WORKING PAPERS IN ECONOMICS

No. 14/08

OTTAR MÆSTAD AND GAUTE TORSVIK

IMPROVING THE QUALITY OF HEALTH CARE WHEN HEALTH WORKERS ARE IN SHORT SUPPLY

Department of Economics

UNIVERSITY OF BERGEN

Improving the quality of health care when health workers are in short supply

Ottar Mæstad*and Gaute Torsvik†

July 2, 2008

Abstract

A number of low- and middle-income countries have a severe shortage of health workers. This paper studies how health workers' choices of labour supply and work effort impact on the quality of health services when health workers are in short supply. We analyse how policy measures such as monetary incentives, monitoring, provisions of quality-enhancing inputs, and the building of professionalism and organisational identity can improve the quality of health care in the presence of a health worker shortage. We find that to pay health workers based on the number of patients may have a positive impact on the quality of health care even if quality does not affect demand. Furthermore, provision of quality-enhancing drugs and equipment may reduce health workers' effort in delivering quality care, thus diminishing the positive impact of such interventions. Our most surprising result is that if the actual quality of health care is far below a professional standard, measures to build a professional mindset among health workers may reduce the quality of care.

*Chr. Michelsen Institute. E-mail: Ottar.mestad@cmi.no

†University of Bergen, Department of Economics and Chr. Michelsen Institue. E-mail: Gaute.Torsvik@econ.uib.no

1 Introduction

Inadequate health worker performance is a widespread problem in many low and middle income countries (Rowe et al., 2005). Poor performance, such as weak compliance with clinical guidelines, is a threat to population health in these countries, not only because low quality health services may be harmful to the patients, but also because poor quality will reduce the utilisation of health services in general.

The reasons for inadequate health worker performance are poorly understood. Historically, much attention has focussed on the lack of sufficient knowledge and skills in the health workforce. Recent evidence suggests, however, that many health workers provide services with a quality level significantly below that of which they are capable, given their actual level of knowledge and the physical infrastructure they have at their disposal (Das and Hammer, 2007; Leonard et al., 2007). The existence of such a know-do gap suggests that more training is not the only, and perhaps not the best, way of improving health worker performance.

One fact which might explain both a low absolute level of performance and a know-do gap is the severe shortage of health workers in many low income countries (Barber et al., 2007). The problem is particularly acute in Sub-Saharan Africa, which has 24 percent of the global burden of disease but only 3 percent of the health workforce (WHO, 2006). If health workers are too few, they may simply not have sufficient time to provide services of adequate quality. The current health worker shortage calls for a reassessment of policies for improving the quality of health services in low income countries. To this end, we develop a model to analyse how health worker performance and the quality of health services are affected by the shortage of health personnel, and how quality can be improved in such a setting.

One important issue is how an increased supply of drugs, equipment and other "performance-enabling factors" impacts on the quality of health services when health workers are time-constrained at the outset. Better drug supply, for instance, may attract a higher number of patients to the health facilities and thus increase the workload. Health workers may then be forced to spend even less time with each patient, which is likely to reduce quality. Such behavioural responses to increased supply of equipment and drugs are not well understood. The first contribution of this paper is to explore possible linkages between the supply of "performance-enablers" and health worker behaviour in a situation with a health worker shortage.

If low quality is a consequence of a health worker shortage, a natural policy response is to train and recruit more health workers. While it is certainly important to increase the number of health workers, other policy options also seem available. Recent studies from several low income countries have found high rates of absenteeism in health facilities, up to 40% in some areas (Chaudhury et al., 2006; Banerjee and Duflo, 2006). If low health worker performance is caused partly by an excessive workload, quality may improve if absenteeism is reduced. Hence, policies which primarily aim at reducing absenteeism might deserve a place among the tools that can be used for improving health service quality. The second contribution of this paper is to examine how labour supply decisions and labour supply policies affect the quality of health services in the presence of a health worker shortage.

A high workload is unlikely to be the only explanation for the observed know-do gap. Recent literature suggests that motivational problems also play a crucial role in explaining low levels of performance (Leonard et al., 2007; Das and Hammer, 2007). Motivational problems can partly be ascribed to lack of economic incentives or the lack of other extrinsic sources of motivation, such as unclear career paths and unpredictable patterns of promotion (Manongi et al., 2006). Others have emphasised that the internal motivation or vocation of health workers is under stress in many low income countries (WHO, 2006). In focus group discussions that we conducted with Tanzanian health workers, many participants pointed both at lack of monetary rewards and a decline in the sense of vocation as important reasons for sub-standard performance.

Performance-based pay has recently been advocated as a promising tool for improved health system performance in low income countries (Meessen et al., 2006; NORAD, 2007). Performance-based pay can be seen as a way to address the problems caused by low extrinsic and intrinsic motivation among health workers. In practice, performance-based pay in low income countries boils down to some kind of output-related reward (e.g., a reward related to the number of patients). At first glance, output-based pay may seem quite misplaced in health systems which are severely capacity constrained by a shortage of health workers. However, since capacity limits are not necessarily absolute, not even in the short run (e.g., due to high levels of absenteeism), the scope for output-based pay might be larger than it seems at the outset. The third contribution of this paper is to investigate the potential role of output-based pay for the quality of services in a setting with both a health worker shortage and endogenous labour supply. This represents an extension

of the work by Ma (1994) who discusses the impact of output-based financing on the quality of health care when labour supply is exogenous.

Monetary incentives are not the only way of addressing motivational problems. Nor are they an ideal measure, as the quality of health services will often be extremely costly to observe and verify on a regular basis. Therefore, much emphasis has traditionally been placed on building professional attitudes among health workers. This can be seen as an attempt to make the pursuit of certain quality standards part of the intrinsic motivation of the health workers (i.e., to make the health workers willing to incur personal costs in order to adhere to the quality standards even if deviation is non-observable). While the possibilities to change workers' objectives have been discussed within sociology for decades (Barnard, 1938; Selznick, 1957) and are an important part of current management theories, these options have been almost neglected within the economics literature. Health worker behaviour in the presence of professionalism has been discussed by Woodward and Warren-Boulton (1984) and Gaynor et al. (2004), among others, but in these papers, the strength of professional attitudes is taken as exogenous parameters. Our fourth contribution is to introduce the possibility that the degree of professionalism might be influenced by policy makers. We show, somewhat surprisingly, that a strengthening of professional attitudes might reduce the quality of health services when quality at the outset falls far short of the professional standard.

