Pacific Economic Review

Pacific Economic Review, 17: 4 (2012) pp. 514–524
doi: 10.1111/j.1468-0106.2012.00597.x

INFORMAL PAYMENTS IN DEVELOPING COUNTRIES' PUBLIC HEALTH SECTORS

TING LIU* *Stony Brook University*
MONIC SUN *Stanford University*

Abstract. In many developing countries, public patients offer payments to their doctors outside the official payment channels. We argue that the fundamental reason for these informal payments is that formal prices cannot fully differentiate patients' various needs. We compare patient welfare and social efficiency when informal payments are allowed with the scenario when they are banned. Patient heterogeneity plays a central role in the comparison. Contrary to conventional wisdom, allowing informal payments always improves social efficiency when patients do not face income constraints. Moreover, allowing informal payments improves patient welfare if patients' willingness to pay differs significantly.

1. INTRODUCTION

A World Bank report by Lewis (2000, p. 5) begins with: 'Informal payments in the health sector in Eastern Europe and Central Asia are emerging as a fundamental aspect of health care financing and a serious impediment to health care reform'. By definition, such informal payments are those made to individuals or institutions in cash or in kind outside official channels for services that are meant to be covered by the public health-care system. In China, for example, informal payments for medical services are often given in 'red packets. Such payments have become a pressing social issue. The Chinese Government treats these payments as bribes and has already imposed a national policy that whenever a doctor is found to accept informal payments, his or her license is immediately suspended by the Ministry of Health. Nevertheless, patients are still offering such payments. The Ministry of Health reports that in 2004, Chinese doctors returned to patients or turned in to the state informal payments totalling RMB41.36m (roughly $US5m). As there is little incentive for doctors to give up the informal payments, the actual amount of informal payments may be much higher than reported.[1]

Many health-care professionals believe that patients offer informal payments to induce more effort from doctors, whereas others posit that the purpose is to conform with the social norm. As long as patients are rational economic agents,

**Address for Correspondence*: Economics Department, SUNY at Stony Brook University, Stony Brook, NY 11794, USA. E-mail: tingliu78@gmail.com. This paper started when Ting Liu and Monic Sun were PhD students at Boston University. We thank Albert Ma, Jacob Glazer, Bart Lipman, Dilip Mookherjee and participants at Boston University Micro Theory Workshops for helpful conversations. We also thank the Associate Editor for valuable suggestions.

[1] Lewis (2000) lists the frequency of informal payments in some other countries as follows: Armenia (1999) 91%; Vietnam (1992) 81%; Azerbaijan (1995) 78%; Poland (1998) 78%; Kyrgyz Republic (1999) 75%; and Russian Federation (1997) 74%. Note that Armenia data is based on a non-representative national sample inpatient care only. Data for Poland is for inpatient care only.

they must be paying for *something valuable*, which could be a higher level of effort by doctors, the choice of a better doctor or a better position on the waiting list. In any case, there must exist some mechanism that ensures that patients recieve better services when they make greater informal payments. For example, doctors are concerned about their reputation in a repeated game: if they do not react to informal payments in one period, they lose all future payments. Alternatively, they might simply feel guilty for not investing more effort when being paid more.

It is not our research objective to characterize this mechanism in a supergame. Instead, we take the mechanism as effective and simply assume that doctors react to informal payments. To fix ideas, we model better services as the option of seeing a more capable doctor, and informal payments serve as a device for patients to compete for better services.

We take the stance that patients have more information about doctors than the administrators. The Ministry of Health in China, for example, ranks doctors into different categories (experts, chief doctors and ordinary doctors) and sets a uniform price for seeing doctors in each category. The ranking criteria include medical degree, years of practising, publications and number of patients that they have treated. Ideally, doctors' salaries should be based on their skills. However, because doctors' skills are not directly observed by the hospital administrators, doctors' salaries are set according to their rank, which serves as an approximate of their skills.[2] Doctors of the same rank, however, often vary in their skills. Patients, in contrast, might be able to make better judgements regarding a doctor's skill. They can gather information about the doctor from their own personal experience, their friends' recommendations, or even online reviews. Such first-hand information is crucial for doctors to build up a reputation among patients.

