out-of-pocket private sector is not the sole preserve of the richer members of the society but extends to almost all economic levels of the population, offering price and quality combinations reflecting differentiated levels of ability to pay (Standing and Bloom, 2002; Mackintosh and Tibandebage, 2007; Mbaga Kida, 2009).

This scenario contrasts to a middle and high-income country scenario where it is realistic to seek universal coverage of essential services through a wholly public sector strategy. In the absence of a significantly higher public health budget than any low-income country achieves, even with donor support providing the largest share of the expenditure, universal coverage with even basic services requires private finance which can either be channelled through public or private providers (Rannan-Eliya and Sikurajapathy, 2009).

While universal user fees in the public sector and private provider strategies focused on a single price-quality combination service exist, they restrict either public or private provider to a limited segment of the potential demand and in consequence, either as business strategies or as strategies to achieve greater inclusion of public services, two-tier provision has found multiple manifestations. The introduction identifies some of these manifestations: exemption systems from public sector user fees; insurance agencies provision of multiple policies with price and quality combinations (broadly defined to include co-payment and deductible levels; choice of provider; inclusion and exclusion of conditions and services); pharmaceutical companies’ pricing strategies. The situation in which a provider of health related services in a low-income country restricts their offer to a single segment of the potential demand may even be the exception rather than the rule. In that case, understanding the incentives implicit in targeting multiple demand segments simultaneously is a significant gap. The remainder of this paper considers the extent to which the model developed for two-tier hospital provision can assist in understanding other two- or multi-tier scenarios by focusing on a few examples of those.

*Insights generated into exemption policy*

Where user fees are introduced for public health services it is frequently recognised that a segment of demand is consequently excluded and exemptions policies are introduced as a measure to mitigate this exclusion. These range from the exemption of those judged by a means test unable to pay, to the exemption of those presenting with particular conditions judged of high public health importance to the exemption of demographic groups, such as children under 5 or pregnant women (Bitran and Giedion, 2003).

In principle, this appears to produce an analytically relevant situation for the two-tier model. Two groups of patients attempt to access services from the same provider but offering different levels of payment. The model would predict that the levels of quality offered to exempted and non-exempted patients would diverge.

The realities of the implementation of this policy are often complex. A number of empirical studies suggest that means tested exemptions are in practice rarely offered despite the presence of a policy that they should be (Willis, 1993; Gilson et al., 1995; Nyonator and Kutzin, 1999; James et al, 2006), or where offered, that those intended are rarely effectively targeted (Kutzin, 1994). In part this seems to reflect administrative difficulties. Systems to
identify the poorest are rudimentary and not trusted (Blas and Limbambala, 2001). In part, this may reflect a simpler incentive mechanism than that proposed by our model: where some pay and others do not, the incentive is to refuse treatment to non-payers altogether, in a sense an extreme quality restriction.

Some evidence suggests that the incentive driving the failure to grant exemption is as much related to the presence of payers, as it is to the non-payment of the potentially exempted, reinforcing the insights of the model. In 1990s Zimbabwe, an unsustainably high proportion of users was exempted (Hecht et al., 1993). The difference in Zimbabwe’s arrangements to those where almost no-one was exempted is that those providers making exemption decisions did benefit directly from the user charge revenues – hence their incentives were identical to those of a universal free service.

The need to structure incentives so that exemption is offered in practice, as well as the need to ensure the adequacy of funds so that generalised quality attrition is not associated with higher exemption levels, have led to the policy proposal that exemptions should be funded. Health equity funds such as have operated in Cambodia (Hardeman et al., 2004) and target group focused programmes such as the maternal health exemption programme in Ghana (Witter et al., 2007) reflect this understanding. The model would predict that operating these programmes as a per-unit subsidy rather than a flat-rate subsidy would be the more effective mechanism and this is generally the strategy applied. While paying for a specific service (such as delivery) might appear to protect the subsidy from leaking to other programmes, in practice this would be difficult to enforce, and the model predicts that incentives would exist to use the subsidy to increase the demand for priced services - hence per unit subsidy would in principle also be preferred for a programme such as Ghana’s. Under means testing, there is in principle, no need to secure the voluntary willingness to pay (more) of the higher rate paying group and it would be feasible to fund exemption at the same level (in the form of a per-unit subsidy) as the user fee, and eliminate incentives for quality divergence.

Further complexity is introduced by the simultaneous presence of formal and informal fees in many contexts. Exemption from the formal fee need not imply exemption from the pressures to pay an informal fee. Among these pressures is the withholding of elements of quality of care from non-payers. Accounts from Tanzania and Uganda, for example, suggest that the two types of fees are not easily distinguished by users (Mamdani and Bangser, 2004; McPake et al., 1999). It can be consequently difficult analytically to identify formal fee payers, informal fee payers, those exempted from either or both and users’ responses to the inherent incentives in the dual fee situation.