The usual way of modelling professional attitudes among health workers is to assume that their utility is reduced when actual performance differs from some ideal level of performance. This modelling may not capture very well the attitude of genuine patient care, which has also received much attention in the medical profession, such as, for instance, in the Hippocractic Oath. To care for the patients, or to maximise the health of the population, could be seen as the ultimate goal of the health sector, and thus as the "organisational goal" of any health care institution. Following Akerlof and Kranton (2005), one way of addressing motivational problems in organisations is to build an "organisational identity", i.e., to align the goals of individual workers with the goal of their organisation. We argue that a natural operationalisation of the concept of organisational identity in the health sector, or genuine patient care, might differ from the concept of professionalism. Our fifth contribution is to explore how the impacts on the quality of health services of a stronger organisational identity might differ from the effects of stronger professionalism.

2 Health worker motivation, behaviour and the quality of health services

This section develops an analytical framework that can be used to examine policy options for improving the quality of health services when health workers are in short supply and the rate of absenteeism in health facilities may be large. The analytical framework highlights the importance of individual health workers' choices for the quality of health services. We start by formalising how the quality of health services is affected by the level of effort that health workers put into the production of health services and by their labour supply. Thereafter, we specify the motivational factors which, together with the incentive structure facing the health workers, explain their behaviour.

2.1 Determinants of health service quality

When health workers are in short supply, both their choice of work effort and work hours may have an impact on the quality of the health services. We define effort as any activity that improves the clinical quality of the services, including thorough history-taking and physical examination, but also activities that increase the patient's feeling of convenience, comfort and education about medical conditions (see Wedig et al., 1989). The quality of health services is also influenced by a number of factors beyond the control of the individual health worker, such as the supply of equipment and drugs, the quality of their education etc. Variable x denotes these external factors, e is the level of effort, and q represents the quality of health care. The quality of health services can then be written as $q = q(e, x)$. We assume that quality is non-decreasing in both e and x (i.e., $q_e(e, x) \geq 0$, $q_x(e, x) \geq 0$).

Proper history taking, examination, and medication of patients are time consuming. Following Ma and McGuire (1997), we therefore assume that the level of effort e per patient is equivalent to the time spent with each patient. Accordingly, we assume that there exists a minimum level of effort $\underline{e}$ such that $e \geq \underline{e} > 0$. The maximum time that can be spent on each patient is equal to the number of hours l that the health worker spends at the health facility divided by the number of patients n. The level of effort is thus bounded from above by the constraint $e \leq l/n$. In the following, health worker shortage is defined as the case when this constraint is binding, i.e., when $e = l/n$. Hence, when there is a health worker shortage, because there

are few health workers and/or because those who are there are often absent from work, there will be a direct association between the health workers' supply of labour and their effort levels, and thus between labour supply and the quality of health services.

The association between labour supply on the one hand and the level of effort/quality on the other is somewhat more involved than suggested so far, because the number of patients n may increase with the quality of the health services. Following McGuire (2000), we assume that the number of patients is a function of the net benefits they receive from health services. We formalise net benefits as the difference between the quality q and the user fee p. Hence, the number of patients will be

$$n = n\left(q(e, x) - p\right).$$

In the case of a health worker shortage, the constrained level of effort $\hat{e}$ is then implicitly defined by the relationship

$$\hat{e} = \frac{l}{n\left(q(\hat{e}, x) - p\right)}. \tag{1}$$

We assume that the user fee is unresponsive to the number of patients, as user fees in many low income countries are defined by government authorities and do not normally serve the role as a market clearing device. The relationship between labour supply and effort per patient can then be described as follows

$$\frac{d\hat{e}}{dl} = \frac{1}{n\left(1 + \varepsilon_{n,e}\right)} > 0, \tag{2}$$

where $\varepsilon_{n,e} \equiv \frac{\partial n}{\partial q}\frac{\partial q}{\partial e}\frac{e}{n}$ is the elasticity of the number of patients with respect to the level of effort. Effort (and hence the quality of health services) is monotonically increasing in labour supply. But the higher the number of patients, the smaller will be the gain in effort for a given increase in hours worked, because the increased amount of available time has to be shared across a larger number of patients. Thus, with a positive demand response (i.e., $\partial n/\partial q > 0$), the effort level will tend to be a concave function of labour supply.[1]

[1] In theory, the function $\hat{e}(l)$ may also have non-concave segments, for instance if the demand response is non-continuous in the sense that there is level of effort $\tilde{e}$ such that $\partial n/\partial e >> 0$ for $e < \tilde{e}$, while $\partial n/\partial e = 0$ for $e \geq \tilde{e}$.

2.2 Health worker motivation

In health economics, as in economics more generally, it is still common to assume that work behaviour is driven solely by narrow self-interest.[2] It is often acknowledged that many health workers' motivation extends beyond their narrow self-interests, but these motivations are seldom explicitly modelled, and there is little agreement about how it should be done (McGuire, 2000). One possibility is to model other-regarding concerns as a constraint on behaviour. Ma and McGuire (1997) follow this line when they assume that physicians must provide health benefits above a certain threshold. Another alternative is to assume that health personnel make trade-offs between selfish and other-regarding concerns, as in Woodward and Warren-Bolton (1984). We follow the latter approach by assuming that a health worker trades off her narrow self-interests against two kinds of other-regarding concerns that seem to be of particular relevance in the health sector; professionalism and organisational identity. We start by modelling the health worker's narrow self-interests.

Self-interest The health worker in our model enjoys income from the facility but finds it unpleasant to exert effort above some critical level. She also takes account of how the time spent at the facility prevents her from pursuing other activities. To be more specific, the health worker allocates a fixed time endowment $\bar{l}$ between health services production and other activities (e.g., other work or leisure). Time spent in health service production generates an income I and some personal effort costs $c(\cdot)$. Time spent on other activities $(\bar{l}-l)$ generates a constant benefit z per unit of time. Income from health service production consists of a fixed salary w and maybe also an output-based salary component related to the number of patients seen. The output-based component can for instance be calculated as a share of the user fees charges, or it can be an output-bonus paid for by the employer without any relation to the user fees. Let ϕ denote the additional pay per patient. Income is then $I = w + \phi n$.