In our model, the public insurance and the health-care providers are integrated and they set the formal prices for doctors' services that are paid by patients.[3] Because the formal prices for seeing doctors are low and do not reflect the difference in doctors' skills, patient are willing to pay more out of their own pocket to be treated by more capable doctors. This is the foundation of our model: the actual quality of care varies among different doctors who are paid the same through the formal channels. By modeling patients' competition through informal payments, we discuss the welfare implications when informal payments are banned and when they are allowed. Social efficiency does not depend on any transfers, and, thus, the amount of informal payments, whereas patient welfare depends crucially on such payments.

A crucial factor in welfare analysis is patient heterogeneity: informal payments should be allowed if and only if patients' incremental willingness to pay is

[2] In China's public hospitals, doctors' salaries have three components: base salary, diagnosis fee and treatment fee. The last two depend on doctors' rank.

[3] China's public hospitals are partially fund by the government. The prices for doctors' services are set too low to cover the treatment costs. The gap is partly subsidized by the government and partly from hospitals' revenue from selling medicine.

heterogeneous. Intuitively, allowing informal payments improves allocation efficiency because seriously ill patients are willing to pay more and, hence, are more likely to be treated by the more capable doctors. However, banning informal payments helps patients to save money. When patients differ greatly in their incremental willingness to pay, achieving the optimal allocation is most important. If patients differ little, the competition becomes wasteful as the allocation is barely better than random. At the extreme, when patients are identical, they offer the same informal payments for the more capable doctors. As a consequence, providing informal payments will not improve anyone's chance of seeing the more capable doctors; this will be a waste of the patients' money.

Although we focus on informal payments in the public health sector, our model applies to more general situations where individuals or firms compete by bribing a bureaucrat for a limited resource. For example, one may modify the model to describe two customers competing for promptness in passport delivery. Each customer knows his or her own willingness to pay and can also observe the anxiousness of the other customer. The model would predict that if the two customers' willingness to pay is sufficiently different, the existence of bribes to service officers could improve the aggregate customer welfare.[4] In this perspective, our paper fits in with the literature on corruption (Leff, 1964; Lui, 1985; Beck and Maher, 1986) that argues that corruption is the much-needed grease for the squeaking wheels of a rigid administration.[5]

Published literature on informal payments in the health sector is quite limited. Lewis (2000) points out that informal payments arise to alleviate the mismatch between specialties needed and specialties provided. Garcia-Prado (2005) considers the severity of doctor punishment and the bargaining structure between patients and doctors in determining the equilibrium amount of informal payments. Garcia-Prado does not model competition among patients. Biglaiser and Ma (2007) and Gonzalez (2004) study 'moonlighting', a related phenomenon where public sector doctors work part time for private hospitals. They focus on how doctors divide their labor supply between the public and private sectors, in which reimbursement schemes are different.

As we assume that patients' informal payments are not refunded even if they do not get to see the better doctor, our model is essentially an 'all-pay auction'.

[4] It is important to note that aggregate welfare is only one of the many goals of an average social planner. Equity across different parties in the economy, for example, is at least as important. We do not discuss equity and other social goals in the current paper.

[5] Our model formalizes the idea of Leff (1964) that corruption may improve efficiency when the government and bureaucracy fail to allocate resources efficiently. In a competitive bidding environment, Beck and Maher (1986) show that the government will always award the contract to the low-cost firm because it can offer the largest bribe. In contrast to Beck and Maher (1986), agents in our model differ in their incremental willingness to pay for high quality service. Similar to Beck and Maher (1986), in our model, the agent who values the high quality service the most will have a greater chance of getting the service. Lui (1985) argues that corruption will minimize the waiting cost of the queue and, hence, improve efficiency. The efficiency gain from corruption in Lui's (1985) model is the reduced waiting time. In contrast, in our model, the gain from corruption is the correct allocation of the more capable doctor. Bardhan (1997) offers a survey of the literature on corruption.

The analytical framework is conceptually similar to the 'menu auction' used in Bernheim and Whinston (1986) and Grossman and Helpman (1994). Bernheim and Whinston (1986) describe influence-seeking as an example of a 'menu auction' game. In a menu auction, each of several principals who will be affected by an action offers a bid to an agent who will take that action. These bids take the form of schedules that associate a payment to the agent with each feasible option. Once the agent chooses an action, all of the principals pay the bids stipulated by their schedules. Bernheim and Whinston (1986) define an equilibrium in a menu auction as a set of contribution schedules such that each one is a best response to all of the others, and an action by the agent that maximizes their utility given the schedules that confront them. In our model, bids take the form of a simple one-dimensional offer rather than a schedule. Hillman and Riley (1989) study political rents and transfers in an all-pay auction similar to ours.

