Physician dual practice: a welfare-improving mechanism to match patients with appropriate treatment methods

Jiwei Qian
Physician dual practice: a welfare-improving mechanism to match patients with appropriate treatment methods

Jiwei Qian¹

Abstract

In the literature, physician dual practice is argued to be a self-selecting mechanism for physicians. High skilled or selfish hospital physicians are more likely to have private practice. For developing countries, dual practice is also considered as an informal way to compensate public hospital physicians whose salary is low. However, dual practice remains a controversial issue since physicians may take advantage of hospital resources for their own interests. In this paper, we consider the question whether dual practice should be allowed in the context of the policy objective that patients should receive their care in the treatment setting that is most efficient. Hence, the focus of the paper is on patient heterogeneity and differences in the cost of care in a hospital and an outpatient setting, depending on the nature of the patient's illness. In an environment in which there is information asymmetry between doctors, on the one hand, and patients and local health authority on the other hand, we analyze equilibria in which doctors have an incentive to exploit their information advantage over patients and health authority to increase their net income. We show that, in general, such equilibria will not be economically efficient in the sense that doctors will not have an incentive to refer patients to the most efficient place of treatment. However we find allowing dual practice can lead to a second-best improvement in efficiency. Dual practice can be interpreted as an alternative mechanism for sorting patients in terms of both illness severity and switching costs.

Keywords: Dual practice, Physicians’ incentives, Information asymmetry, Public-private

JEL Classification: I12, I18, L15

¹ Email: jiwei.qian@nus.edu.sg
Introduction

A central issue in much of the health economics literature is that of information asymmetry. Much of the literature on this issue focuses on the interaction between doctors and patients in a primary-care setting in which the doctor is paid via fee-for-service and hence is likely to have an incentive to provide (and charge for) a large volume of services to each patient. As a result of the information asymmetry between patients and doctors, patients typically have to rely on their doctors' advice in deciding on the type of services to purchase. In this situation, the result of the information asymmetry may be supplier-induced demand, i.e., an outcome in which patients utilize, and pay for, a larger volume of services than they would have done if they had been better informed.

Doctor-patient information asymmetry can also lead to supplier-induced demand on the part of specialist doctors who supply secondary and tertiary care, and who may treat inpatients as well as outpatients. In many countries, however, supplier-induced demand is not a relevant issue in the case of hospital-based doctors, because such doctors are paid on the basis of a fixed salary and hence have no incentive to induce patients to utilize more services than they would if they were fully informed. However, in the case of hospital-based salaried doctors another type of information asymmetry may be relevant, namely that between the doctors who provide services to patients and the hospital managers who employ them, and whose financial results or promotion prospects may depend on the type and volume of services the hospital produces. Although hospital managers may be knowledgeable about medical conditions and the technologies available for treating patients, the nature of health care can make it difficult for managers to closely monitor the illness condition of individual patients, and the nature of the treatment that doctors provide. The result of this type of information asymmetry may be that doctors shirk (that is, spend less time and effort in treating patients than they would if they had a stronger incentive to provide more and higher-quality care).

In many countries where salaried doctors treat patients in government hospitals, patients also have the option of seeking care in private hospitals or outpatient clinics instead. When privately provided care is supplied on the basis of fee for service, doctors who practice privately have a stronger incentive to provide care of high quality than salaried doctors in government hospitals, since their reputation for
providing high-quality care is likely to influence the future demand for their services and hence increase their expected net income. In contrast, the income of salaried doctors in government hospitals does not depend directly on the volume of services they supply, and hence is less dependent on an individual doctor's reputation for providing high-quality care. In such cases, a tendency may arise for the best doctors to gravitate toward private practice, leaving less experienced or talented doctors in the government hospitals.

One way of overcoming this problem is to allow “dual practice”, that is, to allow salaried doctors in government hospitals to also practice privately on a part-time basis. In some countries, specialist doctors in dual practice may actually treat their patients in designated beds in the government hospitals, although a more common arrangement is for them to treat private patients in their own outpatient clinics. By allowing dual practice, hospital managers hope to be able to retain experienced doctors capable of providing high-quality service in government hospitals so as to avoid the tendency for such doctors to leave the hospitals for full-time private practice.

Arrangements to allow dual practice are often controversial. One reason is that additional monitoring may be required to ensure that salaried part-time employees actually fulfill their contracted working hours in the government hospital (and do not spend some of the time that they are supposed to be in the hospital in private practice instead). In India, for example, dual practice is one of the major reasons for a large scale of absence of health workers in public hospitals (Singh 2008, Yip and Mahal 2008). This may be particularly relevant in cases when doctors in dual practice can send patients who have been diagnosed in the hospital to their private clinics for treatment. Another reason why dual practice is controversial is that it often exists in situations where there is an excess demand and hence lengthy waiting lists for subsidized treatment in government hospitals, and private practice offers persons with high income an opportunity to bypass the waiting lists by paying for unsubsidized private care. This outcome is regarded as undesirable in countries such as Canada where there is an ideological commitment to the principle that all citizens should receive health care on similar terms, regardless of their income. At the same time, however, dual practice can be advantageous because it has the indirect effect of shortening the waiting lists for subsidized care, or because it raises the average quality of the
care provided in government hospitals as it reduces the incentives for skilled practitioners to switch to full-time private practice.

In this paper, we consider the question of whether dual practice should be allowed in the context of another policy objective in the health care system: patients should receive their care in the treatment setting that is efficient for them. Thus the focus in the paper is on patient heterogeneity and differences in the cost of care in a hospital and an outpatient setting, depending on the patient’s illness condition. In an environment in which there is information asymmetry between doctors and patients on the one hand, and doctors and hospital managers/local health authority on the other hand, we analyze equilibria in which doctors have an incentive to exploit their information advantage over patients and managers to increase their net income (utility). We show that in general, such equilibria will not be economically efficient in the sense that doctors will not have an incentive to refer patients to the most economically efficient place of treatment. However we find that under some conditions, allowing dual practice can lead to a second-best improvement in efficiency.

**Literature Review**

Wolinsky (1993) studies different equilibriums for informed experts' services when buyers can search for the services they need. A general survey on interaction between informal and formal activities in health sector is provided by Ensor and Thompson (2006). Barros and Siciliani (2011) is an excellent survey on the interaction between public and private sector in the context of healthcare.

Socha and Bech (2011) and Eggleston and Bir (2006) are very good surveys on dual practice literature. García-Prado and González (2007) review the variety of policies which government can make responding to physician dual practice. Jan, et al., (2005) is a survey on various regulations of physician dual practices and welfare impacts in several developing countries.
Assuming heterogeneous physicians, Bir and Eggleston (2003) argue that higher skilled physicians are more likely to stay in public hospitals when they can have additional income from private practice. Then, social welfare sometimes could be improved if dual practice is allowed. Biglaiser and Ma (2007) evaluate welfare effects with heterogeneous physician's morality: moonlighters who shirk and provide poor quality treatment in hospital and non-moonlighters who unselfishly provide good quality service in hospital. In their paper, since quality in private sector is assumed to be audited by regulators more effectively, social welfare will always be improved with dual practice since moonlighters will provide higher quality service in private sector.