We allow for the possibility that the discovery of illegitimate absenteeism may trigger some kind of sanctions. Sanctions may result in reduced future

[2]This practice takes place despite a fast growing heap of evidence that individual behaviour is swayed by other-regarding motivations; see Fehr and Schmidt (2006) for labaratory experiments on social behaviour and other-regarding preferences, and Rotemberg (2006) for a discussion of social motivation at workplaces.

income opportunities, for instance through less promotions or even through a termination of the work relation. Proper modelling of such sanctions would call for a multi-period model, but we limit ourselves to a reduced form specification by assuming that income is received with a probability $\pi(l) \leq 1$. π is increasing in l as long as there is illegitimate absenteeism. π equals one if there is no illegitimate absenteeism.[3]

Personal effort costs $c(\cdot)$ are increasing ($c' > 0$) and convex ($c'' > 0$) in the aggregate effort exerted per working day, which is equal to the effort per patient multiplied by the number of patients seen.

If we let S capture the health worker's narrow self-interests, we can now write;

$$S \equiv \pi(l)I(e) - c(ne) + z(\bar{l} - l). \tag{3}$$

Professionalism One relevant motivation extending self-interest is a preference for adhering to a professional standard. We model a health worker's professional attitude as a loss in utility whenever actual quality q deviates from some "ideal" level of quality q^N. We interpret q^N as reflecting clinical guidelines or some other medical standard. Following Gaynor et al. (2004), we further assume that the utility loss is higher the larger the number of patients treated at sub-standard quality levels. A simple way of capturing these aspects is to express the professional concern P as follows;

$$P \equiv -n \left| q^N - q \right|. \tag{4}$$

We will assume throughout that $q < q^N$.

Organisational identity (or altruism) A professional attitude geared at achieving some medical standard does not necessarily fully capture a motivation to care for the well-being of the patients. For instance, a health worker with a professional attitude as specified in Eq. (4) would experience a loss in utility whenever the number of patients increases (as long as q differs from q^N). This specification certainly has some merit, but it fails to capture the fact that many health workers would also experience some kind of utility

[3] π can also be a function of the level of effort in case there is a probability that unacceptable low efforts will be discovered and sanctioned. The level of effort is however far more difficult to observe and verify on a routine basis than the level of absenteeism.

gain by treating a larger number of patients. Such altruistic attitudes towards the patients, which have been strongly promoted in the health sector, would require a different specification.

The degree of altruism towards patients can be viewed as a personal trait which is more or less unchangeable. In our opinion this interpretation is too narrow. Following Akerlof and Kranton (2005), we believe that it is possible to develop behaviours which are consistent with altruism by strengthening the health workers' "organisational identity", i.e., the degree to which they adopt the goals of the health care sector as their own goals. A central goal of any health care institution is to improve patient welfare. Building organisational identity in the health sector would therefore likely make health workers behave more altruistically. In fact, we believe that organisational identity is easier to sustain in the health sector and other "idealistic" organisations than in many other businesses.

In our set-up, the objective of the health sector O is to improve patient welfare, as represented by the number of patients treated times the average quality of the treatment;

$$O \equiv nq$$

In sum, then, we portray a health worker who varies her work hours and work effort so as to maximise

$$U(l,e) = S(l,e) + \alpha O(l,e) + \gamma P(l,e),$$

where α and γ represent the degree of organisational identity (and/or altruism towards patients) and the degree of professionalism.

Julian Le Grand (2003) has presented a motivational dichotomy for professional workers where he distinguishes between knights and knaves. Knights are honourably committed to deliver high quality services to the public they serve, while knaves are selfish and care only about their personal gains. In our model, a knave would be characterised by $\alpha = \gamma = 0$. Such a person will be denoted a *selfish* health worker. Our model has no true knights, but we have health workers who trade off their narrow self-interest against professional attitudes and/or organisational goals (i.e., α and/or γ are greater than zero). Such a person will be denoted an *ethical* health worker.

2.3 Choices: The selfish health worker

A selfish health worker chooses her labour supply and effort to maximise

$$U(l,e) = [w + \phi n(e)]\,\pi(l) - c(ne) + z(\bar{l} - l)$$

subject to the constraints

$$\begin{aligned} l &\leq \bar{l} \\ \underline{e} &\leq e \leq \frac{l}{n}. \end{aligned}$$

2.3.1 No health worker shortage

It is instructive to start with the case without a health worker shortage, i.e., when $e \leq l/n$ is non-binding, implying that e can be chosen independently of l. Assuming an interior solution, the first order conditions for optimal choice of working hours l^* and effort e^* are then given by

$$\frac{\partial U}{\partial l} = 0 \Leftrightarrow I\frac{\partial \pi}{\partial l} - z = 0, \tag{5}$$

$$\frac{\partial U}{\partial e} = 0 \Leftrightarrow \phi\frac{\partial n}{\partial e}\pi(l) - c'\left[\frac{\partial n}{\partial e}e + n\right] = 0. \tag{6}$$

The working time allocated to health service production is decided by equalising the marginal increase in expected income (through a reduction in the probability of sanctions related to illegitimate absenteeism) with the marginal benefits z from alternative activities.

The level of effort is decided by equalising the marginal increase in incomes from higher efforts with the marginal increase in effort costs. Income is increasing in effort if there is a demand response ($\partial n/\partial e > 0$) and if income is linked to the number of patients ($\phi > 0$). Effort costs are increasing in the level of efforts both because it is personally costly to treat each patient more carefully and because the health worker must treat more patients as higher quality causes an increase in the number of patients n. If there is no output-based income ($\phi = 0$), the selfish health worker chooses the minimum level of effort $\underline{e}$.

With no health worker shortage, it is unlikely that there will be any association between labour supply and the quality of health services, even

though l appears in Eq. (6). Since there is no shortage, health workers will spend some of their working time just waiting without attending any patients. With excess labour supply it seems unlikely that a small reduction in l will affect the likelihood of sanctions being imposed due to illegitimate absenteeism. Hence, π can be treated as a constant, and e will be independent of l.