Hence, an absence of clear empirical evidence of quality divergence as predicted by the model can be explained. Nevertheless some limited supportive evidence arose from a study in which it happened that a health centre near a prison was included in a study (McPake et al. 1999). In this case the option of refusing treatment appeared unavailable (perhaps because of the authority represented by accompanying prison officials) and the inducing of informal payment would be untenable. Prisoners without any recourse to funds were reported to be treated with watered down antibiotic injections.

Policy debates concerning exemption have tended to focus on the difficulties of effecting the right to a free service by those who qualify. What this discussion suggests is that were successful mechanisms to support this discovered, they might prove insufficient without
careful parallel attention to the protection of the quality of care levels of the exempted. The model supports the external funding of exempted patients at the same level as the user charge to achieve both objectives, a conclusion that is supported by other commentators (for example Bitran and Giedion, 2003).

**Insights generated into informal charging**

Informal charging itself can present two (or multi) tier scenarios. Unlike in the model, informal charge levels are not set by an external agent but the provider instead resembles a monopolist (single seller) or oligopolist (one of a few sellers) who sets prices, qualities and service volumes simultaneously.

It might be expected that models exist in the economics literature to explain this more common scenario – similar to that in the airline industry where first, business and economy class packages are constructed in terms of price, amenity and seating volumes simultaneously, or to that in car manufacturing where a range of models from basic to luxury models are simultaneously designed, priced and manufactured in given volumes.

The economics literature uses a number of labels that recognise different aspects of what we are terming a two-tier scenario. Market segmentation refers to the strategy that responds to the range of demands for similar products with differentiated products, multiproduct firms are those that supply more than one of these (although they may also supply products in more than one product range) and price discrimination is the strategy that seeks to elicit the maximum willingness to pay from each demand segment, even if for an identical product, although in practice, at least slight product differentiation is often required to elicit that willingness to pay. A separating menu is a product range with differentiated price and quality combinations aiming to separate consumers with heterogeneous preferences into demands for different products.

There are considerable literatures on all these phenomena, yet there is little attention to the determination of the range of quality levels. For the situation of a monopolist, Mussa and Rosen (1978) show that the development of different price-quality combinations is the predicted response to heterogeneity of consumers. In comparison to competitive scenarios, they suggest that quality levels will be lower for each consumer, and that those with the lowest willingness to pay for quality may be excluded from the market altogether.

Developments of these ideas clarify the circumstances under which separating menus will not be selected (for example Kim and Kim, 1996; Acharya 1998), and further explore the market conditions under which market coverage, or the inclusion of those consumers with lower willingness to pay will be achieved (Gabszewicz and Thisse, 1980; Shaked and Sutton, 1982; Constantatos and Perrakis, 1997), largely confirming the importance of larger numbers of firms for that to occur.

These predictions seem plausible in comparison to documented realities in informal charging contexts (McPake et al., 1999; Mackintosh and Tibandebage, 2007; Mbaga Kida, 2009). Effective monopolists in rural areas, where consumers have little choice, offer low quality and exclude the poorest or least willing to pay (McPake et al., 1999). In more competitive urban areas, more low price-low quality combinations are offered by both public and private providers (Mackintosh and Tibandebage, 2007; Mbaga Kida, 2009). While further exploration of the relationship between price and quality ranges and market structure seem relevant,
there would appear a prima facie case for considering market structure in the design even of public health service delivery strategies where informal charging is a recognised reality.

*Insights generated into insurance packaging*

In traditional Latin American health systems, there was separation of market segments such that insurer-provider partnerships each targeted a separate segment of the market. Under reformed models such as in Colombia, Chile and Brazil, insurers and providers have been separated to different degrees while consumer segments have been pooled in an attempt to universalise provision. For example in Colombia, reforms enacted in 1993 created Health Promotion Enterprises to sell insurance packages (including subsidised packages to poorer groups) and contract with a range of provider institutions for the provision of services. The result is that provider institutions such as hospitals contract with a range of Health Promotion Enterprises, negotiating different price-quality combinations with each (Londoño, 1996; Yepes, 2000; McPake et al, 2003).

In principle, hospitals face differential incentives associated with different patient groups and may respond by varying quality levels. Some differentiated quality response is formally negotiated such as through stated exclusions associated with alternative packages. These are observable to consumers at the time of insurance package purchase. As with informal charges, these are likely to provide a range of price-quality combinations responding to the heterogeneity of consumer preferences, and abilities to pay, and responding more completely, the more competitive the market. Hence it should not be expected that such reforms will result in universal provision in the sense of a common quality of provision for all.

What our model also suggests is that further quality differentiation may result where one Health Promotion Enterprise pays a lower price for the same service than another. However, to our knowledge, no research has been conducted that explores this issue and only hypothetical further exploration is possible as to what kinds of manifestations of this behaviour might be plausible. Since patients for whom different prices attached are probably not physically separated, it is not clear whether individual clinicians would recognise patient groups for whom different prices applied. Recognition would be a precondition for discretionary clinician behaviour affecting quality elements such as time spent in examination of patients or in observing basic civilities of conversation. If incentives to discriminate were strong enough, it would be expected that hospitals would evolve systems to allow such discrimination, such as separate wards or other patient labelling mechanisms. Insurers paying higher prices might demand such labelling in return for continued willingness to pay.