Section 2 introduces the model. Section 3 compares allowing and banning informal payments. Section 4 summarizes the welfare analysis. Section 5 discusses patients' income constraints and Section 6 concludes.

2. THE MODEL

There are two patients and two doctors. Each doctor can treat only one patient. One patient's illness is serious (H) and the other's is common (L). One doctor is more capable and has a good reputation among patients (G); the other doctor is ordinary (B). Whether each doctor is more capable or mediocre is known to the patients but not to the social planner. Therefore, the formal price is the same for a patient seeing any of the two doctors and is normalized to zero.

If patient $i \in \{H, L\}$ is treated by doctor $j \in \{G, B\}$ after paying p amount of informal payment, patient i's utility is $v_i^j - p$, where v_i^j is patient i's benefit from being treated by doctor j. Both patients receive greater benefit from being treated by the more capable doctor: $v_i^G > v_i^B$ for $i \in \{H, L\}$. In addition, the seriously ill patient has a larger increase in benefit when he or she is treated by the good doctor rather than the ordinary doctor:

$$v_H^G - v_H^B > v_L^G - v_L^B, \quad or \quad \Delta H > \Delta L, \tag{1}$$

where $\Delta H = v_H^G - v_H^B$ and $\Delta L = v_L^G - v_L^B$ are the seriously ill patient's and the common patient's (incremental) willingness to pay for being treated by the more capable doctor, respectively.

The cost of treating a patient is also normalized to zero. Doctors commit to selecting the patient who offers greater informal payment. This motivation of this assumption is twofold. First, the more capable doctor has concerns for future profits. By committing to treat the patient who pays the most, the doctor gives future patients a strong incentive to offer informal payments. Second, the doctor wants to minimize the risk of patient retaliation. In China as well as many other developing countries, taking informal payments is considered a

breach of ethics, even if it is not illegal. Doctors may have their licenses suspended once patients report the bribing event to the hospital or the relevant government officials.[6] The patient who has paid the greater informal payment is more likely to retaliate if he or she does not receive high quality service.

The two patients offer informal payments to attract the good doctor and neither of them offers any informal payment to the general doctor. When there is a tie in the two offers of informal payments, the more capable doctor randomly selects a patient. In addition, informal payments can never be refunded. In reality, doctors do one of three things upon receiving informal payments: hand in the money to hospital administration, keep the money or take the money but then return it after treatment. As informal payments are illegal, it is not possible to find statistics on the frequency of each option. To the authors' best knowledge, offering informal payments has become a norm in most public hospitals, at least in China, and they are often not refunded. A patient is unlikely to confront the doctor to request a refunded as the patient and their family might need to see the doctor again in the future.

If the seriously ill patient is matched with the more capable doctor, total patient welfare is $v_H^G + v_L^B$. In contrast, if the common patient is matched with the more capable doctor, patient welfare is $v_L^G + v_H^B$. The first best allocation is to have the seriously ill patient matched with the more capable doctor. If there is a private market of health care in which a Walrasian auctioneer sets price P_j for the service of doctor j, $j \in \{B, G\}$, any pair of prices (P_B, P_G), with $P_B = 0$ and $P_G \in (\Delta L, \Delta H)$ sustains a Walrasian equilibrium. To see this, the seriously ill patient's utility from buying the service of the more capable doctor is $v_H^G - P_G$; his or her utility from buying the service of the less capable doctor is v_H^B. Hence, the seriously ill patient will pay for the more capable doctor if and only if $P_G < \Delta H$. The common patient's utility from buying the service of the more capable doctor is $v_L^G - P_G$; his or her utility from buying the service of the less capable doctor is v_L^B. The common patient will not pay for the more capable doctor if and only if $P_G > \Delta L$. Therefore, the Walrasian equilibrium achieves the first best allocation when $\Delta L < P_G < \Delta H$.