Assuming that the reputation from practicing in public sector has effects on the demand for the services of a given physician in the private sector, González (2004) argues that dual practice may improve social welfare since physicians have incentives to provide quality services in hospitals to enhance their reputation, given a carefully designed set of hospital regulations. In another paper, González (2005) discusses the negative effects of cream skimming behavior in dual practice when physicians follow a policy of transferring patients from hospital to their private clinics. Iverson (1997) constructs a model in which waiting time in a hospital would be longer after allowing physician dual practice. In the model built in González and Macho-Stadler (2013), doctors have different levels of ability and a health authority chooses the optimal regulation for dual practice. González and Macho-Stadler (2013) conclude that dual practice is useful to keep more skilled doctors in the public sector.

Our paper is different from other papers by focusing on how physician dual practice can be useful to match treatment technologies and different types of illnesses under an environment where information is asymmetric, when physicians have homogeneous quality/morality. The context of our model is close to Blomqvist and Leger (2005), who considered the question when physicians and patients can decide where patients will be treated (i.e., in a hospital or a clinic), when information is asymmetric. In addition, the configuration of our model is close to Biglaiser and Ma (2007). They considered the situation in which the private sector has a better audit technology on service quality than the public sector, while our paper assumes that the public hospital has better technology to treat patients compared with the private sector.
The rest of the paper is organized as follows. The next section introduces the environment of our model. A benchmark equilibrium with full information is shown afterwards. The equilibriums without dual practice, when information is perfect or not, are discussed in the following two sections. Then, we analyze the dual practice’s welfare effect before we conclude the paper.

**Environment**

There are three kinds of agents in this model: patients, physicians and a local planner. Let $N$ be the number of patients who need treatment in each time period. For these patients, there are two sorts of diseases: serious or mild. The probability to have mild disease is $\nu$ and the probability of serious disease is $1-\nu$. Patients do not know the severity of illness before they are diagnosed by a doctor.

Patients with serious illness may want to go to another doctor for treatment or a second opinion. If they do so, they incur a cost. The costs for switching providers consist of two parts. The first part is the costs for re-diagnosis. We assume that, for a patient with serious disease, doctors will insist on doing their own diagnosis to know more about the patient's characteristics before treating the patient. We believe this is a reasonable assumption for many types of specialized medical treatment, such as an operation. The second part of the cost is the (subjective) opportunity cost of patients' time, effort, and worry when treatment is delayed as a patient with serious disease changes his service provider. Patients with different income and preferences have different switching costs. For example, switching costs for patients with higher income may be higher since opportunity cost for time spent shopping around for providers is higher. This subjective switching cost for patients with serious disease is distributed uniformly and independently between $[0,\tau]$. An important asymmetry that we impose on the model is that for mild cases, going to a second provider will not entail a switching cost. Also, mild cases need only to be diagnosed once.
Total costs for patients include treatment costs, diagnosis costs and expected subjective switching costs. Patients cannot infer what their illness severity was even after receiving treatment. However, patients can observe whether the treatment outcome, as measured by a health status variable $\bar{h}$, is lower than a threshold value $\bar{h}$. Patients' utility can be defined as:

$$\begin{cases} 
-\infty, & \text{if treatment outcome is less than } \bar{h}; \\
W, & \text{if treatment outcome is no less than } \bar{h}. \end{cases}$$

$W$ is positive and denotes income less treatment and diagnosis fees, and potential subjective switching costs.

Health service market is perfectly competitive with an infinitely elastic supply of physicians. A physician's utility depends only on his income and time inputs in treating patients. A physician can work in either the public hospital or a private clinic. A physician in a clinic is paid by patients on the basis of fee-for-service while a hospital physician works as a salaried employee. Physicians' utility function $u(Y, T)$ is in a quasi-linear form as follows:

$$u(Y, T) = Y - \varphi(T)$$

Where $Y$ is income for a physician, $T$ is the aggregate amount of time spent on treating patients and $\varphi(T)$ is the disutility from the number of working hours. $\varphi(T)$ is a strictly convex function increasing on $T$ and $\varphi(0) = 0$. Physicians' reservation utility is $\overline{u}$.

Treatment outcome is jointly produced by a non-physician input (i.e., perhaps some kind of specialized equipment, or drugs) and the amount of time the physician takes to treat the patient. Clinic physicians choose both input levels and the number of patients. Physicians use two inputs to produce treatment for a given patient: working hours $t$ and non-physician inputs $x$. The following functions $g$ and $b$ are production functions for the treatment of mild and serious diseases in a clinic:

$$g(x_{cg}, t_{cg}), b(x_{cb}, t_{cb}).$$

Subscripts $g$ and $b$ refer to mild and serious cases. Subscripts $c$ denotes clinic.
The capacity of the hospital (denoted by $H$) is the number of patients the hospital can treat per unit of time. $H$ is less than the expected number of serious cases in the population (i.e., $(1 - \nu)N > H$). We assume that non-physician inputs are more productive when used in a hospital. This assumption may reflect, for example, superior technology in the use of some kinds of specialized equipment that is present in the hospital, but which doctors in private clinics cannot afford to acquire. Coefficient $\sigma > 1$ implies that non-physician inputs are utilized more efficiently in the hospital. The following functions $g$ and $b$ are production functions in the hospital: $g(x_{hg}, t_{hg}), b(\sigma x_{hb}, t_{hb})$ where $\sigma > 1$ and subscripts $h$ denotes hospital.

Treatment outcome is not verifiable in general. The only thing can be verified is that whether treatment outcome is higher or lower than threshold value $\bar{h}$. Treatment outcome level is effectively translated to two domains: worse than threshold value or not. A physician will be punished if the treatment outcome results in the patient attaining a health status $h$ that is less than $\bar{h}$. The punishment if the actual outcome of a treatment episode is less than $\bar{h}$ is so high that no doctor will find it is optimal to undertreat a patient. Also, because it is never optimal for a doctor to treat a patient so that the outcome $h$ is less than the critical value, the inputs in clinics and the hospital are effectively constrained by the conditions $g(x_{cg}, t_{cg}) = b(x_{cb}, t_{cb}) = g(x_{hg}, t_{hg}) = b(\sigma x_{hb}, t_{hb}) = \bar{h}$.

When dual practice is not allowed, hospital physicians work full timely in the hospital. If dual practice is allowed, hospital physicians can practice part time in private clinics. We assume that there is no switching cost if patients (both serious and mild cases), diagnosed in the hospital, are treated by dual practicing physicians in in their own private clinics, since patients in this case are treated by the same physician who made the diagnosis in the first place.

