April 9, 2018

A Model of Hospital Congestion in Developing Countries

Damien Besancenot*, Nicolas Sirven† and Radu Vranceanu‡

Abstract

This paper explains the observed hospital congestion in developing countries as the result of the interaction between ambulatory care physicians who refer patients to hospitals, and hospitals which must detect the severity of the incoming patients' disease. In an imperfect information environment, physicians might refer to top-tier hospitals patients with mild diseases that could be properly addressed by regular hospitals, just to fulfill patients' demand for the best care. Yet, the triage capability of top-tier hospitals declines if the hospital is subject to congestion, which, in turn, provides incentives to physicians to refer more patients to these hospitals. The model presents two equilibria, one with perfect triage, and another with triage errors and hospital congestion. In this last equilibrium, a higher hospital size raises the likelihood of congestion.

JEL Classification: I11, I15, D82
Keywords: Hospital congestion, hospital size, referral system, health policy, developing countries.

*University of Paris Descartes and LIRAES, 45 rue des Saints Pères, 75270 Paris. E-mail: damien.besancenot@parisdescartes.fr.

†University of Paris Descartes, IRDES and LIRAES, 45 rue des Saints Pères, 75270 Paris. E-mail: nicolas.sirven@parisdescartes.fr.

‡Corresponding author. ESSEC Business School and THEMA, 1 Avenue Bernard Hirsch, 95021 Cergy, France. E-mail: vranceanu@essec.fr.

1 Introduction

Large hospitals overcrowded with patients who could be treated in smaller and less sophisticated facilities is a common feature of poorly functioning referral systems in developing countries (Stefanini, 1994; Akande, 2004; Murray and Pearson, 2006; Gyedu et al., 2015; Goodman et al., 2017; Arrieta and Guillen, 2017). One reason for this situation is that the gatekeeping system is not efficient: some physicians refer patients to top-tier (or referral) hospitals to meet patients' preference for the highest quality of care, even if this referral is not medically justified (Donohoe et al., 1999; Ringberg et al., 2014). Patients in developing countries have a perception of a better quality of health care in referral hospitals (Kowalewski, 2000; Juliani et al., 2017), and considerable patient leeway in deciding the type and level of health care facility care from which to seek treatment (Mwabu 1989; Hajjaj et al. 2010). Benevolent physicians might meet patients' calls for "abnormal care" under the stress of competition with alternative providers of medical care, including traditional healers (Hammer & Jack, 2002). In practice, a formal referral document is often unrequired (Goodman et al., 2017), or incomplete (Eskandary et al., 2013; Gyedu et al., 2015), in some cases the gatekeeper does not even issue a referral form, which leads to patients' self-referrals. Quite often patients bypass the lower levels of health care and present themselves to higher levels, regardless of the gravity of their medical situation (Abdi, 2015; Nshimirimana et al., 2016). As an upshot of these considerations, Kapoor et al. (2017) argue that "[an] effective referral/counter-referral system is essential to resource-limited health care systems to maximize efficiency, decrease unnecessary resource utilization, and provide safe and timely patient care within a tiered system. The challenges of successfully developing such a system are multifactorial since limited resources, multiple layers of communication, education, and relationships among providers and patients are all involved."

To the best of our knowledge, there is no theoretical framework for the documented hospital congestion in the context of developing countries where the referral system is poorly functioning and where perceived quality of care is higher in referral hospitals. This paper develops a model to analyze the interaction between physicians (ambulatory care providers) in their referral role,

patients, and hospitals. We consider a standard assumption regarding the two-tier organization of the health care system in the developing countries, featuring top-tier (or referral) hospitals with high quality of equipment and high staff expertise, and primary care (or regular) hospitals at the district or regional level (Jack, 1999; Zakus and Bhattacharyya, 2007). In the model, an acutely ill patient is always referred to the top-tier hospital to benefit of the highest level of care. However, as a manifestation of a poorly functioning gatekeeping system, physicians can also send patients with less than severe conditions to top-tier hospitals, just to fulfill their demand for the best level of care. Hospitals have some triage capacity, i.e. medical procedures to determine the priority of patients' treatments based on the severity of their condition. This in-hospital diagnostic is a mechanism intended to reveal asymmetric information from the ambulatory care physician who made the referral , or to directly asses the patients' health in case of self-referral. The efficiency of the triage mechanism depends on the number of arriving patients. The higher congestion, the higher the likelihood that patients with mild conditions will be accepted by the hospital (i.e., admission of a false positive), a phenomenon also emphasized by Alizamir et al. (2013). Congestion and screening errors can lead some patients with severe diseases to be redirected towards regular hospitals that will not be able to deliver the needed care. Furthermore, it is these errors due to congestion which create the incentive for inappropriate referral behavior, in a self-fulfilling mechanism.

One original contribution of the paper is the modelling device used to represent the patients' in-hospital triage, i.e. the process of determining the priority of patients' treatments based on the severity of their condition at the hospital level. Instead of using a microeconomic model of the diagnostic process, we adopt a more parsimonious statistical perspective, and assume that the hospital aiming at an efficient use of its care capacity accepts patients with a probability equivalent to the frequency of severe condition in the set of referred patients. Any other assumption would involve that the hospital rejects patients with severe illness while hospital beds are unoccupied, or that it accepts more patients than allowed by its treatment capacity.[1]

Our paper contributes to the existing literature on the efficiency of health care systems in developing countries in two main ways. First, we provide an original explanation of top-tier

[1] A similar modelling device was used by Besancenot and Vranceanu (2005) as applied to labor courts in France.

hospitals congestion as the result of the rational behavior of the physician in a context where hospitals can implement only an imperfect triage mechanism. The existing literature (Nardo and Juliani, 2012; Kapoor et al., 2017; Juliani et al., 2017) has already revealed that technical issues that impede referral communication (lack of computerized information, non-standardized referral form, etc.) nurture poor hospital triage. We develop the alternative and complementary argument that poor triage is explained by hospital congestion, itself driven by physicians' rational decision to sent patients with light diseases to top-tier hospitals. We assume the physician's referral decision is driven by a combination of medical concern (the probability that the patient is accepted in a top-tier hospital) and his expected payoff maximization, essentially based on the gratitude of his patient. Whilst consistent with the stylized facts that we presented beforehand, this assumption has been neglected so far by existing literature, maybe because of the underlying ethical challenges associated with the physicians' behavior as modelled in this paper.