2.3.2 Health worker shortage

We have defined a health worker shortage as the situation where the level of effort exerted per patient is effectively constrained by the condition $e \leq l/n$. That is, a shortage exists when, at the actual level of labour supply, the health workers do not have sufficient time to exert the level of effort that they would have chosen if labour were in more abundant supply. This definition, by taking into account the labour supply decision of the existing workforce, represents a more compelling operationalisation of the concept of a health worker shortage than the standard approach of simply focusing on the number of health workers relative to some predefined standard. In particular, this operationalisation takes into account the important role of high rates of absenteeism in understanding the degree of health worker shortages in several low income countries. At the same time, it points to the fact that with high rates of absenteeism, health worker shortages may be at least partly addressed through increased labour supply from the existing workforce.

The health worker's utility in the case of a health worker shortage is defined as

$$U(l) = [w + \phi n(\hat{e}(l))]\,\pi(l) - c(l) + z(\bar{l} - l)$$

The optimal choice of working time $\hat{l}$ (and hence effort $\hat{e}$) is given by the first order condition

$$\frac{\partial U(l)}{\partial l} = 0 \Leftrightarrow I\frac{\partial \pi}{\partial l} - z + \phi\frac{\partial n}{\partial \hat{e}}\frac{d\hat{e}}{dl}\pi(l) - c' = 0 \tag{7}$$

The marginal benefits of increased labour supply consist in this case not only of the increase in expected income due to reduced absenteeism, but also of the higher incomes that may be generated through a positive demand response when the level of effort and the quality of services increase. The marginal costs of increased labour supply now include not only the marginal

value of other activities z, but also the marginal costs of increased effort levels c'.

Despite the increase in both marginal benefits and marginal costs of labour supply, it is straightforward to demonstrate that a health worker shortage will cause an increase in the labour supply, i.e., $\hat{l} > l^*$. The level of effort per patient will decline, i.e., $\hat{e} < e^*$.[4]

The optimal choice is illustrated in Figure 1.

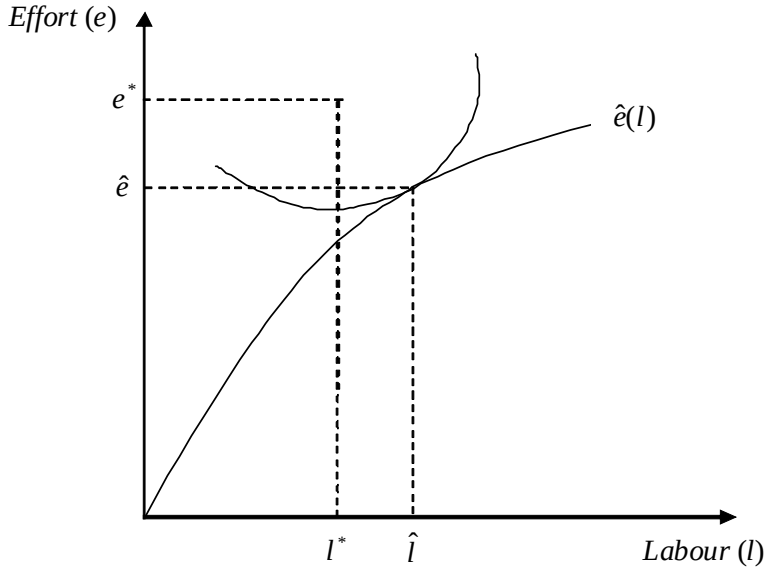

Figure 1: Unconstrained (l^*, e^*) and constrained optimum $\left(\widehat{l}, \widehat{e}\right)$.

[4]This follows from the fact that l^* and e^* solve the first order condition for an interior optimum, which means that the first order effect of a small adjustment in l and e away from l^* and e^* is zero. Hence, the health worker will adjust both variables in order to satisfy the constraint imposed by the health worker shortage.

Formally, the result can be shown by the first order condition, which can be written as $I\frac{\partial \pi}{\partial l} - z + \left[\phi \frac{\partial n}{\partial \hat{e}} \pi(l) - c' \left[\frac{\partial n}{\partial \hat{e}} \hat{e} + n\right]\right] \frac{d\hat{e}}{dl}$. Assume that $\hat{e} < e^*$. Then, the term in square brackets, which is the first order condition for e^* in the unconstrained case, is bound to be postive. Since $d\hat{e}/dl$ is positive as well, the sum of the two first terms, which is the first order condition for l^* in the unconstrained case, must be negative, implying that $\hat{l} > l^*$. By the same argument, $\hat{e} > e^*$ would imply $\hat{l} < l^*$, which is an impossibility given the constraint imposed by the health worker shortage. Hence, $\hat{e} < e^*$ and $\hat{l} > l^*$.

Whenever (e^*, l^*) lies above the $\hat{e}(l)$ curve, there will be a health worker shortage. The health workers will then choose the point at the $\hat{e}(l)$ curve which maximises their level of utility, represented by the indifference countors circling around the unconstrained optimum (e^*, l^*).

2.4 Choices: The ethical health worker

Let us now introduce a health worker who is motivated not only by narrow self-interest, but also by professionalism and/or a desire to deliver high quality health services to many patients (i.e., organisational identity or altruism). In this section, we confine the discussion to the case with a shortage of health workers. Health workers then choose their working time l in order to maximise their utility

$$U(l) = S(l, \hat{e}(l)) + \alpha n(\hat{e}(l))q(\hat{e}(l)) - \gamma n(\hat{e}(l)) \left[q^N - q(\hat{e}(l))\right]$$

where $S(\cdot)$ captures all the narrowly self-interested concerns discussed above.

The first order condition for an interior optimum is

$$\frac{\partial S}{\partial l} + \alpha \left(\frac{\partial n}{\partial \hat{e}} \frac{d\hat{e}}{dl} q + \frac{\partial q}{\partial \hat{e}} \frac{d\hat{e}}{dl} n \right) - \gamma \left(\frac{\partial n}{\partial \hat{e}} \frac{d\hat{e}}{dl} \left[q^N - q\right] - \frac{\partial q}{\partial \hat{e}} \frac{d\hat{e}}{dl} n \right) = 0 \qquad (8)$$

The first term captures how a small increase in working time affects self-interested concerns, as discussed above. The second term captures the increase in utility that comes from the fact that an increase in time spent at the clinic both increases the number of patients treated and the quality of the service. The last term captures how spending more time at the clinic affects the costs of deviating from a professional standard. This term includes both a positive and a negative component. More time spent at the clinic has a positive effect on utility since it increases effort per patient and thus brings the quality closer to the professional norm. But higher quality also attracts more patients to the facility, which implies a utility loss whenever the treatment does not satisfy the professional standards (i.e., when $q < q^N$).