Decisions for the public hospital are made by a local planner. The local planner chooses the non-physician input level for hospital patients, the number of physicians to be hired in the hospital and the
salary for physicians. The planner also sets prices in the hospital. The local planner's objective is to maximize social welfare. Given the quality of treatment out is fixed at the level $\bar{h}$, local planner's objective is simply to minimize costs. The expenditure of hospital physicians' salary, non-physician input in the hospital and the fees paid to private clinic physicians are both costs for the local planner. The fees paid to private clinic physicians is considered as social costs since physicians in the hospital or clinics are considered by the planner as outsiders whose utilities do not enter into social welfare function. Anyway, a doctor can always reach the reservation level of utility at equilibrium: if a doctor is not hired by local hospital or works in a local clinic, he can migrate somewhere else to get his reservation utility. The treatment and diagnosis fees paid at the hospital are not included into the social cost since they are interpreted as a transfer.

It is always welfare improving by treating more patients suffering from serious disease in the hospital since the hospital has superior technology. Also, given that hospital capacity is not large enough to treat all patients with serious disease, it is optimal then for the planner to decide to only treat patients with serious diseases in the hospital. Patients diagnosed as mild cases in the hospital will be referred to clinics and it will not reduce welfare since no switching cost will incur for these patients.

In the imperfect-information version of the model, the planner cannot verify physicians' time inputs in the hospital and that patients cannot observe physician time inputs, or the input of non-physician services, in either the hospital or in private clinics.

The timing of the model is as follows: nature decides the types of diseases. Then patients seek for diagnosis in the hospital or a clinic. Patients diagnosed as mild cases in the hospital choose clinics for treatment. Other patients, after learning result of diagnosis report, can decide to continue treatment by the same doctor who conduct the diagnosis or switch to another doctor. A doctor then treats patients. When dual practice is allowed, a dual practicing doctor will decide whether to refer a hospital patient to his own clinic or not.
The benchmark model with full information

When information is perfect, patients can tell not only whether treatment outcomes are less than $\bar{h}$ or not, but they can also observe non-physician inputs and physician's time inputs in private clinics. However, patients do not know their illness severity before diagnosis.

If the physician could mislead the patient, he would have an incentive to tell people with a mild form of the disease that they have the serious form. If he did this, he could supply the smaller input quantities required for the mild case, but charge the patients as if they had received treatment for the serious case. However, patients can indirectly infer the severity of their illness from the observed input level if there is full information. This rules out the possibility that a physician can supply a patient with the inputs necessary to treat the mild form of illness while claiming that the patient suffers from the serious form, and charging the patient the (higher) fee corresponding to treatment of the severe illness form.

In this model, there are three types of market-determined fees: the fee charged for diagnosis alone $p_d$, the fees charged for treating patients with the mild $p_{cg}$ and serious $p_{cb}$ types of the disease. When information is perfect, we can show that there are only three strategies that can exist in equilibrium. The first is that a doctor announces that he will diagnose and treat both types. The second is that a doctor announces that he will diagnose all patients, but only treat those with the severe disease. In an
equilibrium when this happens, there must also be some doctors who are willing to treat patients with mild cases that have been diagnosed by another doctor.

With given market prices, a clinic physician can choose total working hours under any strategy, by choosing the amount of time he spends treating each patient, and choosing the number of patients he will diagnose and treat. Note that physicians choose number of patients but not the case mix in his clinic. Lemma 1 below shows that there is a one-to-one relationship between the doctor's return per unit of time supplied and his utility.

**Lemma 1:** We have following market equilibrium conditions:

\[
\frac{\partial g}{\partial x_g} \bigg|_{g=h} = \frac{\partial b}{\partial x_b} \bigg|_{b=h} = \frac{1}{\varphi'(T^*)}
\]  

\[p_{ci} = x_{ci} + t_{ci} \varphi'(T^*), \quad i = g, b\]  

\[p_d = \bar{d}\varphi'(T^*)\]  

\[T^*\varphi'(T^*) - \varphi(T^*) = \bar{u}\]  

Proof: see appendix.

Note that \(\frac{\partial g}{\partial x_g} \bigg|_{g=h}\) and \(\frac{\partial b}{\partial x_b} \bigg|_{b=h}\) in (1) denote marginal rate of technical substitution when treatment outcome is \(\bar{h}\). This marginal rate of technical substitution is equal to the relative price between non-physician and physician input at the margin (i.e. \(\frac{1}{\varphi'(T^*)}\) in (1)). (2) and (3) show that prices are equal to expected marginal costs for treatment and diagnosis in equilibrium. (4) is the market clearing condition.

With full information, the hospital only treats serious cases. The hospital receives enough patients so that, given the probabilities, it only treats \(H\) serious cases and refers all the mild cases to clinics. No patient needs to pay switching costs since each patient only has to be diagnosed once and no patients with serious disease need to switch to a second provider.
The local planner’s problem is to minimize social costs, which include all fees paid to private sector plus non-physician input in public sector and expenditure on physician’s salary. The local planner’s optimization problem is:

$$\min_{x_{hb}, t_{hb}, q, s} \{ Hx_{hb} + sq + vNp_{cg} + [(1 - v)N - h]p_{cb} + (N - \frac{H}{1-v})p_{d} \}$$  \hspace{1cm} (5)$$

where $q$ is the number of hospital physicians. First item of (5) $Hx_{hb}$ denotes non-physician inputs for $H$ hospital patients and the second item $sq$ refers to the total salary expenditure for the hospital. The third and fourth items $[(1 - v)N - h]p_{cb} + \left(N - \frac{H}{1-v}\right)p_{d}$ denote the treatment and diagnosis fee paid to private clinics. The salary constraint for a hospital physician is: $s = \varphi(T^*) + \bar{u}$.

Effectively, the local planner chooses the level of non-physician input in the hospital, the number of hospital physicians to be hired, and the salary they will be offered. In this equilibrium, the hospital treats exactly $H$ patients. Since health service market is competitive, price for serious cases must be equal between a clinic and the hospital at equilibrium (i.e. $p_{hb} = p_{cb}$). Otherwise, there will be too many or too few patients treated in the hospital. Also, treatment outcomes in the hospital is $\bar{h}$.

Social planner’s problem can be reduced to finding the optimal level of $x_{hb}$ and $t_{hb}$. Differentiating the objective function with respect to $x_{hb}$, we obtain:

$$\frac{\partial b}{\partial x_{hb}} |_{b = \bar{n}} = - \frac{1}{\varphi(T^*)}$$  \hspace{1cm} (6)$$

Both non-physician and physician inputs are determined by (6), given the equilibrium condition for treatment outcomes. Compared with (1), the right hand side of (6) is smaller, which means that, for a patient with serious disease, physicians’ time input is less in the hospital than in a clinic, which reflects the physician input is relatively more expensive in the hospital given the superior technology in utilizing the non-physician input in the hospital.
Whether to allow dual practice or not can be shown to be not relevant when information is complete. A physician has no incentive to refer a hospital patient, who has serious disease, to his own clinic. With competitive market and full information, there is no additional profit by treating this patient. Every player's (i.e., physician, planner and patients) action does not change after allowing dual practice and case mix in the hospital does not change also.