*Insights generated into pharmaceutical pricing*

It has been argued that cross-national price differences in pharmaceuticals reflect the practice of Ramsay pricing (for example Danzon and Towse, 2003). This implies that pharmaceutical companies set prices reflecting willingness to pay in different economies with generally equitable outcomes: lower pharmaceutical prices pertain in poorer economies.

However, cross-national studies of pharmaceutical prices present mixed evidence on the existence of this phenomenon. It may help to explain why prices in the USA appear higher
than in other OECD countries (Danzon and Furukawa, 2003) but there are also many documented examples of high pharmaceutical prices in very low-income countries and weak correlation between income and price levels (for example Barton, 2001; Scherer and Watal, 2002) that appear to contradict the theory.

One explanation is that markets need to be effectively separated for Ramsey pricing to be a profit maximising strategy for pharmaceutical companies. Concern for the potential for re-export of cheap pharmaceuticals is an acknowledged constraint to low prices in low income countries (Danzon and Towse, 2003). A strategy applied to support market separation is quality differentiation. If the same active ingredient of a pharmaceutical is differentially packaged or labelled in a way that affects quality perceptions, it may have a sufficient depressing effect on the potential demand in a richer country to which it might be re-exported that market separation can be achieved. Such strategies are explicitly used. For example, a malaria drug is simultaneously supplied a Coartem in poor, malaria endemic countries and as Riamet to high income economies (Grace, 2003).

Hence the price discrimination that would benefit poor consumers in low income countries with lower prices is contingent on the cross-price and quality elasticities of the demands of consumers in high income countries, with parallel interdependencies to the two-tier model. It is unclear what levels are required of these elasticities to allow for effective market separation or how they are related to income levels and other economic characteristics of high income economies. Important insights could be generated by further research in this area. To provide one example, how might global recession affect the Ramsay-like or other pricing strategies of pharmaceutical companies?

**Conclusion**

The two-tier model highlights the potential policy significance of the incentives inherent in arrangements that result in differential prices for different consumers of the services of a single provider. This paper argues that such arrangements are widespread and have many manifestations. They may even be the norm in some low income settings.

The two tier model fits some of these situations poorly and the first gap highlighted is therefore in model development. More theoretical understanding is required of the incentives inherent in a range of two-tier scenarios and to better understand how to more successfully wrest equity, quality and coverage outcomes from systems with multiple manifestations of two tier provision.

Nevertheless, the model is able to directly generate predictions about behaviour in a wider range of situations than it was designed for. While some of these predictions match existing evidence of patterns of provision, in most cases there is no research that addresses the issues raised and therefore a lack of supporting evidence of the operation of the model. A main concern of this paper is therefore to highlight this research gap and its importance for the understanding of the relationship between health system and market arrangements including new forms of public private mix and equitable access to quality health care.

In a number of respects it is clear that debates overlook these issues with potentially negative implications for policy. A few examples illustrate this point.
Most equity analyses assume the same service is received from the same provider and equity is increased when there is a more even pattern of utilisation of a given mix of providers. Benefit Incidence Analysis (BIA) provides one example of this (ref). BIA treats an episode of health service utilisation as a unit of achieved use of resources. Differential use of resources within the facility is ignored, which would not be a problem if such differentiation were not systematic. Our analysis suggests it may be. There is some recognition of this limitation in the orthodox approach. For example, study in Bangladesh re-estimated BIA in order to understand the implications of gender differentiated service receipt (Ensor et al., 2003).

Similarly, standard approaches to quality of care measurement assume that it is adequate to use the provider facility as the unit of analysis. Donabedian (1980) provides the standard framework for quality of care measurement. The starting point of this measurement is ‘structure’ which measures the availability of infrastructure in a health facility, a measurement which is then combined with those of process and outcome to produce quality scores that can only be disaggregated to provider facility level. Greater attention to intra-facility quality differentiation will require the development of new quality of care measurement strategies.
References


http://www.icrier.org/pdf/barton63.PDF


http://healthaff.highwire.org/cgi/reprint/hlthaff.w3.521v1.pdf


Donabedian, A. (1980) The definition of quality and approaches to its assessment, Health Administration Press

http://books.google.co.uk/books?q=benefit-incidence-analysis+gender+Ensor&ots=sCx_oYdP2X&sig=Q7i9SLLaX0zggsUg4ldNiAx4k#v=onepage&q=benefit-incidence&f=false


Willis, C. (1993) *Means testing in cost recovery for health services in developing countries*. Major applied research paper no. 7, Health Financing and Sustainability Project