3. THE INFORMAL PAYMENT GAME

Accepting informal payments is illegal in many developing countries. Now, we discuss patient welfare when the social planner can use certain methods to successfully ban informal payments. In this case, the allocation is random, resulting in patient welfare:

[6] In most provinces in China, a doctor, upon being reported to have accepted informal payments, would be punished (fine, demotion or even unemployment), while the patient who is found to have offered an informal payment normally does not receive any substantial punishment. See a relevant Chinese commentary at http://news.sohu.com/20120718/n348486688.shtml.

$$\frac{1}{2}(v_H^G + v_H^B) + \frac{1}{2}(v_L^G + v_L^B).$$

Obviously, patient welfare is lower than in the first best allocation.

What happens if the social planner allows informal payments? Patients compete for the more capable doctor by offering informal payments. We assume that a patient's health status and, hence, his or her willingness to pay is known to the other patient, but unknown to the doctors.[7] The assumption that a patient observes his or her competitor's health status is motivated by the fact that informal payments are offered mostly by in-hospital patients.[8] In many developing countries, inpatients share their wards due to limited space. Although it is not reasonable to assume that every patient observes every other patient's health condition, it is often the case that two or three inpatients with the same illness are assigned to the same ward. We model the competition among these patients by analyzing the following game:

Stage 1 Patients simultaneously offer informal payments, P_i, to the more capable doctor before their treatments. Once a patient pays informal payments, the money cannot be refunded.
Stage 2 The more capable doctor commits to treating the patient who offers more informal payments. When both patients offer the same informal payments, the more capable doctor randomly selects one patient to treat.

We solve for Nash Equilibria of this game.

PROPOSITION 1. *There is no pure strategy Nash equlibrium.*

PROOF. Given the patients' willingness to pay, $P_H \leq \Delta H$ and $P_L \leq \Delta L$. There is a pure strategy equilibrium (P_L^*, P_H^*). First, suppose $P_L^* = P_H^*$. If the seriously ill patient deviates to offer $P_H^* + \varepsilon$, he or she receives the more capable doctor for sure and suffers a payment loss of ε. As long as $\varepsilon < \Delta H/2$, the seriously ill patient recieves greater utility. Second, suppose $0 < P_L^* < P_H^*$. The common patient benefits from deviating to offer zero informal payment. Third, suppose $0 = P_L^* < P_H^*$. The seriously ill patient benefits from deviating to offer $P_L^* + \varepsilon'$, as long as $\varepsilon' < P_H^* - P_L^*$. Fourth, suppose $P_L^* > P_H^*$. The seriously ill patient's utility is $v_H^B - P_H$. He or she benefits from deviating to offer $P_L^* + \varepsilon''$ as long as $\varepsilon'' < \Delta H + P_H^* - P_L^*$. Summarizing the four cases, we conclude that there does not exist a pure strategy Nash equilibrium.

[7] If we assume that each patient observes his or her own willingness to pay, but not that of his or her competitor, a mixed strategy Nash equilibrium similar to the one characterized in the current paper still exists. A proof is available upon request. In addition, assuming doctors know the patients' willingness to pay does not change the result, as the more capable doctor commits to treating the patient who offers more informal payments.

[8] Eggleston *et al.* (2008) discuss the scale and sources of informal payments in China.

This result comes from the continuity in patients' offers of informal payments. Each patient wants to outbid the other by only an infinitesimal amount and, hence, no pure strategy equilibrium can be sustained. We now turn to mixed strategy equilibria.

Let $F_i(x)$, with $i \in \{L, H\}$, denote patient *i's* cumulative distribution function of offering informal payments.

PROPOSITION 2. *The unique mixed strategy Nash equilibrium is:*

$$F_L(x) = \begin{cases} 1 - \dfrac{\Delta L}{\Delta H} + \dfrac{x}{\Delta H}, & 0 \le x \le \Delta L, \\ 1, & x > \Delta L; \end{cases} \quad F_H(y) = \begin{cases} \dfrac{y}{\Delta L}, & 0 < y \le \Delta L, \\ 1, & y > \Delta L. \end{cases}$$

The proof of the proposition is involved and, hence, is placed in the Appendix.

Proposition 2 has five implications. First, the lower is $\dfrac{\Delta L}{\Delta H}$, the more likely it is that the common patient offers *no* informal payment. When the seriously ill patient is willing to pay a great deal more for the more capable doctor than the common patient is, the seriously ill patient offers a bulky 'red packet'. The common patient has little hope of winning the competition, and his or her incremental utility from being treated by the more capable doctor is low. As a result, the common patient would rather quit the competition and save some money.