**Imperfect information without dual practice**

Now, consider the case where information is imperfect and dual practice is disallowed. The planner can no longer verify physicians' time inputs in the hospital. Patients cannot observe what inputs they receive, either of non-physician or of physician services in clinics and the hospital. The planner as well as patients can only observe whether the treatment outcome is lower or higher than a minimal value $\bar{h}$. Thus in this case, it is possible for a physician to misrepresent the patient's medical condition.

Since the hospital planner cannot observe the actual time input of hospital physicians, the salary offered by the hospital to a physician can be set on the basis of observable criteria: the number of patients being treated and diagnosed in the hospital by each physician. Since information is asymmetric between patients and physicians, a physician can take advantage of his superior information. By misrepresenting a mild case as a serious case, a clinic physician can charge a higher service while a hospital physician can save his time input since treating a mild case takes less time compared with a serious case. In both clinics and the hospital, it is to the doctor's advantage to diagnose and treat patients as serious cases, even if they are mild cases.

**Pure Fraud Equilibrium**

A pure fraud equilibrium may emerge, which is referred to an equilibrium where every mild case is treated as a serious case. Optimal input condition (1) still holds and the price in a clinic is:
\[ p_{cb} = vx_{cg} + (1 - v)x_{cb} + [vt_{cg} + (1 - v)t_{cb}]\phi'(T^*) \]  \tag{7}

(7) shows that equilibrium price in clinics is equal to expected marginal cost for treatment. The diagnosis cost is also equal to the marginal cost and determined by (3). Under this equilibrium, no patient will switch to a second physician. This is because patients anticipate that every physician is going to misrepresent patients with mild diseases. The private sector is efficient, given that price is equal to marginal cost and no patient switches.

Hospital physicians classify all patients as severe cases at the equilibrium. The hospital treats exactly \( H \) patients. Since hospital physicians classify all patients as serious cases, only \( H \) patients are admitted, and provided with the same level of non-physician input \( x_{hb} \), even though some of them are mild cases. Under this equilibrium, no patient switches to a second physician. This is because patients anticipate that every physician misrepresents mild cases. We can show that the number of working hours is the same for a hospital and a clinic physician (i.e. \( T^* \)).

Under the pure fraud equilibrium, every patient is diagnosed as a serious case. There is inefficiency in this equilibrium because the hospital treats some mild cases. Since the planner cannot observe the types of individual patients, the planner has to prescribe the same amount of \( x_{hb} \) for all cases. However, this amount is chosen knowing that some patients are misrepresented. The planner’s problem is:

\[
\min_{x_{hb}, q} \{Hx_{hb} + sq + (N - H)(p_{cb} + p_d)\}
\]

The third item \((N - H)(p_{cb} + p_d)\) denotes the fee paid to private clinics (i.e., every patient is treated and diagnosed as a serious case). Optimal inputs in the hospital with respect to \( x_{hb} \) is:

\[
(1 - v) \frac{\partial b}{\partial t_{hb}} \bigg|_{b = \bar{b}} + v \frac{\partial g}{\partial t_{hg}} \bigg|_{b = \bar{b}} = -\frac{1}{\phi'(T^*)} \tag{8}
\]

Where \( t_{hg} \) is defined as the physician input in the hospital to treat a mild case, subject to \( g(x_{hb}, t_{hg}) = \bar{t} \). Compared with (6), there is less capital input per patient in the hospital in (8). Intuitively, the planner can infer that there are some mild cases being treated in the hospital. Hospital physicians will
work less time if planner implements the capital input level in the case of full information. In this fraud equilibrium, the planner can save more social costs by reducing non-physician inputs, compared to the full information case.

**Specialization Equilibrium**

In the pure fraud equilibrium, patients who suffer from the mild version of the disease are worse off than they would be in the full information equilibrium, since they are paying the higher (common) price for treatment. Thus patients who suffer from the mild version of the disease would be better off if they could somehow establish this fact, and then obtain (and pay for) the lower-cost treatment necessary for the mild disease. It would in fact be possible for them to do so if some doctors commit in advance to a strategy of only treating patients with the mild version of the disease, while those diagnosed with the severe version would be referred to doctors who were willing to treat both versions of the disease (or the severe version only). Below we refer to doctors who follow a strategy of only treating patients with the mild version of the disease as General Practitioners (GPs).

However, a patient strategy of going to a GP (rather than to a non-specialized clinic) is not cost-free. If the patient turns out to have a severe condition, he has to consult another doctor. He will therefore incur the subjective switching cost, and will also have to pay for an additional diagnosis (recall our assumption that when a patient with the severe form of the illness switches to another provider, he has to undergo a second diagnosis).

A specialization equilibrium refers to an equilibrium where some physicians (i.e. GP) can commit to treating mild diseases only in their clinics and patients choose where to be diagnosed/treated on the basis of expected costs. For a patient whose switching cost is lower than a critical value, he first visits a GP committed to treating mild cases only. If the patient is diagnosed as a serious case, he will switch to another provider (non-GP). For a patient whose switching cost is higher than the critical value mentioned above, he visits a non-GP provider.
Whether a patient goes to a GP or not depends on the value of switching cost. The critical value of the switching cost is $m\tau$, where $m \in [0,1]$; a patient with this switching cost is indifferent between visiting a GP or a non-specialized doctor. $m$ is defined by

$$p_{cb} = v p_{cg} + (1 - v)(p_{cb} + m\tau + p_d)$$  \hspace{0.5cm} (9)$$

The left-hand-side of (9) is the expected cost for patients who choose to visit a non-GP physician and the right-hand-side is the expected cost for patients to have their first visit to a GP. $(1 - v)(p_{cb} + m\tau + p_d)$ denotes expected costs when a patient has to seek for a second provider after being diagnosed as a serious case. For patients whose switching cost lies between $(m\tau, \tau]$, they visit a non-GP first while for those patients whose switching cost lies between $[0, m\tau)$, they visit a GP first instead.

We may interpret $m$ as a degree of specialization for health service markets. The higher $m$ is, the more people are diagnosed by a GP physician. When $m = 0$, no patient visits a GP in the first place. In this case, every patient is diagnosed as a serious case (i.e., pure fraud equilibrium). When $m = 1$, at the other extreme, every patient visits a GP for diagnosis.

A specialization equilibrium will emerge if physicians who commit to treating mild cases only can attain at least the reservation utility. Formally, 

**Lemma 2**: Given $p_{cb}$ defined by (7), if $\exists p_{cg} < p_{cb}$, a specialization equilibrium will emerge if we have

$$(p_{cg} - x_{cg})vm_{cg}N + p_d m_{cg}N - \varphi (vm_{cg}N t_{cg} + m_{cg}N d) > \bar{u},$$

Where $m_{cg} = \min (\bar{m}_{cg}, \frac{v(p_{cb} - p_{cg})}{(1-v)\tau} - \frac{p_d}{\tau})$ and $v p_{cg} + p_d = v x_{cg} + (v t_{cg} + d) \varphi (v m_{cg}N t_{cg} + \bar{m}_{cg}N d)$. $m_{cg}$ denotes portions of patients who are treated by a GP. $x_{cg}, t_{cg}$ are defined by optimal input condition (1).