Second, our model sheds its own light on the measures needed to fight hospital congestion. So far, hospital congestion has been explained by the inaccurate decision in the choice of non-substitutable inputs that have, at least temporarily, reached full capacity (Brailer, 1992). In this context, congestion can be relieved by increasing the binding inputs in the production function. On the other hand, in our model hospital congestion is the consequence of the emergence of a "bad equilibrium", itself grounded in the features of the referral system and the imperfection of the patient diagnostic mechanism under congestion. Our results indicate that by increasing the hospital binding inputs (e.g., increase the number of beds) may actually exacerbate congestion as a result of increased likelihood to get more false positive patients admitted to the top-tier hospital. Alternative options to lower congestion include increasing the hospital triage capacity (i.e. producing a second diagnostic, and possibly counter-refer patients to second-tier health care facilities), which has been found elsewhere to have a beneficial effect beyond that which could be achieved by redeploying resources for gatekeeping at lower-level referring facilities (Freeman et al. 2017).

The paper is organized as follows: Section 2 develops the main assumptions of the model. Section 3 presents the two equilibria of the game and provides a discussion of their implications.

The last section concludes the paper.

2 Section 2: The model

We consider a population of ambulatory care physicians (hereafter ACP or simply 'physicians') of dimension one. Each physician is supervising one patient that must be sent to a hospital for the appropriate treatment. There are two types of patients: the $S-$type (S for "Severe") are patients with severe illnesses that need to be addressed only by the top-tier hospitals (advanced equipment, highly qualified medical teams) and the $M-$type (M for "Mild" or "Moderate") are patients with lighter illnesses that can be treated in any regular hospital. In order to simplify the presentation, hereafter we will refer to as a $S-$type physician (resp. $M-$type) any ACP supervising a $S-$type (resp. $M-$type) patient. The frequency of $S-$type patients (resp. $M-$type) in the general set of patients is denoted by α (resp. $1-\alpha$). With a population of physicians of dimension one, α is also the number of $S-$type patients in the total patient population.

We assume that patients' pathology is sufficiently severe to require hospital care, regardeless of their type. The physician who perfectly knows the type of his patient has to refer his patient to one of the two available hospitals. He may refer this patient to a top-tier hospital ($T-$strategy) or send him to a regular one ($R-$strategy). To avoid unnecessary complexity, we make the plausible assumption according to which a physician supervising a patient with a severe illness will always refer him to a top-tier hospital, i.e. a $S-$type physician always plays the $T-$strategy.[2]

Regular hospitals have no capacity constraint. They can accept all incoming patients. Thus, the $R-$strategy always leads to an immediate hospitalization. The ACP referring a $M-$type patient to the regular hospital reaches the certain reward W_R. This reward encompasses the gratitude of his patient, the good reputation of the physician, etc...

The treatment capacity of the top-tier hospital is limited. In particular, it is limited by the number of hospital beds b. We assume that this number of beds is strictly equal to the number of $S-$type patients, $b = \alpha$.[3] In other words, if the matching between patients and hospitals

[2] This assumption has no effect on the results as, for a S-type physician, the R-strategy is dominated by the S-strategy.

[3] This assumption will be discussed at the end of the paper.

were perfect, all $S-$type patient would find a place in the top-tier hospital, and no bed would be empty. However, if the number of incoming patients to the top-tier hospital exceeds α, the available beds will not be sufficient and, after a diagnostic phase, some of them will be redirected to regular hospitals. Thus the ex-post reward for a physician playing the $T-$strategy relies on the first-diagnostic and medical decision of his colleagues in the top-tier hospital.

If the patient is accepted, he will get the best possible care and his gratitude to his ACP will be high, leading in the case of the moderate disease to a reward W_T for the ACP, with $W_T > W_R$ (a $S-$type ACP with a patient accepted in the $T-$hospital obtains S_T). However, if the patient's pathology is perceived as too benign for the top-tier hospital and, accordingly, the patient is redirected to a regular hospital (a.k.a. a counter-referral procedure), the physician will have a cost to pay for his unsuccessful recommendation (related to a partial loss of reputation, the time spend to build a unconvincing medical file, the anger of the patient who wasted his time before getting the proper care, etc...). Let us denote by c the cost paid by a physician in case of patient redirection, the reward for a $M-$type ACP in case of redirection is therefore $W_R - c < W_T$ (the reward of a $S-$type physician in case of a patient redirection is $S_R - c$ as the patient suffering from an acute disease will not receive the appropriate care).

The triage mechanism that leads to the hospitalization in the top-tier hospital is at the heart of this paper. We assume that the precision of the patients' selection by the $T-$hospital is decreasing with the number of patients referred to it. When this number remains low, the specialists can implement all medical tests requested by an in-depth examination of the patient pathology and the diagnostic is perfect. Only patient with severe diseases are accepted. However, when this number rises, due to a shortage of time and resources, the perfect diagnostic becomes impossible and the hospital may make triage mistakes, accepting some $M-$type patients and refusing $S-$type ones.