Second, when the two patients have similar willingness to pay, they both offer a strictly positive informal payment amount. In this case, their random offers turn into a wasteful competition as the allocation is such that each patient gets the more capable doctor with probability 0.5. If the two patients can both commit to not paying informal payments, both are better off.

Third, the seriously ill patient is more likely to be treated by the more capable doctor. The probability of the first best allocation is:

$$\Pr(P_H > P_L) = 1 - \frac{\Delta L}{\Delta H} + \int_0^{\Delta L} \left(\int_x^{\Delta L} \frac{1}{\Delta L} dy \right) \frac{1}{\Delta H} dx = 1 - \frac{1}{2} \frac{\Delta L}{\Delta H} > \frac{1}{2}$$

The stronger the heterogeneity in the patients' willingness to pay, the higher the probability of the seriously ill patient being treated by the more capable doctor.

Fourth, the ratio of the two patients' expected informal payments equals the ratio of their willingness to pay.

Finally, the total amount of informal payments, $E(P_L) + E(P_H) = \dfrac{\Delta L^2}{2\Delta H} + \dfrac{\Delta L}{2}$, increases with the common patient's willingness to pay and decreases with the seriously ill patient's. As the common patient pays more, the seriously ill patient also pays more, and the total amount of informal payments increases. In contrast, when the seriously ill patient pays more, the common patient is more likely to quit the competition, which reduces the total amount of informal payments.

4. WELFARE ANALYSIS

When informal payments are allowed, the expected utility of the common patient is v_L^B and that of the seriously ill patient is $v_H^G - \Delta L$. Recall that, when informal payments are banned, each patient has half the chance to see the more capable doctor and, hence, patient *i*'*s* utility is $\frac{1}{2}(v_i^G + v_i^B)$, $i \in \{H, L\}$. Clearly, the common patient is always worse off when informal payments are allowed. In contrast, allowing informal payments improves the seriously ill patient's chance of seeing the more capable doctor, while costing him or her more payments as well. When $\frac{1}{2} < \frac{\Delta L}{\Delta H}$, the gain from allowing informal payments is dominated by the increment in cost, and the seriously ill patient is worse off.

The analysis of *aggregate* patient welfare depends on the tradeoff between a more efficient match and larger payments. Specifically, when $\frac{1}{3} < \frac{\Delta L}{\Delta H}$, that is, willingness to pay is not too far apart for the two patients, allowing informal payments decreases aggregate patient welfare. The welfare comparison provides a rationale for some developing countries' banning informal payments: when the social planner's goal is to maximize patient welfare, he or she should ban informal payments whenever patients do not differ much in their willingness to pay for the more capable doctors.

To summarize, allowing informal payments always improves social welfare: the more capable doctor is allocated to the seriously ill patient with a higher probability. It does not, however, always improve patient welfare. On one hand, there is a higher probability of achieving efficient allocation. One the other hand, patients have to pay more.

5. DISCUSSION

5.1. *Income constraints*

We have assumed that both patients can offer as much informal payment as they want. In reality, patients might have income constraints. In particular, patients with serious problems might not be able to offer enough informal payments to attract the more capable doctor. Suppose the seriously ill patient's income, I_H, is less than ΔL; his or her income constraint is binding.

Consider the game in Section 3 again. As before, there is no pure strategy Nash equilibrium. The unique mixed strategy Nash equilibrium now becomes:

$$F_L(x) = \begin{cases} 1 - \frac{I_H}{\Delta H} + \frac{x}{\Delta H}, & 0 \le x \le I_H. \\ 1, & x > I_H. \end{cases} \quad F_H(y) = \begin{cases} 1 - \frac{I_H}{\Delta H} + \frac{y}{\Delta L}, & 0 < y \le I_H. \\ 1, & y > I_H. \end{cases}$$