Proof: See Appendix.
The proportion of mild cases in the hospital as well as a non-GP clinic \( v^s \) is:

\[
v^s = \frac{(1-m)v}{(1-m)v+1-v} = \frac{(1-m)v}{1-mv}
\]  

(10)

The numerator on (10) \((1-m)v\) denotes the proportion of mild cases in the hospital and non-GP clinics. The denominator \(1 - mv\) denotes the proportion of patients visiting the hospital and non-GP clinics since \(mv\) patients go to a GP clinic. If \(m \in (0,1)\), \(\frac{v^s}{v} = \frac{1-m}{1-mv} < 1\) and \(v^s < v\). There is a lower proportion of mild cases treated in the hospital compared with the pure fraud equilibrium if \(m \in (0,1]\). If \(m = 0\), \(v^s = v\).

At equilibrium, we can show that the number of working hours for a non-GP clinic is the same as that for a GP (i.e., \(T^*\) in (4)). It is straightforward to see that the optimal input condition in a GP clinic is the same as full information case (i.e. equation (1)). For a GP and non-GP physician, treatment prices are equal to marginal costs and they get their reservation utility. Treatment fee in a GP clinic is subject to (2) and treatment fee for a serious case in a non-GP clinic is set by (11):

\[
p_{cb} = v^s x_{cg} + (1 - v^s)x_{cb} + [(v^s t_{cg} + (1 - v^s)t_{cb})\phi'(T^*)]
\]  

(11)

**Proposition 1:** there exists a stable fraud equilibrium. Also, there may exist multiple stable specialization equilibrium.

**Proof:** see appendix.

From proposition 1 fraud equilibrium can be stable, depending on parameter values. Also, there may be more than one specialization equilibrium even with the same set of parameter values. Intuitively, the proposition 1 implies that which equilibrium will emerge may depend on how the equilibrium was initially established for local health service markets (i.e., path dependence on the status quo).

The planner’s optimization problem is as follows:

\[
\min_{x_{hh}, q} \{\Gamma + Hx_{hh} + sq + vmNp_{cg} + (N - H)p_d + \left[(1 - m) + (1 - v)m - \frac{u}{N}\right]Np_{cb}\}
\]  

(12)
Where \( \Gamma = m v N \left[ \frac{(1-v)m \tau}{2} + (1-v)p_d \right] \)

\( \Gamma \) indicates aggregate switching costs. There are \( m v N \) patients who visit a GP first and for these patients, expected subjective switching cost is \( \frac{(1-v)m \tau}{2} \) while \( (1-v)p_d \) indicates the expected amount of diagnosed costs. The last item of (12) \( (1 - m) + (1-v)m - \frac{H}{N} \) \( N p_{cb} \) denotes the fee received by clinics for treating patients diagnosed as serious cases.

Physician’s expected time input for a patient in the hospital is \( v^s t_{bh} + (1-v^s) t_{hb} + \bar{d} \) since every patient diagnosed in the hospital does not switch to a second provider. The planner then can infer the number of physicians’ working hours from the expected time input for a patient and the number of patients admitted. We can show that a hospital physician' total number of working hours is \( T^* \), which is equal to the number of working hours with full information.

Optimal input condition in the hospital is as follows:

\[
(1 - v^s) \left. \frac{\partial b}{\partial g} \right|_{g = \bar{g}} + v^s \left. \frac{\partial g}{\partial g} \right|_{g = \bar{g}} = -\frac{1}{\sigma \varphi(T^*)}
\]

From (13), the amount of capital input per patient in the hospital is smaller compared with full information case (6). However, the amount of capital input per patient in the hospital is larger than in (8). This is because a larger proportion of serious cases are being treated in the hospital, compared to the case of fraud equilibrium.

The specialization equilibrium is inefficient since there are some mild cases treated in the hospital. In the private sector, there are additional switching costs compared to fraud equilibrium. Under a specialization equilibrium, patients with low switching costs are willing to search for a second provider.

**Proposition 2:** There is welfare loss under specialization equilibrium compared with welfare under full information case.
**Proof:** see Appendix.

Under pure fraud equilibrium, there is no inefficiency in the form of switching costs, but there is inefficiency in the sense that some mild cases are be treated in hospital, and the use of non-physician input in the hospital is not be set at its optimal level. Under the specialization equilibrium, there is be inefficiency in the form of switching costs, but because many mild cases are treated by GPs, there are a smaller proportion of mild cases seeking treatment in the hospital. The latter effect indirectly contributes to more efficient utilization of the hospital’s capacity. Thus, under specialization equilibrium, the hospital can take better advantage of superior technology.

**Imperfect information with dual practice**

Suppose now dual practice is allowed. Patients who are treated in dual practice clinics do not need to pay a switching cost since they are referred to the dual practice clinics by the same hospital physician. Patients cannot observe inputs from a dual practicing physician. Patients are indifferent to being treated in the hospital or in a dual practicing clinic as long as there are no price differences. After allowing physician dual practice, a hospital physician has incentives to treat hospital patients with mild diseases in his own private clinic. Some hospital patients, who are diagnosed in the hospital as serious cases, are referred to be treated in a dual practice clinic.

There are eventually less mild cases treated in the hospital compared with the regime disallowing dual practice. This gives us a central result of this paper: a planner can achieve second best welfare improvement by allowing dual practice. This is true for a pure fraud equilibrium or a (mixed) specialization equilibrium.

**Pure Specialization equilibrium with dual practice**

Consider first the pure specialization equilibrium. Everybody visits a GP first (i.e., $m = 1$). Allowing physician dual practice will not change the case mix in the hospital since there are only serious cases in
the hospital. The optimal input condition in the hospital is the same as first best. The social welfare is strictly less than the first best for this case since every patient diagnosed as the serious case has to switch providers (whether they go to a non-specialized clinic or a hospital). In comparison with full information case, the welfare loss comes from additional switching cost: \((1 - \nu)\left(\frac{r}{2} + p_d\right)N\).

**Pure fraud and Specialization equilibrium with dual practice**

For Pure fraud and Specialization equilibrium, \(m \in [0,1]\). The planner sets the price in the hospital at the same level as a non-GP clinic to eliminate excess demand. The equilibrium condition for clinic physicians including both GP and non-GP is the same as for the case with no dual practice. Market prices for non-GP private clinics are set by (11).

The total number of non-GP physician’s working time and a GP’s total working time are also the same as the regime without dual practice (i.e., \(T^*\)). Optimal inputs condition remains the same as (1). Therefore, (9) still holds and the critical value \(m\) remains the same as the regime disallowing dual practice. However, compared to the regime disallowing dual practice, more patients are diagnosed in the hospital since some of them will be referred to dual practicing clinics after diagnosis.

After allowing dual practice, there are two possible equilibria, depending on a corner solution or an interior solution of the optimal self-referral problem for the dual practicing physician.