By utilising Eq.(2), the first order condition (8) can be rewritten as

$$\frac{\partial S}{\partial l} + \frac{\partial q / \partial \hat{e}}{1 + \varepsilon_{n,e}} \left[(\alpha + \gamma) (\varepsilon_{n,q} + 1) - \gamma \varepsilon_{n,q} \frac{q^N}{q} \right] = 0 \qquad (9)$$

where $\varepsilon_{n,q}$ is the elasticity of the number of patients with respect to the level of quality.

One important observation from Eq. (9) is that it is not obvious that an ethical health worker will work more and exert higher efforts per patient than a health worker who is concerned solely with her narrow self-interest. The optimal levels of l and e will be higher only if

$$(\alpha + \gamma)(\varepsilon_{n,q} + 1) - \gamma\varepsilon_{n,q}\frac{q^N}{q} > 0 \tag{10}$$

This condition is satisfied as long as the actual quality of health services q is not too different form the professional norm q^N. Hence, in those cases where the actual performance does not deviate much form professional standards we can expect that the ethical health worker will work more and exert higher efforts than the selfish health worker. However, if the actual quality falls far below the professional norm, a health worker which experience a large loss in utility by treating patients in a sub-standard way may actually choose to work less and thus exert less effort than the selfish health worker. The reason is that the number of patients receiving inadequate treatment will decline when her working time is reduced. In other words, if it is impossible to achieve more than an extremely low quality of service, the professional health worker may rather choose to stay away.

It is straightforward to show that if there is no shortage of health workers, the ethical health worker would choose the same working time l as the selfish health worker, simply because the choice of l has no effect neither on the number of patients treated nor on the quality of the service. Whether or not the ethical health worker would exert higher efforts would depend on exactly the same factors as in the case of a health worker shortage (see (10)).

3 Alternative ways of improving the quality of health services

We will use our framework to analyse how different policy measures will affect health worker behaviour and the quality of health services. The policy instruments we consider fall into three categories. First, there are policies directed at increasing the supply of various "performance enabling factors", such as the supply of equipment and drugs, the level of knowledge and skills, etc. Second, there are policies which affect monetary incentives and the degree of monitoring and control. The third group of policy instruments is

those which attempt to change the health workers' motivational structures, such as the degree of professionalism and/or organisational identity.

3.1 Performance enabling factors

The traditional approach to improving the quality of health services in low income countries has focussed much on the insufficient supply of various performance enabling factors (e.g., drugs, equipment, knowledge and skills). In our model, these factors are captured by the parameter x, and they are assumed to have a direct positive impact on the quality of health services.

One often neglected aspect in discussions of the impact of performance enabling factors, is their potential indirect effects through changes in health worker behaviour (and even in their motivation). Our analytical framework allows us to study how an increase in x may affect the health workers' supply of labour l and level of effort e. Obviously, if there are both a shortage of health workers and a positive demand response, the first order effect of an increase in x is to reduce the time available per patient and thus reduce the level of e (see Eq. (1)). In other words, there will be a crowding-out of the quality of the services through reduced effort per patient. However, if the labour supply also increases when x increases, the crowding-out effect will diminish and might perhaps disappear. It is therefore of great interest to discuss how labour supply might respond to an increase in x.

In general, the effect of x on $\hat{l}$ can be found as $d\hat{l}/dx = -U_{lx}(\hat{l})/U_{ll}(\hat{l})$, where $U_{ij} \equiv \partial^2 U/\partial i \partial j$. By the second order condition, $U_{ll}(\hat{l}) < 0$. The sign of $d\hat{l}/dx$ thus equals the sign of $U_{lx}(\hat{l})$.

Consider the case of a selfish health worker in the presence of a health worker shortage. $U_l(\hat{l})$ is then given by Eq. (7), implying that the sign of $d\hat{l}/dx$ will be given by

$$sgn\frac{d\hat{l}}{dx} = sgn\left[\phi\left(\frac{\partial n}{\partial x}\frac{\partial \pi}{\partial l} + \frac{\partial^2 n}{\partial e \partial x}\frac{d\hat{e}}{dl}\pi + \frac{\partial n}{\partial e}\frac{d^2\hat{e}}{dldx}\pi\right)\right] \tag{11}$$

There are three ways in which changes in x may affect the labour supply (and thus the effort per patient) of a selfish health worker. First, with higher x and a positive demand response, the number of patients will increase. This will increase health worker income if income is positively related to the number of patients, and it will therefore weaken the incentives for illegitimate absenteeism and stimulate labour supply (provided $\partial\pi/\partial l > 0$). Second, the

level of x may affect the "effectiveness" of effort in producing quality services. For instance, if x and e are complements (i.e., $\partial^2 n/\partial e \partial x > 0$), an increase in x might strengthen the incentives for work because higher effort now has a stronger impact on quality and thus on the number of patients. Third, an increase in x will affect the slope of the $\hat{e}(l)$ curve, i.e., the marginal rate of transformation of labour supply into effort per patient. The first order effect is a reduction in the slope of the $\hat{e}(l)$ curve, as a higher number of patients will reduce the attainable level of $\hat{e}$ for a given supply of labour. A reduction in the slope the of $\hat{e}(l)$ curve pulls towards a reduction in labour supply, as a marginal increase in labour supply will have a smaller impact on effort and hence on the number of patients.

Note that all these three mechanisms will impact on labour supply only if the utility of the health worker is a positive function of the number of patients. For a selfish health worker, therefore, x will impact on labour supply only if there is an output-bonus ($\phi > 0$). As can be seen from Eq. (11), $d\hat{l}/dx = 0$ if $\phi = 0$. A positive demand response in the wake of an increase in x will then translate directly into a reduction in the effort per patient.