Both patients now have a positive probability of offering no informal payment. The probability of achieving the first best allocation becomes:

$$Pr(P_H > P_L) = \left(1 - \frac{I_H}{\Delta H}\right)\frac{I_H}{\Delta L} + \int_0^{I_H} \frac{I_H - x}{\Delta L}\frac{1}{\Delta H}\,dx$$

$$= \frac{I_H}{\Delta L}\left(1 - \frac{I_H}{2\Delta H}\right)\begin{cases} < \frac{1}{2}, & \text{if } 0 \le I_H < \Delta H - \sqrt{\Delta H(\Delta H - \Delta L)}; \\ \ge \frac{1}{2}, & \text{if } \Delta H - \sqrt{\Delta H(\Delta H - \Delta L)} \le I_H \le \Delta L. \end{cases}$$

Recall that with no income constraints, social welfare is always higher when informal payments are allowed. When I_H is small, which means that the seriously ill patient is poor, informal payments do not always improve efficiency. Banning informal payments in this case improves both patient and social welfare. In general, for a seriously ill patient, having a binding income constraint makes allowing informal payments less attractive.

Although we did not explicitly model horizontal equity, the argument above is in the same direction as the equity argument would go. A seriously ill patient might be poor, and might not be able to afford the informal payments that are needed to guarantee appointments with the more capable doctor. In such cases, the efficiency gain of allowing informal payments is reduced.

6. CONCLUSION

Informal payments for health care in the countries of Central and Eastern Europe and China are widespread. They are widely condemned on moral grounds and governments in those countries are urged by the public to take effective actions to ban informal payments. The present paper studies the welfare implications of allowing informal payments and banning informal payments.

Our analysis has several implications. First, whether the social planner should allow informal payments depends crucially on patient heterogeneity. When the difference in patients' willingness to pay is high, informal payments can work to improve patient welfare. When patients are willing to pay more or less the same, informal payments are futile.

Second, privatizing the public health sector, as proposed by some policy analysts, might not be a good idea. The analysts argue that in a free market of health care, price would efficiently allocate resources and social welfare is maximized. We agree with this argument but pay more attention to patient welfare, which may shrink severely in a free health-care market. This helps to explain why few countries adopt a purely private health-care system. Essentially, doctors would have strong bargaining power over their patients and if the more capable doctor were to set the price, he or she would make it as high as possible. Whether patient welfare can be improved through privatization depends again on the tradeoff between improvement in the allocation efficiency and the loss from payments.

Third, one popular view is that informal payments result from the low level of doctors' wages in the public health sector. Consequently, raising the average

doctor's wage has been proposed to eliminate informal payments. Our analysis suggests that the proposal would not succeed. An important reason for the existence of informal payments is social planners' lack of information. As long as doctors' wages do not fully incorporate patients' information, informal payments will not completely disappear.

REFERENCES

Bardhan, P. (1997) 'Corruption and Development: A Review of Issues', *Journal of Economics Literature* 35, 1320–46.
Beck, P. and M. Maher (1986) 'A Comparison of Bribery and Bidding in Thin Markets', *Economics Letter* 20, 1–5.
Bernheim, D. and M. Whinston (1986) 'Menu Auctions, Resource Allocation, and Economic Influence', *Quarterly Journal of Economics* 101, 1–31.
Biglaiser, G. and A. C. T. Ma (2007) 'Moonlighting: Public Service and Private Practice', *Rand Journal of Economics* 38, 1113–33.
Eggleston, K., L. Li, Q. Meng, L. Magnus and A. Wagstaff (2008) 'Health Service Delivery in China: A Literature Review', *Health Economics* 17, 149–65.
Garcia-Prado, A. (2005) 'An Analysis of Under-Reporting and Informal Payments among Physicians', *Departamento de Economia, Universidad Publica de Navarra Working Paper* 8, 67–98.
Gonzalez, P. (2004) 'Should Physicians' Dual Practice Be Limited? An Incentive Approach', *Health Economics* 13, 505–24.
Grossman, G. M. and E. Helpman (1994) 'Protection for Sale', *American Economic Review* 84, 833–50.
Hillman, A. and J. Riley (1989) 'Politically Contestable Rents and Transfers', *Economics and Politics* 1, 17–39.
Leff, N. (1964) 'Economic Development through Bureaucratic Corruption', *American Behavioral Scientist* 8, 8–14.
Lewis, M. (2000) 'Who Is Paying for Health Care in Europe and Central China?', *The World Bank*.
Lui, F. T. (1985) 'An Equilibrium Queueing Model of Bribery', *Journal of Political Economy* 93, 760–81.