**First equilibrium: the interior solution**

If the benefits for misrepresenting a hospital patient with mild diseases are not high enough, a dual practicing physician will not refer all the mild patients to his own clinic (i.e., the interior solution). \(T_d\) denotes the number of working hours for a dual practice physician, whose optimization problem is:

\[
\max_{m_d, x_{cg}^d, r_{cg}^d} \{s + (p_{hb} - x_{cg}^d)m_d - \varphi(T_d)\}
\]
Where $T_d = [(1 - v^d)t_{hb} + v^d t_{hg} + \bar{d}] \frac{H}{q} + m_d(t_{cg}^d + \bar{d})$ and $v^d = v^s - \frac{(1 - v^s)m_d q}{H}$

$m_d$ is the number of patients being self-referred from the hospital to a private clinic by a dual practicing physician. The proportion of mild cases in the hospital is $v^d$. The first order condition with respect to the number of self-referral $m_d$ is:

$$p_{hb} = x_{cg}^d + [(1 - v^s)(t_{hb} - t_{hg}) + t_{cg}^d + \bar{d}] \varphi'(T_d)$$ (14)

$t_{cg}^d$ and $x_{cg}^d$ denote the inputs for a dual practice physician. To clear the market, $p_{hb} = p_{cb}$ and $p_{cb}$ is determined by (11). Therefore, $m_d$ is determined by (14) after substituting $p_{cb}$ with (11).

The planner’s problem is

$$\min_{x_{hb}, q} \{ \Gamma + Hx_{hb} + sq + vmNp_{cg} + (N - vmN - H)p_{cb} + [N - H - m_d q]p_d \}$$

s.t. $\overline{u} = s + (p_{hb} - x_{cg}^d)m_d - \varphi(T_d)$

There are $H + m_d q$ patients being diagnosed by the hospital. We can show that dual practicing physicians’ total number of working hours $T_d$ is the same as previous equilibria (i.e., $T^*$), which implies that the optimal time inputs and non-physician inputs are the same as the case with full information.

The optimal input combination in the hospital is:

$$(1 - v^d) \frac{\partial b_{hb}}{\partial p_{hb}} \bigg|_{b_{hb} = \bar{b}} + v^d \frac{\partial g_{cg}}{\partial q} \bigg|_{q = \bar{q}} = \frac{1}{\sigma_{\varphi(T^*)}}$$ (15)

In (15), non-physician input per patient in hospital is smaller compared with full information case (6).

However, the proportion of mild cases in the hospital $v^d = v^s - \frac{(1 - v^s)m_d q}{H} < v^s < v$ is less than the regime disallowing dual practice. The amounts of non-physician inputs in the hospital are higher than the regime disallowing dual practice.

**Proposition 3**: For the interior solution equilibrium, the social welfare is higher when dual practice is allowed, compared with the regime disallowing dual practice with imperfect information.
Proof: see Appendix.

There are two reasons for welfare improvement: first, the welfare improvement is due to the saving of salary expenditure. The intuition here is that after allowing dual practice, dual practicing physicians are willing to accept a lower salary from the hospital since they have additional revenue from their own clinics. Second, since more serious cases are treated in the hospital, non-physician inputs for serious cases in the hospital are more efficiently utilized in general.

**Second equilibrium: the corner solution**

If the benefits for misrepresenting a mild patient in the hospital are high enough, a dual practicing physician will refer all the mild patients to his own clinic (i.e., the corner solution).

The planner's problem is defined by:

\[
\min_{x_{nh}, q} \left\{ \Gamma + Hx_{nh} + sq + vmNp_{cg} + (N - vmN - H)p_{cb} + \left[ N - \frac{H}{1-\psi} \right] p_d \right\}
\]

The last item denotes total diagnosis fee \( N - \frac{H}{1-\psi} \) paid to the private sector since there are \( \frac{H}{1-\psi} \) patients who are diagnosed in the hospital. The first order condition with respect to \( q \) is (4). Total number of working hours for a physician is the same as full information case. Also, it is straightforward to show that planner's optimal input condition is defined by (6) (i.e., the same as full information case). The intuition is that there is no patient with mild disease being treaded in the hospital under this “corner solution” equilibrium and the input combination is efficient in the hospital as the full information case.

**Proposition 4:** For the corner solution equilibrium, in general, dual practice is a second best welfare improvement when information is imperfect. Welfare achieves the first best level when \( m = 0 \).

**Proof:** See appendix.
Again, the improvement of welfare after allowing dual practice is due to the saving of salary expenditure and more efficiently usage of non-physician inputs in the hospital. In contrast to our intuition, if the switching cost is very high (i.e. \( m \) is close to 0), social welfare increases with switching cost. This is because fewer people are willing to switch to second providers if the cost to do so is higher. For both of the corner and interior equilibria, it is interesting that the welfare improvement is not due to reduction of switching costs, given the number of patients who choose to go to a GP for diagnosis is the same as the regime without dual practice.

**Conclusions and discussion**

In the literature, models have been constructed in which physician dual practice improves welfare under the assumption that physicians are heterogeneous in morality or quality. We show in this paper that allowing dual practice can improve welfare even when physicians have homogeneous quality/morality. When information is asymmetric among physicians, patients and the planner, dual practice can be conceived as a tool to improve welfare in two ways: first, and most importantly, resource allocation within the hospital is more efficient; second, allowing dual practice can save salary expenditure for the hospital.

Multiple stable equilibrium may emerge given heterogeneous switching costs among patients. Dual practice can improve welfare in general under different equilibrium. The specialization equilibrium can offer some additional insights. As long as health service market is not fully specialized (i.e., the hospital still treats some people with mild diseases), dual practice can be interpreted as a substitute institution to achieve specialization among public and private providers. Further, in our model, people with high opportunity cost (high income) in switching providers will be more likely to go to the hospital while people with low opportunity cost in switching providers (low income) will visit the GP first. Hence, after allowing dual practice, rich patients with mild diseases are more likely to be referred to private clinics from the hospital. Patients with serious conditions, or low income patients, are more
likely to be treated in the hospital. Therefore, dual practice can be interpreted as an alternative instrument for sorting patients in terms of both illness severity and switching costs.