Let us denote by $\varphi < 1/2$, the maximum number of patients for whom the top hospital can set a perfect diagnostic. In the following, we restrict our analysis to the non trivial case where $\alpha \leq \varphi$. If only patients with a severe illness are sent to the top-tier hospital, the latter can perfectly assess their medical situation.

With these assumptions, the model can be cast as a game featuring two players: the physician,

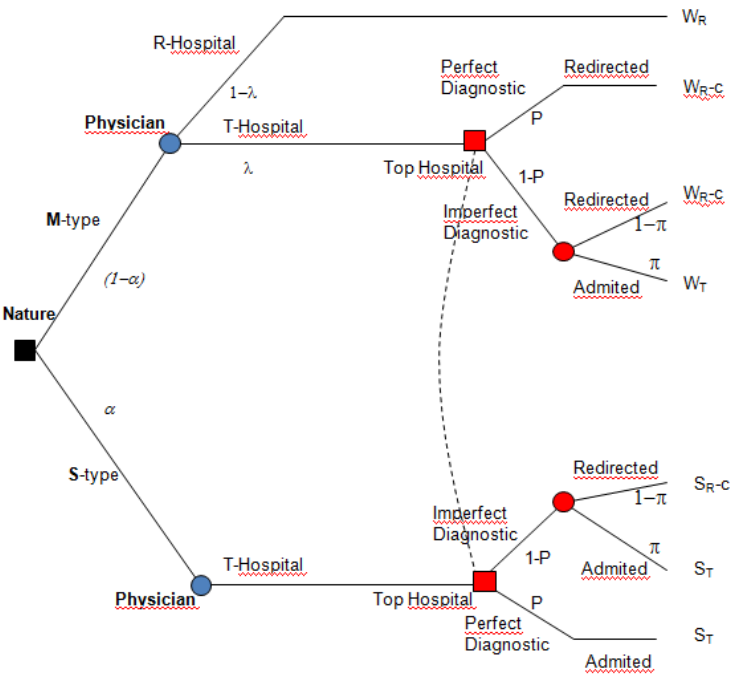

Figure 1: Decision Tree

who has to decide to which type of hospital he will refer his patient, and the hospital, which decides to admit or not the patient. The typical sequence of decisions is the following (Figure 1 presents the decision tree):

Time 1. Nature chooses the type of patient and assigns it to a ACP.

Time 2. The $M-$type ACP chooses to which of the two types of hospital he will refer his patient ($S-$type physicians have no choice but to refer their patient to the best hospital). He can recommend his patient to the regular hospital ($R-$strategy) and he therefore receives W_R (in this case, the game is over), or he may send it to the top hospital ($T-$strategy); in this case, his reward depends on the subsequent decision of the hospital.

Time 3. Specialist physicians at the top-tier hospital assess the patient's degree of illness and make a decision whether to accept the patient or to redirect him to the regular hospital.

Let us denote by n the number of patients that the hospital has to assess (by definition, $n \geq \alpha$), by p the probability of perfect diagnostic by the hospital and by π the probability of admission in case of imperfect diagnostic. The probability for one patient to see his pathology assessed

perfectly is: $p = \min\{\varphi/n, 1\}$. When $n < \varphi$, each patient's pathology is perfectly identified, $p = 1$, patients with a severe illness are admitted (their ACPs receive S_T) and $M-$type patients will be redirected toward the regular hospital (ACPs receive $W_R - c$).

When $n > \varphi$, congestion occurs. A number φ of patients will get the right diagnostic, but the hospital's physicians may underestimate or overestimate the severity of the illness of the remaining $(n - \varphi)$ patients. Let us denote by λ the probability that a $M-$type ACP plays the $T-$strategy and by $(1 - \lambda)$ the probability of the $R-$strategy. Given that all $S-$type patients are send to the top hospital, the number of patients managed by this hospital is given by $n = \alpha + (1 - \alpha)\lambda$ and the probability for a $S-$type patient to receive a perfect diagnostic is:

$$p = \min\{\varphi/n, 1\} = \min\left\{\frac{\varphi}{\alpha + (1-\alpha)\lambda}, 1\right\}. \tag{1}$$

For the sake of simplicity, we assume that when a patient is subject to imperfect assessment, he will be admitted by the hospital with a probability π equal to the frequency of $S-$type patients in the set of patients referred to the top hospital (a value also equal to the observed frequency of admission in the set of the perfectly assessed patients):

$$\pi = \frac{\alpha}{\alpha + (1-\alpha)\lambda}. \tag{2}$$

The overall probability of admission for a $S-$type patient, noted Π_S is thus:

$$\begin{aligned} \Pi_S &= p + (1-p)\pi \qquad (3) \\ &= \min\left\{\frac{\varphi}{\alpha + (1-\alpha)\lambda}, 1\right\} + \left\{1 - \min\left[\frac{\varphi}{\alpha + (1-\alpha)\lambda}, 1\right]\right\}\left(\frac{\alpha}{\alpha + (1-\alpha)\lambda}\right). \end{aligned}$$

And, for a $M-$type patient the probability of admission, noted Π_M, is:

$$\begin{aligned} \Pi_M &= (1-p)\pi \\ &= \left\{1 - \min\left[\frac{\varphi}{\alpha + (1-\alpha)\lambda}, 1\right]\right\}\left(\frac{\alpha}{\alpha + (1-\alpha)\lambda}\right). \end{aligned} \tag{4}$$

The admission rates Π_S and Π_M can be seen as an extremely simplified representation of the admission process. However, despite its simplicity, this modelling approach presents two interesting properties that properly matches with effective hospital triage rules.