In general, the effect of x on labour supply $\hat{l}$ is ambiguous. Consider an example where $n = q - p$, $q = \kappa e + v(x)$, $p = 0$, and $\pi = 1$ in the relevant range of $\hat{l}$. Then,

$$sgn\frac{d\hat{l}}{dx} = sgn\left[\phi\frac{\partial n}{\partial e}\frac{d^2\hat{e}}{dldx}\right] = sgn\left[\phi\kappa\frac{\frac{v'}{n^2}\left(-n^2+\kappa\hat{l}\right)}{\left(n+\frac{\partial n}{\partial e}\frac{\hat{l}}{n}\right)^2}\right]$$

which clearly can be either positive or negative. Figure 2 below illustrates how an increase in x might lead to a reduction in the optimal labour supply in the case of a health worker shortage and thus reinforce the crowding-out of effort per patient. The increase in x tilts the $\hat{e}(l)$ curve downwards. The increase in the number of patients combined with a substitution effect pulling towards a lower level of $\hat{l}$ implies that labour supply and effort decline from $(\hat{l}_0, \hat{e}_0)$ to $(\hat{l}_1, \hat{e}_1)$.

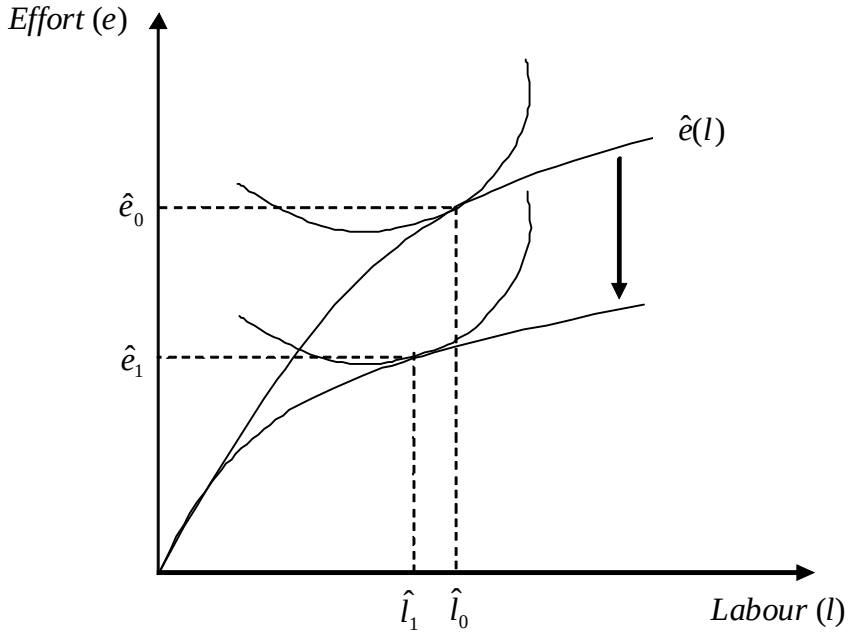

Figure 2: Higher x may cause a reduction in $\widehat{l}$.

For the ethical health worker, an increase in x will produce additional both positive and negative effects on the marginal benefits of increased labour supply. For instance, a higher level of q and n will have a direct positive impact on the marginal benefits of labour supply both for the professional health worker and for the health worker with a strong organisational identity (see Eq. (8)). But there are also other mechanisms at work. A decline in the slope of the $\hat{e}(l)$ curve, for instance, will pull in the opposite direction.

Finally, note that a negative effect of x on effort e is not unlikely even without a health worker shortage. In that case, all factors which contribute to a higher number of patients will increase the marginal effort costs (see the last term in Eq. (6)). A positive effect on e is however also conceivable, provided there are complementarities between e and x so that $\partial^2 n/\partial e \partial x$ is positive.

3.2 Monetary incentives and monitoring

It is often prohibitively costly for relevant principals to observe and verify the quality of the health services on a regular basis. Therefore, it is difficult to design effective incentive mechanisms to improve the quality of the services. However, as argued by Ma (1994) and others, if improved quality generates a positive demand response, quality levels may be improved by letting health workers' income increase with the number of patients, i.e., through output-based financing.

In our model, output-based financing is implemented when $\phi > 0$. With no shortage of health workers, it is easily seen from Eq. (6) that a higher output-bonus will translate into higher efforts per patient and thus into higher quality of the health services, which is a direct parallel to the results of Ma (1994).

Obviously, this result no longer holds if there is a health worker shortage and labour supply is exogenous, as it will be if the existing workforce spends all available time in the production of health services (i.e., $l = \bar{l}$). Output-based pay ($\phi > 0$) will then only increase the salary of the health workers and not change their behaviours. This is a trivial result, but it nevertheless deserves some attention at a time when aid donors are pushing for performance-based pay in health sectors which appear to be capacity constrained through a low number of health workers.

However, in health systems with both a health worker shortage and high rates of absenteeism, there may still be a rationale for implementing output-based financing. In this case, an increase in ϕ will induce health workers to spend more time at their working stations, because health service production becomes more valuable compared to alternative activities (see Eq. (7)). The health worker will then have more time per patient, which will enable her to exert a level of effort per patient closer to her first-best (unconstrained) effort level e^*. Quality will then increase.

Unlike in Ma (1994), a positive effect of output-based financing on the level of quality can occur in our model even in the absence of a positive demand response. This can be seen from Eq.(7) by realising that an increase in ϕ will raise the income level I and hence stimulate to increased labour supply even if $\partial n/\partial \hat{e} = 0$. A higher level of income will increase labour supply because it becomes more important to avoid sanctions due to illegitimate absenteeism. For the same reason, even an increase in the fixed salary w may lead to increased effort and higher quality of health services in a situation

with both health worker shortage and absenteeism.

Higher income will not influence labour supply if there are no effective mechanisms for monitoring and sanctioning illegitimate absenteeism. Without such mechanisms, a change in labour supply will not have any impact on expected incomes (the probability π would be unaffected by l, i.e., $\partial\pi/\partial l = 0$). One way of formalising how the probability π might depend both on l and on the degree of monitoring m would be as follows

$$\pi(l,s) = \begin{cases} 1 - \frac{\beta - l}{\beta} m & if \quad l < \beta \\ 1 & if \quad l \geq \beta \end{cases}$$

Here, β can be interpreted as the number of working hours specified in the employment contract. π is equal to one as long as the health worker is present at work as much as she is supposed to ($l \geq \beta$). However, if her labour supply falls below β, there is a probability of some kind of sanctions ($\pi < 1$) insofar as the level of absenteeism is monitored (i.e., $m > 0$).