This paper is built upon the assumption that the health service market is perfectly competitive with respect to physician fees. A possible extension for this paper may be to treat the health service market as monopolistic competitive, which can be applied to a wider context. In this paper, the level of quality can only take two values (i.e., smaller or not smaller than the threshold value). In the real world, health service providers may compete in continuous quality, which may not be observable for the planner. Consequently, a dual practice physician's behavior may change. For example, in India, dual practice physicians provide services with relatively low quality in public hospitals so that high income patients are willing to pay more to be treated at physicians' private clinics (Yip and Mahal 2008). Also, Physicians' behavior and the welfare implications of dual practice may be different when patients have insurance. We leave these issues for the future research.
References


Blomqvist, Å and Leger, P (2005). "Information asymmetry, insurance, and the decision to hospitalize" *Journal of Health Economics*, 24(4), 775-93,


González, P (2005), "On a policy of transferring public patients to private practice", *Health Economics*, 14, 513—27


Jan, S, et al (2005) "Dual job holding by public sector health professionals in highly resource-constrained settings: problem or solution?*. *Bull World Health Organ* 83(10)


**Appendix:**

**Lemma 1: Proof:** Since there may be some physicians who only treat serious cases, case mix in a clinic is not equal to $v$. The proportion of mild cases in a clinic is denoted by $r$, where $0 \leq r < 1$. For physicians who specialized in serious case, $r = 0$. $m_c$ is the number of patients in a clinic.

$$ \max_{m_c, x_{ci} | c_i = b, g} \{(p_{eg} - x_{eg})rm_c + (p_{cb} - x_{cb})(1 - r)m_c + p_dm_c - \varphi(T_c)\} $$
We have the first order condition on (1) with respect to $T_c$.

\[
\left(p_{eg} - x_{eg} - t_{eg}\phi'(T_c)\right)r + \left(p_{cb} - x_{cb} - t_{cb}\phi'(T_c)\right)(1 - r) = 0
\]  

(A1)

It is easy to see that the above equation is a weighted summation of (2). $rp_{eg} + (1 - r)p_{cb}$ is the expected price for a patient before seeing a doctor.

Suppose the equilibrium price for mild cases is larger than price in (2). A physician can extract profit for treating a mild case. This cannot be an equilibrium since a physician can deviate to reduce price for mild cases by a very small number $\epsilon$ where $\epsilon > 0$ and increase the price for serious case to maintain a same expected price for patients. In this case, since all patients, who are diagnosed by physicians specializing in treating serious case, already know they are mild cases, these patients will prefer to be treated in this deviating clinic with lower fee for mild cases. Deviating physician will have extra profits from these patients. Hence, every other physician will follow this deviating strategy until (2) is fulfilled.

Suppose the equilibrium price for the mild case is less than the price in (2). Therefore, a physician cannot cover marginal cost when treating a mild case. A deviating physician can increase the fee for mild cases and decrease the price for serious cases. All cases referred from physicians specializing in serious cases will choose other providers and deviating providers will make profit. Hence, every other physician will follow this deviating strategy until (2) is fulfilled. Substituting (2) and (3) into physician’s utility function, we have condition (4) to clear the market.

In (4), when $T_c = 0$, $T_c\phi'(T_c) - \phi(T_c) = 0$. $T_c\phi'(T_c) - \phi(T_c)$ is monotonically increasing in $T_c$.

Hence, only a unique $T_c$ can satisfy (4), which we denote as $T^*$.

There is no unique equilibrium. The reason is that the number of clinic physicians who only treat serious cases cannot be determined. Q.E.D.
Lemma 2: Proof: If it is profitable to deviate to be a GP physician (i.e., utility is not less than the reservation utility), \((p_{cg} - x_{cg})vm_{cg}N + p_d m_{cg}N - \varphi (vm_{cg} N t_{cg} + m_{cg} N \bar{d}) > 0\): A pure fraud equilibrium breaks down and an equilibrium with specialized GPs emerges. Also, we show below that a physician’s utility is increasing in \(m_{cg}\) if expected fee charged for a patient \(vp_{cg} + p_d\) is higher than the expected marginal cost for treatment and diagnosis: \(vx_{cg} + (vt_{cg} + \bar{d})\varphi (vm_{cg} N t_{cg} + m_{cg} N \bar{d})\)

\[
\frac{\partial [(p_{cg} - x_{cg})vm_{cg}N + p_d m_{cg}N - \varphi (vm_{cg} N t_{cg} + m_{cg} N \bar{d})]}{\partial m_{cg}}
\]

\[
= N [v(p_{cg} - x_{cg}) + p_d - (v t_{cg} + \bar{d}) \varphi (vm_{cg} N t_{cg} + m_{cg} N \bar{d})] > 0
\]

if \(vp_{cg} + p_d > vx_{cg} + (vt_{cg} + \bar{d})\varphi (vm_{cg} N t_{cg} + m_{cg} N \bar{d})\)

In this case, a deviating doctor will admit \(
\frac{v(p_{cb} - p_{cg})}{(1 - v)\tau} - \frac{p_d}{\tau}
\) from (9).

Define \(\overline{m}_{cg}\) where \(vp_{cg} + p_d = vx_{cg} + (vt_{cg} + \bar{d})\varphi (v\overline{m}_{cg} N t_{cg} + \overline{m}_{cg} N \overline{d})\), where marginal cost is equal to marginal revenue for admitting a patient. A doctor will admit \(\overline{m}_{cg}\) patients if \(\frac{v(p_{cb} - p_{cg})}{(1 - v)\tau} - \frac{p_d}{\tau}\) > 0.

From above, the number of patients admitted is \(\min(\overline{m}_{cg} \frac{v(p_{cb} - p_{cg})}{(1 - v)\tau} - \frac{p_d}{\tau})\)

Q.E.D

Proposition 1: Proof: As we show in the proof of Lemma 2, for a GP physician’s utility is

\[
\min(\overline{m}_{cg} \frac{v(p_{cb} - p_{cg})}{(1 - v)\tau} - \frac{p_d}{\tau}).
\]

If \(\exists \tau\), to keep the demand for a deviating doctor’s clinic small enough to let the deviating doctor’s utility is less than \(\overline{u}\), the pure fraud equilibrium is stable.

Also, there is one stable specialization equilibrium. Substituting (2), (10), (11) into (9), we have

\[
\frac{1}{1 - m r} [(x_{cb} - x_{cg}) + (t_{cb} - t_{cg}) \varphi'(T^*) - (m r + p_d)] = 0
\]

(A3)
(A3) can be transformed into a quadratic function of \(m\) as follows:

\[
\nu \tau m^2 - m(\tau - v p_d) + v[(x_{cb} - x_{cg}) + (t_{cb} - t_{cg})\phi'(T^*)] = 0
\]

Two roots \(m_1\) and \(m_2\) for above equation are solved as follows:

\[
m_1 = \frac{1}{2\nu \tau} (\tau - v p_d - \sqrt{\tau^2 v^2 - 4\tau v^2[(x_{cb} - x_{cg}) + (t_{cb} - t_{cg})\phi'(T^*)]})
\]

\[
m_2 = \frac{1}{2\nu \tau} (\tau - v p_d + \sqrt{\tau^2 v^2 - 4\tau v^2[(x_{cb} - x_{cg}) + (t_{cb} - t_{cg})\phi'(T^*)]})
\]

The following assumption is required to have either of these two roots as solution:

\[
(\tau - v p_d)^2 - 4\tau v^2[(x_{cb} - x_{cg}) + (t_{cb} - t_{cg})\phi'(T^*)] \geq 0
\]

It is straightforward that \(m_1 \in [0, \frac{1}{2\nu}]\) if above inequality holds. We assume \(p_d\) is very small compared to \(\tau\). We normalize \(p_d\) as zero we can show that the root \(m_2\) is located between (0,1] if

\[
\tau \leq \frac{v}{1-v}[(x_{cb} - x_{cg}) + (t_{cb} - t_{cg})\phi'(T^*)]
\]