First, our assumption allows to study the impact of an increasing number of $M-$type patients sent to the top hospital on the efficiency of admission choices. According to Eq. (3) and (4), the probability of admission is always higher for $S-$type patients i.e., $\Pi_S > \Pi_M$. For low values of n ($n \leq \varphi$) there is no congestion and we have $\Pi_S = 1$ and $\Pi_M = 0$. In the case of congestion ($n > \varphi$), hospital specialists are compelled to rely on a lower number of complementary examinations or a lower diagnostic time and should accept a lower precision in the medical diagnosis. In this case, the admittance of $M-$type patients as well as the rejection of $S-$type patients can both be feasible. As long as p remains close to 1, the probability of accepting a $M-$patient or of rejecting a $S-$patient are low. However, it is easy to check that an increase in the number of $M-$type patients sent to the hospital reduces the probability of admission for $S-$type patients ($d\Pi_S/d\lambda \leq 0$) and fosters the chances of admission for $M-$types patients ($d\Pi_M/d\lambda \geq 0$ as long as $\lambda > (2\varphi - \alpha)/(1-\alpha)$).

Second, the definition of π allows a constant number of hospitalization. Given that, by assumption, $b = \alpha$, it leads to a number of inpatients strictly equal to the number of bed. Any other definition of π would lead to an excess number of beds or an excess number of hospitalized patients and would reveal a bias in favour or against admission. With an admission rate higher than π, the hospital must acknowledge that it will accept $M-$type patients and, on the opposite, an admission rate lower than π implies that the hospital knows for sure that it will reject patients with severe diseases. When π is equal to the frequency of the $S-$type patients in the set of patients managed by the $T-$hospital, the latter provides a bed for each of the sorted patients and displays some rationality in his triage procedure.

Finally, the expected payoffs $E[U_S|j-strategy]$ for a $S-$type physician playing the $T-$strategy are :

$$E[U_S|T-strategy] = S_T - (1-\pi)(1-p)(S_T - S_R + c) \tag{5}$$

and the expected payoffs for a $M-$type physician are given by:

$$\begin{cases} E[U_M|R-strategy] = W_R \\ E[U_M|T-strategy] = (W_R - c) + (1-p)\pi(W_T - W_R + c) \end{cases}. \tag{6}$$

2.1 Section 3: Equilibria

An equilibrium of this game corresponds to a situation where M-type ACPs choose their optimal referral strategy given the likelihood that patients are admitted in the top-tier hospital, and the likelihood that patients are admitted in the top-tier hospital correctly reflects the best use of the triage mechanism given the number of arriving patients.[4]

The set of equilibria of the game includes a separating, a pooling, and a hybrid equilibrium. Because the pooling equilibrium appears to be a special case of the hybrid one, it will be presented in the Appendix.

2.1.1 Separating equilibrium

In this equilibrium, ACPs in charge of patients with a severe disease send them to the top-tier hospital, while ACPs in charge of patients with moderate diseases refer them to the regular hospital. Thus, $n = \alpha \leq \varphi$ and $p = 1$. In this equilibrium, the severity of every pathology is perfectly identify by the top-tier hospital.

For a physician in charge of a $M-$type patient, the expected rewards are $E[U_M|R-strategy] = W_R$ in case of the $R-$strategy and $E[U_M|T-strategy] = (W_R - c)$ in the other case. For such a physician, the $R-$strategy is optimal as sending his patient to the top-tier hospital would lead to an automatic redirection.

Remark that this equilibrium is always feasible under the sufficient condition: $n \leq \varphi$, i.e. if $p = 1$. This condition states that, as long as the hospital is able to have a perfect assessment of the patients, any $M-$type patient would be rejected by the top hospital. There is no incentive for a $M-$type physician to send such a patient to the top hospital. There is no congestion, patients selection process is efficient and the top hospital receives and treats only $S-$type patients. Unfortunately, this is not the only equilibrium of our model.

2.1.2 Hybrid equilibria

The model also presents an equilibrium in which $M-$type ACPs are indifferent between the two strategies, so they randomly adopt the $M-$ or the $T-$strategy. In this equilibrium, a fraction, $\lambda \in$

[4] We recall that $S-$type ACPs always play the $T-$strategy.

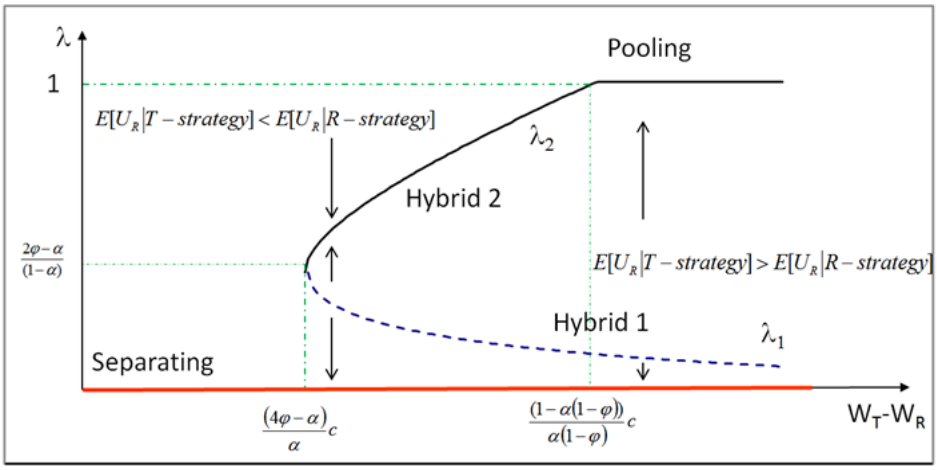

Figure 2: Possible equilibria depending on parameters

$[0,1]$, of the $M-$physicians send their patients to the top hospital. For a $M-$type physician, the indifference between the two strategies implies $E[U_M|R-strategy] = E[U_M|T-strategy]$. Given Eq. (6), we have:

$$W_R = (W_R - c) + (1-p)\pi\left(W_T - W_R + c\right). \tag{7}$$

Note that this condition can only be true if $p < 1$, i.e. if $n > \varphi$. After some calculations, Eq. (7) can be stated as:

$$cX^2 - VX + \varphi V = 0 \text{ with } \begin{cases} X = \alpha + (1-\alpha)\lambda \\ V = \alpha\left(W_T - W_R + c\right) > 0 \end{cases}.$$

If $V > 4c\varphi$, the former equation presents two real roots $X_{1,2}$ which implies two solutions for λ

$$X_{1,2} = \frac{V \pm \sqrt{V^2 - 4c\varphi V}}{2c} \Leftrightarrow \begin{cases} \lambda_1 = \frac{V - 2c\alpha - \sqrt{V^2 - 4c\varphi V}}{2c(1-\alpha)} \\ \lambda_2 = \frac{V - 2c\alpha + \sqrt{V^2 - 4c\varphi V}}{2c(1-\alpha)} \end{cases}, \text{ with } \lambda_1 < \lambda_2 \tag{8}$$

The Appendix at the end of the paper studies the parameters values under which the two roots are defined in the range $[0,1]$. For convenience, we will refer to Hybrid equilibrium 1 or 2 according to the definition of the frequency λ.

- Figure 2 presents the various equilibria according to the values of $(W_T - W_R)$.

In Figure 2, the red line on the horizontal axe indicates that a Separating equilibrium is always feasible regardless of the remuneration spread $(W_T - W_R)$.

The two curves representing λ_1 and λ_2 as functions of $(W_T - W_R)$ define the locus where a $M-$type physician is indifferent between the two strategies ($E[U_M|T - strategy] = E[U_M|R - strategy]$). For $(W_T - W_R) = 0$, it is easy to check that a physician in charge of a patient with a moderate disease would find no interest to send him to the top-tier hospital. This implies that each point on the left of the (λ_1, λ_2) curve describes a situation where $E[U_M|T-strategy] < E[U_M|R-strategy]$ and that points on the right of the curve imply opposite preferences. Intuitively, if for any point to the left of the frontier the $M-$type physicians strictly prefer the $R-$strategy, λ should decrease (descending arrows in Figure 2); symmetrically, for any point on the right of the (λ_1, λ_2) curve, λ should increase (ascending arrows) as $M-$type physicians strictly prefer to send their patient to the top hospital. As a consequence, if the Hybrid 2 equilibrium is stable, the Hybrid 1 equilibrium is not. A small deviation from this equilibrium would degenerate either toward the Hybrid 2 or to the Separating equilibrium with $\lambda = 0$.

In brief, the model exhibits two quite different equilibria: a separating equilibrium and a hybrid one where the number of $M-$type patients with light diseases referred to the top-tier hospital is positive and increases with the spread $(W_T - W_R)$; for high values of this spread, the latter equilibrium may degenerate in a pooling equilibrium where all $M-$type patients are send to the top-tier hospital.

In the separating equilibrium, physicians in charge with patients with a moderate disease correctly foresee that they will be rejected by the top hospital and therefore prefer to send them directly to the regular hospital. This avoids their patients the deception of a counter-referral procedure. In this case, the set of patients arriving at the referral hospital comprises only patients with the most severe conditions. The $T-$hospital receives a low number of patients ($n = \alpha$) and may correctly assess each patient's pathology. Triage is perfect as expected by physician. Finally, each $S-$type patient finds a bed in the top hospital ($n = \alpha = b$), patient allocation is efficient. This equilibrium exists regardless of the values of the parameters.

On the opposite, when physicians believe that patients' triage may be imperfect, an ACP in charge of a $M-$type patient may play the $T-$strategy with the hope that his patient will be accepted by the hospital. As an increase of the number of patients induces the deterioration of the

triage process, the drop in the efficiency of the selection induces more physicians with $M-$type patients to send them to the top-tier hospital, a process that contributes to the congestion of the hospital and fulfils the physicians' expectations.

Taking a public health policy perspective, the separating equilibrium is "good", as the all patients are referred to hospitals that can successfully address their case, while the hybrid equilibrium is "bad", as some patients will not receive the appropriate care, and top-tier hospitals are subject to congestion. The separating equilibrium, at work in many rich countries, can thus be considered as an ideal, or target organization of the health care system for the developing countries, where, in many cases, the health care system is stuck in the bad hybrid equilibrium.

2.1.3 Discussion of the hybrid equilibrium and policy implications

Facing an increase in the number of patients referred by their physicians, the top-tier hospital might try to address the congestion problem by restricting the selection burden. This can be achieved through an appropriate choice of parameters values c, φ and b.

First, an increase in the cost c paid by the physicians in the case of redirection reduces the incentive to play the $T-strategy$. For instance, an increase in the delay before the first consultation which increases the time lost before access to health care rises the cost for redirected patients and therefore reduces the reward of their physician. However, the cost paid by the patients with less serious conditions will also be paid by patients with severe pathologies. A policy that reduces the efficiency of the health care system and would imply the death of some patients with severe diseases cannot be implemented.