In this case, a higher level of monitoring of absenteeism may have a positive impact not only on labour supply but also on the level of effort and the quality of health services. This follows straightforwardly from Eq. (7) and the fact that a higher level of monitoring will make it more profitable for health workers who are sometimes absent to increase their labour supply, i.e.,

$$\frac{\partial}{\partial m}\left(\frac{\partial \pi}{\partial l}\right) = \frac{1}{\beta} > 0 \qquad if \quad l < \beta$$

In summary, the main message from this discussion is that in the case of a health worker shortage, all kinds of incentives which stimulate to increased labour supply will contribute to increased quality of health services. Hence, the number of policy instruments that can be used in order to improve service quality will be much larger than in the case of abundant labour supply.

The mechanisms for improving quality discussed in this section will of course only have an effect to the extent that health workers are motivated by monetary incentives. In our model, there is no difference between the selfish and the ethical health worker in this respect. Both types of workers respond in qualitatively the same way to a strengthening of monetary incentives and the level of monitoring. In reality, there are differences among workers with regards to how sensitive their choices are to monetary rewards.

We started this section by claiming that to link monetary incentives to

the level of quality or effort is not feasible due to the prohibitive costs of monitoring and supervision. This, of course, does not imply that supervision of quality or effort cannot affect the level of performance. But in our opinion such monitoring and supervision can only have a significant impact if they contribute to increased knowledge and skills, or to changes in the underlying motivations of the health workers, which is the issue that we now turn to.

3.3 Professionalism and organisational identity

We have presented two different ways in which health workers may be motivated by factors beyond their narrow self-interest. These alternative motivations represent separate channels through which policy makers may seek to influence the health workers' provision of high quality health services.

There is a strong tradition for promoting professional standards of conduct in the health sector. By communicating these standards to the health workforce, health workers may adopt them as part of their own personal motivations, in the sense that their utility depends on the degree to which they are able to fulfill the professional standards. As professional attitudes in the health sector are very much geared towards providing services of adequate quality, a natural first guess is that a strengthening of health workers' professionalism will improve the quality of health services. We show, however, that this guess may be wrong.

In our model, there are two different ways of strengthening professional standards. First, one can increase the level of γ, which is the degree to which health workers experience a utility loss by deviating from the standard. Second, one may tighten the standard itself by raising the level of q^N.

Consider first the effect of increasing γ. It follows straightforwardly from Eq. (9) that in case of a health worker shortage, a higher γ will increase labour supply and the level of effort and quality if and only if

$$\varepsilon_{n,q} + 1 - \varepsilon_{n,q}\frac{q^N}{q} > 0$$

which also can be written as

$$\varepsilon_{n,q}(\frac{q^N}{q} - 1) < 1$$

It is not obvious that this inequality holds. If the demand response ($\varepsilon_{n,q}$) is large and the actual quality of health services (q) falls considerably short of

the professional standard (q^N), an increase in γ may in fact reduce the quality of health services. On the one hand, a higher γ implies that health workers have stronger incentives to improve the level of quality, but on the other hand, higher γ also implies a higher utility loss of treating an even larger number of patients with sub-standard quality. We believe this to be an important observation: Professionalism is not necessarily a quality-driving motivation if it is difficult for the workers to live up to the professional standard.

In our model, lifting the professional standard to higher levels would in itself have an unambiguously negative impact on quality. As seen from Eq. (9), a higher q^N will reduce the marginal benefits of increasing labour supply (and efforts). This is simply because lifting the professional standard only contributes to a larger utility loss per patient for those health workers who are concerned about reaching the standard.[5]

A strengthening of the health workers' organisational identity, modelled here as a concern for the aggregate health services (nq) provided, have an unambiguously positive effect on labour supply and on the level of effort/quality when there is a health worker shortage. This follows straightforwardly from the fact that a higher level of α increases the marginal benefit of labour supply (and efforts) (see Eq. (9)). This is not a surprising result, but it is nevertheless an important observation. First, it stands as an interesting contrast to the ambiguous effect on the quality of health care of strengthening the degree of professionalism. Second, it provides a useful background for interpreting patterns in empirical data on the level of effort and its correlates. For instance, Das and Hammer (2007) observe a positive correlation between effort and the level of provider skills. In their model, which only considers selfish health workers, this observation is interpreted as evidence that there are complementarities between skills and effort. This would also be a possible explanation within our model, but other interpretations also become visible once we include the possibility that health workers may have other-regarding preferences. For instance, a high degree of organisational identity (or altruism) might cause both a high level of effort and a high level of skill, as we would expect more altruistic people to work harder in the medical schools.

How can policy makers strengthen the organisational identity of health workers? There is no simple formulae they can apply, and a comprehensive

[5]If the utility loss of a professional health worker were a convex function of the difference between actual quality and the quality standard (e.g., $-n(q^N - q)^2$), there would also be an effect drawing towards higher q when q^N increases, as and increase in $(q^N - q)$ would increase the marginal utility gain of improving the level of quality.

answer is beyond the scope of this article. The literature dealing with these issues regularly refers to *respect*, *recognition* and *participation* as important motivators.[6] In order to enhance the employees' organisational identity or self-motivation, workers must be paid respect and not be used simply as means to further the interests of the owners. Self-motivation can also be fostered if owners and/or managers recognise and communicate the importance of the work that is done. Yet another way of strengthening organisational identity may be to include the workers in setting performance targets at their work site.

A distinctive feature of the health sector is that many workers enter with a strong motivation to serve the patients. The job for health sector managers is then to *maintain* this motivation, which probably also requires them to show respect, express interest in and recognise the work that is done, and invite workers to participate in the setting of performance standards. One aspect of paying respect to the workers is to provide a decent remuneration. This issue is particularly relevant in low income countries, where real wages of public sector workers in may places have been declining for years after years. It is also likely that providing health workers with adequate means to do their job can signify respect and recognition and thus cultivate their care for patients. In our model, this would imply that α - the weight that health workers assign to patient care - may be a positive function of the salary w and of the performance enabling factors x, at least over some range. Supervision may play a similar role, insofar as the supervision is performed in a supportive manner which recognises the importance of the work done and involves the health workers in defining performance standards.