APPENDIX

PROOF OF PROPOSITION 2. *A mixed strategy Nash equilibrium is characterized by cumulative density functions of the two patients' offers,* $F_L(x)$ *and* $F_H(y)$*, and the supports,* $[\underline{P_L}, \overline{P_L}]$ *and* $[\underline{P_H}, \overline{P_H}]$*. We prove Proposition 2 in seven steps as follows.*

Step 1. The upper bounds of the two patients' offers are the same: $\overline{P_L} = \overline{P_H} = \bar{P}$. Suppose $\overline{P_i} < \overline{P_j}$, with $i \neq j$ and $i, j \in \{H, L\}$. Then, patient j can profitably deviate to a lower upperbound $\overline{P_j}' = \overline{P_j} - \varepsilon$, with a small $\varepsilon > 0$, and put the probability of offering a payment in the range of $(\overline{P_j} - \varepsilon, \overline{P_j}]$ on $\overline{P_j} - \varepsilon$. The deviation will not reduce patient j's chance of seeing the more capable doctor but will reduce his or her expected payment.

Step 2. The upper bound is smaller than ΔL: $\bar{P} \leq \Delta L$. The common patient can always pay nothing and obtain his or her reservation utility v_L^B. If he or she offers more than ΔL, his or her utility is lower than v_L^B.

Step 3. The lower bounds of the two patients' offers are both zero. If the lower bound of patient i's offer, $\underline{P_i}$, is strictly positive. The other patient, j, would not

offer any amount in $(0, \underline{P_i})$, as $P_j = 0$ strictly dominates any offer in the interval. Given this, patient i can profitably deviate to offer $\underline{P_i'} = \underline{P_i} - \varepsilon$, where $\varepsilon > 0$ is a small number. Compared with $\underline{P_i}$, offering $\underline{P_i'}$ will not reduce patient i's probability of seeing the more capable doctor but will reduce their payment.

Step 4. Both distributions of offers are continuous. Suppose patient i makes an offer $P_i \in (0, \overline{P}]$ with probability $q > 0$ in equilibrium. The patient then makes an offer in the interval $(P_i - \varepsilon, P_i)$ with zero probability, where $\varepsilon > 0$ is a small number. Therefore, patient j also makes an offer in the interval $(P_i - \varepsilon, P_i)$ with zero probability, as any such offer is strictly dominated by $P_j = P_i - \varepsilon$. Given patient j's strategy, patient i has a profitable deviation to $P_i' = P_i - \frac{\varepsilon}{2}$.

Step 5. Any mixed strategy equilibrium is characterized by $\begin{cases} F_L(x) = \dfrac{\Delta H - \overline{P} + x}{\Delta H}; \\ F_H(y) = \dfrac{\Delta L - \overline{P} + y}{\Delta L}. \end{cases}$ In any mixed strategy Nash equilibrium, each patient must be indifferent across their offers. Therefore,

$$(v_i^G - P_i)\Pr(P_i > P_j) + (v_i^B - P_i)\Pr(P_i < P_j) = U_i,$$

where U_i is patient i's equilibrium level of utility. As a result, $F_{P_j}(x) = \dfrac{x - v_i^B + U_i}{\Delta i}$. Now impose the conditions $F_L(\overline{P_L}) = 1$ and $F_H(\overline{P_H}) = 1$. We get:

$$U_H = \Delta H + v_H^B - \overline{P}, \quad U_L = \Delta L + v_L^B - \overline{P}.$$

Therefore, the two distribution functions are $\begin{cases} F_H(x) = \dfrac{x + \Delta L - \overline{P}}{\Delta L}; \\ F_L(y) = \dfrac{y + \Delta H - \overline{P}}{\Delta H}. \end{cases}$

Step 6. At most one patient offers zero informal payment with positive probability. Suppose both patients offer zero informal payment with positive probability. Patient i can profitably deviate by setting that positive probability at $\varepsilon > 0$ instead of zero, where ε is a small number.

Step 7. The upper bound $\overline{P} = \Delta L$. From step 5, we know that $F_L(0) = \dfrac{\Delta H - \overline{P}}{\Delta H}$. Because $\overline{P} < \Delta H$, it must be that $F_L(0) > 0$. From step 6, $F_H(0)$ must be zero, which implies that $\overline{P} = \Delta L$.