Both \(m_1\) and \(m_2\) can emerge as equilibrium. To see whether these equilibriums are stable, by transforming (A3), net benefit \(\pi\) for a patient when he chooses a GP is:

\[
\pi = \frac{v}{1-v}[x_{cb} - x_{cg}] + (t_{cb} - t_{cg})\phi'(T^*)] - (m\tau + p_d)
\]

If \(m\) is stable, a perturbation of \(m\) will always shift \(\pi\) into the opposite direction, which implies \(\frac{\partial \pi}{\partial m} \leq 0\). We can show that, with assumption (A4), \(\frac{\partial \pi}{\partial m} \leq 0\) if \(m \leq \frac{1}{2\nu}\). In this case, \(m_1\) is always stable. If \(m > \frac{1}{2\nu}\), with assumption (A4), \(\frac{\partial \pi}{\partial m} \leq 0\) if only \(\tau \geq \left(\frac{v}{1-v}\right)^2[(x_{cb} - x_{cg}) + (t_{cb} - t_{cg})\phi'(T^*)]. If \(\tau < \left(\frac{v}{1-v}\right)^2[(x_{cb} - x_{cg}) + (t_{cb} - t_{cg})\phi'(T^*)], \(m_2\) is not stable in equilibrium.

With a slight increase of \(m\), a fully specialization equilibrium will emerge. To sum up, there are more than one stable specialization equilibriums if

\[
\frac{v}{1-v}[x_{cb} - x_{cg}] + (t_{cb} - t_{cg})\phi'(T^*)] \geq \tau \geq \left(\frac{v}{1-v}\right)^2[(x_{cb} - x_{cg}) + (t_{cb} - t_{cg})\phi'(T^*)]. Otherwise, there is one stable specialization equilibrium. Q.E.D

**Proposition 2. Proof:** After comparing welfare with the full information case, incremental social cost between the specialization equilibrium and full information is:
mvN \left[ \frac{(1 - v)m_r}{2} + (1 - v)p_d \right] \\
+ H \left[ v^s (p_{cb} - p_{cg}) + x_{hb} - x_{hb}^* + \left[ v^s t_{hg} + (1 - v^s) t_{hb} - t_{hb}^* \right] \psi'(T^*) \right] \\
> \text{mvN} \left[ \frac{(1 - v)m_r}{2} + (1 - v)p_d \right] \geq 0 \tag{A5}

where \( m \in [0,1] \). Variables with superscript * denote the optimal choice variables under full information. Hence, under a specialization equilibrium, the total social cost increases compared with the full information case. The first item of (A5) denotes the additional switching cost under a specialization equilibrium, which is always positive. The second item of (A5) denotes additional resources transferred to physicians in the private sector. The third item with bracket denotes changes of resources spent in the hospital. The inequality of (A5) holds when the second item in (A5) is positive, which can be arranged into:

\[ [v^s p_{cb}^* + x_{hb} + (v^s t_{hg} + (1 - v^s) t_{hb}) \psi'(T^*)] - [v^s p_{cg}^* + x_{hb}^* + t_{hb}^* \psi'(T^*)] > 0 \]

The item within the first brackets denotes social costs to treat \( v^s \) serious cases in a clinic and \( v^s \)'s mild cases together with \( 1 - v^s \) serious cases in the hospital. The item within second brackets denotes the social cost when the hospital treats one serious case and clinics treat \( v^s \) serious cases. Here, \( x_{hb}, t_{hb} \) and \( t_{hg} \) are optimal inputs in specialization equilibrium when the hospital treats \( v^s \) mild cases.

If it is more efficient to treat a serious case in the hospital than in a clinic, resources required to treat a serious case in the hospital and a mild case in clinic are always less than resources required to treat a mild case in the hospital and a serious case in a clinic (i.e., swapping treatment location). Therefore, the above inequality holds. The welfare under specialization equilibrium is less than the full information case. Furthermore, pure fraud equilibrium is a special case of specialization equilibrium when \( m = 0 \). Therefore, from (A5), social welfare under pure fraud equilibrium is less than the social welfare with full information. Q.E.D.
**Proposition 3:** Proof: To prove this proposition, we compare the social welfare function with dual practice and without dual practice. Let \( \{p_{hb}^*, x_{hb}^*, t_{hb}^*, s^*, q^*\} \) be the vector minimizing (12) when dual practice is disallowed. If the planner implements \( p_{hb}^*, x_{hb}^*, t_{hb}^*, s^*, q^* \) but chooses a \( \tilde{s} \), which is different from \( s^* \), when dual practice is allowed. Let \( T_d \) denote the number of physicians’ working hours when planner implements \( p_{hb}^*, x_{hb}^*, t_{hb}^*, s^*, q^* \). The difference of social costs between the regime disallowing dual practice and the regime allowing dual practice is:

\[
 s^*q^* - \tilde{s}q^* = q^*\left[\varphi(T^*) - \varphi(T_d) + (p_{hb} - x_{cg}^d)m_d \right] = q^*\left[\varphi(T^*) - \varphi(T_d) + [(1 - v^s)(t_{hb} - t_{hg}) + t_{cg}^d + \tilde{d}]m_d\varphi'(T_d) \right] \tag{**}
\]

The last equality holds from (14). Since \( \varphi(\cdot) \) is a strictly convex function, after substituting \( T_d \), we have the property for convex function that

\[
\varphi(T^*) - \varphi(T_d) > (T^* - T_d)\varphi'(T_d) = -[(1 - v^s)(t_{hb} - t_{hg}) + t_{cg}^d + \tilde{d}]m_d\varphi'(T_d) 
\]

From above equation, equation (**) can be transformed into

\[
q^*\left[\varphi(T^*) - \varphi(T_d) + [(1 - v^s)(t_{hb} - t_{hg}) + t_{cg}^d + \tilde{d}]m_d\varphi'(T_d) \right] > 0
\]

Therefore, total amount of salary is strictly less after allowing dual practice when \( m_d > 0 \). Total social costs are less after allowing dual practice. Q.E.D.
**Proposition 4: Proof:** When the planner implements the same optimal inputs $x_{hb}, t_{hb}$ as in the case of full information, we can show that additional social costs with dual practicing physicians, compared the social welfare under full information (i.e. (5)), are: $\text{mvN} \left[ \frac{(1-v)mt}{2} + (1-v)p_d \right] > 0$ (i.e. aggregate switching costs).

From proof of Proposition 2 (i.e. (A5)), welfare loss under specialization equilibrium without dual practicing physicians, compared to social cost with full information, is strictly greater than $\text{mvN} \left[ \frac{(1-v)mt}{2} + (1-v)p_d \right]$. Under specialization equilibrium, allowing dual practice then is a second best welfare improvement.

If $m = 0$, a pure fraud equilibrium emerges at the private sector. Private sector is efficient under pure fraud equilibrium. Since all patients in the hospital are serious cases, resource allocation in the hospital is also efficient. Hence, dual practice can improve social welfare at the first best. **Q.E.D**