An important element of the hybrid equilibrium is that a switch from the separating to the hybrid equilibrium requires a minimum of $M-$type strategies. In the model, this switch implies a jump for λ from zero to a value above λ_1.[5] Of course, such a change implies an implicit coordination of the $M-$type physicians who must collectively decide to send their patient to the top-tier hospital. Given the definition of λ_1, this jump is more difficult to achieve when φ takes higher values. An increase in the triage capacities of the top-tier hospital would thus contain the

[5] Note that a jump from 0 to a value above λ_1 is sufficient to engage the convergence toward λ_2.

possibility of a shift from the separating to the hybrid equilibria. On the other side, an increase in the spread between the physicians' reward makes the jump easier as the threshold λ_1 is a decreasing function of $(W_T - W_R)$.

The latter proposition can be illustrated by the fact that expectations in health care responsiveness (i.e. the quality of care delivered) increase with living standards (Robone, Rice, and Smith, 2011; Deaton and Tortora, 2015). Throughout the countries' development process, the demand for higher quality of care (T-strategy) would thus result in the increase of the spread, ceteris paribus. This would suggest that the richer the country, the higher the likelihood of hospital congestion, and the higher the need to improve triage. However, from a historical perspective, successful health care systems in developing countries managed to limit some of these undesired outcomes (e.g. hospital congestion) through improved regulatory and managerial capacity (e.g. better referral system) and investments in health care capacity (Balabanova, McKee, and Mills, 2011; Mills, 2014).

The model is built under the assumption $\alpha \leq \varphi$, i.e., the top-tier hospital is able to perfectly assess the pathology of each $S-$type patients. In the opposite case, when $\alpha > \varphi$, the number of $S-$type patients exceeds this perfect diagnostic capability and the probability for the triage mechanism to be fully efficient is structurally less than one, $p < 1$. This introduces a possibility of mistakes in the measure of illness severity. In this case, one $M-$type physician may expect that his patient will benefit from this mistake and send him to the top hospital. In turn, the presence of patient with less severe pathology in the set of the top-tier hospital patients is sufficient to generate some triage mistakes with a positive probability of admitting a $M-$type patient, $\Pi_M > 0$ and a positive probability of counter-referral for a $S-$type patient to the regular hospital, $\Pi_M < 1$. As the triage procedure may fail, it is easier for a $M-$type practitioner to consider the possibility of a deviation from the $R-$strategy and the hybrid equilibria may more easily appear.[6]

Another important issue to address deals with the number of beds in the top-tier hospital. In the model, we considered a number of beds strictly equal to the number of $S-$type patients.

[6] Formally, with $p < 1$, for any positive value of λ there exists a critical value $(W_T - W_R)^c$ above which $E[U_M|T - strategy] < E[U_M|R - strategy]$. For a high enough reward W_T, the separating equilibrium desappears.

Thus, in the separating equilibrium, only patients with a severe condition obtain a bed in the $T-$hospital. This will not be the case if there are more beds than $S-$type patients.

A top-tier hospital that could choose between the two types of mistakes, i.e., between accepting $M-$type patients or rejecting $S-$type ones, would probably prefer to avoid the second option that entails dramatic consequences for the patients. For so doing, it might strive to increase the number of beds above the number of $S-$type patients ($b > \alpha$) and accept more patients regardless of their type. However, if for narrow economic reasons the health authorities require the top-tier hospital to have a full utilization of its care capacity, a number of beds such as $b > \alpha$ implies that the hospital must accept $M-$type patients. Whatever the equilibrium, ambulatory care physicians in charge of patients with moderate pathologies are therefore encouraged to refer patients to the top-tier hospital. Thus, by definition, $\lambda > 0$ and the possibility of the separating equilibrium is logically excluded. Unless if triage capacities are perfect (which is implausible with a high number of incoming patients) this nurtures triage errors, generates an incentive for practitioners to send their $M-$type patients to the top-tier hospital and may induce hospital congestion. Paradoxically, an oversized hospital system has more chance to be subject to congestion than one with the right capacity.

3 Conclusion

Our analysis concurs with previous in-depth qualitative analysis of health systems in developing countries that revealed major deficiencies in the referral systems (Stefanini, 1994; Akande, 2004; Gyedu et al., 2015; Goodman et al., 2017; Arrieta and Guillen, 2017). We provide an alternative and complementary explanation of hospital congestion in this region that underlines the consequences of the imperfect triage mechanism under "equilibrium congestion", rather than the hospital under-capacity due to a shortage of inputs as suggested by existing literature (Brailer, 1992).

The model features multiple equilibria, including a "good equilibrium" with efficient separation of patient types and patient referral, and "bad equilibrium" characterized by congestion, and a self-fulfilling motivation of physicians to send patients with mild conditions to top-tier hospitals.

The latter is the outcome of physicians sending patients suffering from light illnesses to top-tier hospitals, with the resulting hospital congestion leading to a large error margin of the triage mechanism and a higher chance that these patients be admitted in top-tier hospitals. If admitted, they will benefit of the excellent care system of these hospitals, even if this is not justified from a mere medical perspective. A more dramatic outcome of congestion are type-2 errors, where some patients with severe diseases can be wrongly redirected toward the regular hospitals.

The literature surveyed in the introduction pointed out the extreme congestion of hospitals in the developing countries, suggesting that in these countries the hospital care system might be trapped in the bad hybrid equilibrium. For policymakers in these countries, the target outcome is the separating equilibrium at work in many developed countries. Unfortunately, switching from one equilibrium to another is a challenging endeavour, both in theory, and more so in practice.