4 Conclusion

This paper provides an analytical framework for assessing policies for increasing the quality of health services when health workers are in short supply. In our set-up, a health worker shortage implies that nurses and doctors work longer hours and spend less time per patient than they would have done with more health workers around. Using this constrained optimum as our baseline produces many interesting insights into how the quality of health care can be

[6] Consult Pfeffer (2007), Ellingsen and Johannesson (2007) and Akerlof and Kranton (2005) for a general discussion of organisational identity and worker performance.

improved, especially when we include the other-regarding motivations that often seem to drive health worker behaviour.

One possible extension of our work would be to move beyond the single (representative) health worker and study the interaction among several workers. One interesting issue is how the labour supply (and effort) chosen by one health worker would affect the labour supply of other health workers at the same facility. If one health worker reduces his labour supply at the clinic, there will be more patients for his colleagues. They will then have to increase their working time and/or reduce their time per patient in order to attend to all patients. The quality of the services is likely to fall. We have shown that if quality falls sufficiently below a certain standard, professional health workers may reduce their labour supply. This indicates the possibility of a dynamic process with negative feedbacks which can bring a clinic to an equilibrium with very low labour supply and low quality of health services.

Another possible extension is to include the choice of entering the health workforce. Our model provides an analytical grip on the intuitive idea that when there is a shortage of health workers, there are positive externalities in the recruitment of additional workers. In our analysis, the shortage of health workers makes the existing workers choose a different labour supply and work effort than in the unconstrained case. Hence, their utility from work would increase if they had fewer patients, as would be the case if more health workers were employed. This positive externality implies that policy makers ought to follow a "big push" strategy in order to fill vacancies in the health sector.

References

[1] Akerlof, G. and R. Kranton, 2005. Identity and the Economics of Organizations. *Journal of Economic Perspectives* 19(1): 9-32.

[2] Banerjee, A and E. Duflo, 2006. Addressing Absence. *Journal of Economic Perspectives* 20(1): 117-132.

[3] Barber, S.L., P.J. Gertler, and P. Harimurti, 2007. The Contribution of Human Resources for Health to the Quality of Care in Indonesia. *Health Affairs* 26(3): w367-w379.

[4] Barnard, C., 1938. *The Functions of the Executive.* Cambridge, Mass: Harvard University Press.

[5] Chaudhury, N., J. S. Hammer, M. Kremer, K. Muralidharan, and F.H. Rogers, 2006. Missing in Action: Teacher and Health Worker Absence in Developing Countries. *Journal of Economic Perspectives* 20(1): 91-116.

[6] Das, J., and J. Hammer, 2007. Money for nothing: The Dire Straits of Medical Practice in Delhi, India. *Journal of Development Economics* 83: 1-36.

[7] Fehr, E., and K. Schmidt, 2006. The Economics of Fairness, Reciprocity and Altruism: Experimental Evidence. In: S.C. Kolm and J.M. Ythier (eds.), *Handbook on the Economics of Giving, Reciprocity and Altruism.* Elsevier.

[8] Ellingsen T. and M Johannesson, (2007). Paying respect. *Journal of Economic Perspectives* 21(4)135–149

[9] Gaynor. M., J.B. Rebitzer, and L.J. Taylor, 2004. Physician Incentives in Health Maintenance Organizations. *The Journal of Political Economy* 112(4): 915-931.

[10] Le Grand, J., 2003. *Motivation, Agency and Public Policy: Of Knights & Knaves, Pawns & Queens.* Oxford: Oxford University Press.

[11] Leonard, K.L., M.C. Masatu, and A. Vialou, 2007. Getting Doctors to Do Their Best: the Roles of Ability and Motivation in Health Care Quality. *Journal of Human Resources* 42(3): 682-700.

[12] Ma, C.A., 1994. Health Care Payment Systems: Cost and Quality Incentives. *Journal of Economics & Management Strategy* 3: 93-112.

[13] Ma, C.A., T.G. McGuire, 1997. Optimal Health Insurance and Provider Payment. *American Economic Review* 87: 685-704.

[14] Manongi, R.N., T.C. Marchant, and I.C. Bygbjerg, 2006. Improving motivation among primary health care workers in Tanzania: a health worker perspective. *Human Resources for Health* 4:6.

[15] McGuire, T.G., 2000. Physician Agency. In: Culyer, A.J., and J.P. Newhouse (eds.), *Handbook of Health Economics.* Elsevier.

[16] Meessen, B., L. Musango, J.-P.I. Kashala, and J. Lemlin, 2006. Reviewing Institutions of Rural Health Centres: the Performance Initiative in Butare, Rwanda. *Tropical Medicine and International Health* 11(8): 1303–1317.

[17] NORAD, 2007. *The Global Campaign for the Health Millennium Development Goals.* NORAD Report. http://www.norad.no/default.asp?V_ITEM_ID=9263. Accessed 26 March 2008.

[18] Pfeffer, J., 2007. Human Rescources from an Organizational Behaviour Perspective; Some Paradoxes Explained. *Journal of Economic Perspectives* 21(4) 115 - 134.

[19] Rotemberg, H., 2006. Altruism, Reciprocity and Cooperation at the Workplace. In: S.C. Kolm and J.M. Ythier (eds.), *Handbook on the Economics of Giving, Reciprocity and Altruism.* Elsevier.

[20] Rowe, A.K., D. de Savigny, C.F. Lanata, and C.G. Victora, 2005. How can we achieve and maintain high-quality performance of health workers in low-resource settings? *The Lancet* 366(9490): 1026-1035.

[21] Selznick, P., 1957. *Leadership in Administration.* New York: Harper & Row.

[22] Wedig, G., J.B. Mitchell, and J. Cromwell, 1989. Can Optimal Physician Behaviour be Obtained Using Price Controls? *Journal of Health Policy, Politics and Law* 14(3):601-20.

[23] Woodward, R.S., and F. Warren-Boulton, 1984. Considering the effects of financial incentives and professional ethics on 'appropriate' medical care. *Journal of Health Economics* 3(3): 223-237.

[24] WHO, 2006. *Working Together for Health. World Health Report 2006.* Geneva: World Health Organisation.

Department of Economics
University of Bergen
Fosswinckels gate 6
N-5007 Bergen, Norway
Phone: +47 55 58 92 00
Telefax: +47 55 58 92 10
http://www.svf.uib.no/econ