However, our theoretical model suggests several ways to reduce hospital congestion and improve the referral system in developing countries. It turned out that increasing treatment capacity in top-tier hospitals to reduce congestion would probably trigger the opposite effect, or at least maintain congestion, as it rises chance for patients with moderates illnesses to be admitted in the facility. Thus, efforts should be dedicated to improve the triage capacity of the top-tier hospital, for instance by developing clinical decisions units (CDUs) attached to the emergency department of the hospital, and to which patients can be referred for a more detailed gatekeeping decision. This "second gatekeeping" has been found to have a beneficial effect beyond that which could be achieved by redeploying resources for gatekeeping at lower-level referring facilities in the UK (Freeman et al., 2017).

4 Appendix

This appendix defines the conditions for which λ may exist in the range $[0, 1]$ and studies the influence of the relevant parameters on the frequency λ.

Note first that, in the model, the $T-$strategy may only be considered by a $M-$type practitioner if the condition $p < 1$ is fulfilled. This requires $\varphi < \alpha+(1-\alpha)\lambda$ and induces the necessary condition : $\lambda > \frac{\varphi-\alpha}{(1-\alpha)} > 0$. It is easy to check that this condition is satisfied with any of the two λ values if:

$$\lambda_1 > \frac{\varphi - \alpha}{(1-\alpha)} \Leftrightarrow V - 2c\varphi - \sqrt{(V-2c\varphi)^2 - (2c\varphi)^2} > 0 \tag{9}$$

Therefore, the two roots are positive under the sufficient condition $V - 2c\varphi > 0$. As the two roots exist if $V > 4c\varphi$, they are always positive when the value of the parameters allows their existence. We must now check under which condition these roots are lower than one.

- $\lambda_1 \in [0, 1]$

By definition, $\lambda_1 < 1$ implies $V - 2c - \sqrt{V^2 - 4c\varphi V} < 0$. Remark that if $V - 2c \leq 0$, the inequality $\lambda_1 < 1$ is always true. In this case, as $V > 4c\varphi$, we must have $4c\varphi < V < 2c$. Given the definition of V, the probability λ_1 is therefore defined in the range $[0, 1]$ if the following condition is satisfied (by assumption $\varphi < 1/2$):

$$c\frac{4\varphi - \alpha}{\alpha} < (W_T - W_R) < c\frac{2-\alpha}{\alpha} \tag{10}$$

On the other hand, when $V - 2c \geq 0$, condition $V - 2c - \sqrt{V^2 - 4c\varphi V} < 0$ can be rewritten as $V - \frac{c}{(1-\varphi)} > 0$ which is always true if $V - 2c > 0$. The probability λ_1 is therefore defined in the range $[0, 1]$ when the following condition is satisfied :

$$(W_T - W_R) \geq c\frac{2-\alpha}{\alpha} \tag{11}$$

Finally, mixing the two previous condition, we get the necessary and sufficient condition [NSC1] for $\lambda_1 \in [0, 1]$:

$$c\frac{4\varphi - \alpha}{\alpha} < (W_T - W_R) \tag{NSC1}$$

In this range of parameters, differentiation of λ_1 in Eq. (8) allows to check that an increase of the cost c and that a reduction in the spread $(W_T - W_R)$ would increase the frequency of the $R-$strategy:

$$\frac{d\lambda_1}{dc} = \frac{(1-\alpha)\lambda_1(\alpha + (1-\alpha)\lambda_1) + \alpha\varphi}{(1-\alpha)\sqrt{V - 4c\varphi V}} > 0 \tag{12}$$

$$\frac{d\lambda_1}{d(W_T - W_R)} = -\frac{\alpha}{2c(1-\alpha)}\left(\frac{V - 2c\varphi - \sqrt{V - 4c\varphi V}}{\sqrt{V - 4c\varphi V}}\right) < 0 \tag{13}$$

- $\lambda_2 \in [0,1]$

By definition, condition $\lambda_2 < 1$ is equivalent to $V - 2c + \sqrt{V^2 - 4c\varphi V} < 0$, a condition that can be satisfied only if $(V - 2c)$ is negative and $2c - V > \sqrt{V^2 - 4c\varphi V}$. As this condition is equivalent to $V < \frac{c}{1-\varphi}$, and given that λ_2 exists under the condition $V > 4c\varphi$, λ_2 is therefore defined in the range $[0,1]$ if $4c\varphi < V < \frac{c}{1-\varphi}$, i.e. if the following condition is satisfied :

$$c\frac{4\varphi - \alpha}{\alpha} < (W_T - W_R) < c\frac{2-\alpha}{\alpha} \tag{NSC2}$$

In this equilibrium an increase of the cost c would induce a drop of the $M-$type submissions, and a rise of the spread $(W_T - W_R)$ produces the opposite result:

$$\frac{d\lambda_2}{dc} = -\frac{(1-\alpha)\lambda_2(\alpha + (1-\alpha)\lambda_1) + \alpha\varphi}{(1-\alpha)\sqrt{V - 4c\varphi V}} > 0 \tag{14}$$

$$\frac{d\lambda_2}{d(W_T - W_R)} = \frac{\alpha}{2c(1-\alpha)}\left(\frac{V - 2c\varphi - \sqrt{V - 4c\varphi V}}{\sqrt{V - 4c\varphi V}}\right) > 0 \tag{15}$$

Remark that $\lambda_2 = 1$ when $(W_T - W_R)$ reaches the upper bound. Thus the Pooling equilibrium appears to be a special case of the Hybrid 2 equilibrium. For a low cost c or an important difference between the rewards W_T and W_R, congestion reaches its maximum.

5 References

References

Abdi, W. O., Salgedo, W. B., & Nebeb, G. T., (2015). Magnitude and determinants of self-referral of patients at a general hospital, Western Ethiopia. *Science*, 4 (5), 86-92.

Akande, T. M. (2004). Referral system in Nigeria: study of a tertiary health facility. *Annals of African Medicine* 3 (3): 130-133.

Alizamir, S., De Véricourt, F., & Sun, P., (2013). Diagnostic accuracy under congestion. *Management Science*, 59 (1), 157-171.

Arrieta, A., & Guillén, J., (2017). Output congestion leads to compromised care in Peruvian public hospital neonatal units. *Health Care Management Science*, 20 (2), 157-164.

Balabanova, D., McKee, M., & Mills, A. (2011). *Good Health at Low Cost 25 Years on. What Makes a Successful Health System?*. London, United Kingdom, London School of Hygiene and Tropical Medicine.

Besancenot, D. & Vranceanu, R., (2009), Multiple equilibria in a firing game with impartial justice, *Labour Economics*, 16 (2), 262- 271

Brailer, D. J., (1992). A theory of congestion in general hospitals. University of Pensilvania, *mimeo*.

Deaton, A. S., & Tortora, R. (2015). People in Sub-Saharan Africa rate their health and health care among the lowest in the world. *Health Affairs*, 34 (3), 519-527.

Donohoe, M. T., Kravitz, R. L., Wheeler, D. B., Chandra, R., Chen, A., & Humphries, N., (1999). Reasons for outpatient referrals from generalists to specialists. *Journal of General Internal Medicine*, 14 (5), 281-286.

Eskandari, M., Abbaszadeh, A., & Borhani, F., (2013). Barriers of referral system to health care provision in rural societies in Iran. *Journal of Caring Sciences*, 2 (3), 229.

Freeman, M., Robinson, S., & Scholtes, S., (2017). Gatekeeping under Congestion: An Empirical Study of Referral Errors in the Emergency Department. *INSEAD Working Paper* No. 2017/59/TOM, mimeo.

Goodman, D. M., Srofenyoh, E. K., Olufolabi, A. J., Kim, S. M., & Owen, M. D., (2017). The third delay: understanding waiting time for obstetric referrals at a large regional hospital in Ghana. *BMC Pregnancy and Childbirth*, 17 (1), 216.

Gyedu, A., Baah, E. G., Boakye, G., Ohene-Yeboah, M., Otupiri, E., & Stewart, B. T., (2015). Quality of referrals for elective surgery at a tertiary care hospital in a developing country: an opportunity for improving timely access to and cost-effectiveness of surgical care. *International Journal of Surgery*, 15, 74-78.

Hajjaj, F. M., Salek, M. S., Basra, M. K., & Finlay, A. Y., (2010). Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice. *Journal of the Royal Society of Medicine*, 103(5), 178-187.

Hammer, J., & Jack, W., (2002). Designing incentives for rural health care providers in developing countries. *Journal of Development Economics*, 69 (1), 297-303.

Jack, W. (1999). Principles of health economics for developing countries. World Bank Publications.

Juliani, C., MacPhee, M., & Spiri, W., (2017). Brazilian Specialists' Perspectives on the Patient Referral Process. *Healthcare*, 5 (1): 4.

Kapoor, R., Avendaño, L., Sandoval, M. A., Cruz, A. T., Sampayo, E. M., Soto, M. A., ... & Crouse, H. L., (2017). Initiating a Standardized Regional Referral and Counter-Referral System in Guatemala: A Mixed-Methods Study. *Global Pediatric Health*, 4 : 1-14.

Kowalewski, M., Jahn, A., & Kimatta, S. S., (2000). Why do at-risk mothers fail to reach referral level? Barriers beyond distance and cost. *African Journal of Reproductive Health*, 4 (1), 100-109.

Mills, A. (2014). Health care systems in low-and middle-income countries. *New England Journal of Medicine*, 370(6), 552-557.

Murray, S. F., & Pearson, S. C., (2006). Maternity referral systems in developing countries: current knowledge and future research needs. *Social Science & Medicine*, 62 (9), 2205-2215.

Mwabu, G. M., (1989). Referral systems and health care seeking behavior of patients: an economic analysis. *World Development*, 17 (1), 85-91.

Nardo, L. R. D. O., & Juliani, C. M. C. M., (2012). Ombudsman: evaluating the access to health services. *Northeast Network Nursing Journal*, 13 (3): 613-622.

Nshimirimana, D. A., Kokonya, D., Uwurukundo, J. M. C., Biraboneye, P., Were, F., & Baribwira, C., (2016). Pain Assessment among African Neonates. *American Journal of Pediatrics*, 2 (2), 4-9.

Ringberg, U., Fleten, N., & Førde, O. H., (2014). Examining the variation in ACPs' referral practice: a cross-sectional study of ACPs' reasons for referral. *British Journal of General Practice*, 64 (624), e426-e433.

Robone, S., Rice, N., & Smith, P. C. (2011). Health Systems' Responsiveness and Its Characteristics: A Cross-Country Comparative Analysis. *Health Services Research*, 46(6pt2), 2079-2100.

Stefanini, A., (1994). District hospitals and strengthening referral systems in developing countries. *World Hospitals and Health Services*, 30 (2), 14-19.

Zakus, D., & Bhattacharyya, O., (2007). Health systems, management, and organization in low-and middle-income countries. *Understanding Global Health*, 278-292