A Theoretical Approach to Dual Practice Regulations in the Health Sector*

Paula González[†] and Inés Macho-Stadler[‡]

August 13, 2012

Abstract

Internationally, there is wide cross-country heterogeneity in government responses to dual practice in the health sector. This paper provides a uniform theoretical framework to analyze and compare some of the most common regulations. We focus on three interventions: banning dual practice, offering rewarding contracts to public physicians, and limiting dual practice (including both limits to private earnings of dual providers and limits to involvement in private activities). An ancillary objective of the paper is to investigate whether regulations that are optimal for developed countries are adequate for developing countries as well. Our results offer theoretical support for the desirability of different regulations in different economic environments.

Keywords: Dual practice, optimal contracts, physicians' incentives, regulations.

[*] We appreciate valuable comments by Joan-Ramon Borrell, David Pérez-Castrillo, Nicolás Porteiro and participants at 38th European Association for Research in Industrial Economics (EARIE) Conference, 7th European Conference on Health Economics, XXVI Jornadas de Economía Industrial, 30th Jornadas de la Asociación Española de Economía de la Salud, Universitat de Barcelona and Universidad de Granada. Comments and suggestions from two anonymous referees have improved the paper considerably. Financial support from Ministerio de Ciencia y Tecnología (ECO2008-04321 and ECO2009–07616), Generalitat de Catalunya (2009SGR-169) and Junta de Andalucía (SEJ-02936 and SEJ-04992) is gratefully acknowledged. Inés Macho-Stadler is MOVE fellow.

[†] Corresponding author: Department of Economics, Universidad Pablo de Olavide. Email: pgonzalez@upo.es.

[‡] Department of Economics, Universitat Autònoma de Barcelona and Barcelona GSE. Email: Ines.Macho@uab.es

1 Introduction

In most countries with mixed health care systems it is common practice for physicians to work simultaneously in public hospitals and private facilities. Despite the need for better documented evidence, existing data show that this phenomenon is widespread in many developed and developing countries. In Europe, for instance, physicians working in the public sector often operate in private facilitates under their public contracts. Important examples are Austria, where almost 100% of the senior specialist hospital doctors work in both sectors, Ireland, where more than 90% of the doctors employed in state hospitals also have private practice privileges, and the UK, where 60% of public doctors also work for the private sector. Outside Europe, there is available data for Australia and New Zealand, where 79% and 43% of public sector doctors, respectively, hold some job in the private sector. In developing countries, low public-sector salaries are making exclusively state-employed doctors disappear and, thus, dual practice is prevalent in African countries (Egypt, Mozambique, Zambia, among others), Asian countries (Cambodia, India, Indonesia, Thailand, Vietnam, etc.), Latin America, and also in Eastern Europe.[1]

Despite the prevalence of dual practice among doctors in the majority of mixed health systems, there is a surprising lack of evidence regarding the potential impact on the efficiency of health care resource management. Most health economists agree that this dual practice has both positive and negative side-effects on the delivery of health services. They argue that, on the one hand, allowing dual practice can serve to reduce waiting times for treatment and lead to improvements in access to health services. But, on the other hand, dual providers may have incentives to skimp on work hours or divert patients to private clinics where they have some financial interest, compromising the efficiency and quality of public health provision. On the whole, dual practice among doctors is subject to social controversy, as there is no consensus on its net effects and there is no unique and simple answer as to whether and how this practice should be regulated.[2]

This lack of consensus is reflected by the fact that there is wide cross-country heterogeneity in government responses to dual practice.[3] While some governments ban it altogether,[4] others regulate or restrict dual practice with different regulatory instruments. The measures implemented include offering higher salaries or other work benefits to physi-

[1] See Ferrinho et al. (2004) for further discussion on the available evidence of dual practice in developing countries, and the recent work by García-Prado and González (2011) for a review on the extent of dual practice in the health sector.

[2] See Eggleston and Bir (2006) and Socha and Bech (2011) for a thorough discussion on these issues.

[3] See García-Prado and González (2007) for a review of these policies.

[4] China (Jingqing, 2006) and Canada (Flood and Archibaldare, 2001) are examples of countries where physician dual practice is forbidden.

cians in exchange for their working exclusively in the public sector,[5] limiting the income physicians can earn through dual practice,[6] and limiting dual practice through government specification of the maximum involvement in private activities.[7] In addition, most of these regulations have been introduced only in developed countries, while in developing countries dual practice remains largely unregulated, although it is attracting more attention from policy makers.

In this paper we provide a theoretical model to study different governmental responses to dual practice. The aim of the paper is two-fold. First, we analyze from a theoretical point of view different regulations that are currently employed to deal with dual practice. Secondly, we investigate whether the regulatory policies that are optimal for developed countries are adequate for developing countries as well, or whether a different policy mix is needed. As discussed below, there are no existing works in the literature that provide a uniform theoretical framework to evaluate the desirability of one or another regulation on dual practice. We believe our results shed new light on these questions.

We construct a simple model in which a Health Authority contracts physicians in order to provide public health care and designs the regulatory regime regarding dual practice. Physicians have different levels of ability, interpreted as their capacity to provide adequate health services to patients, and they can choose, given the regulatory regime and available contracts, whether to work solely for the public sector, as dual practitioners, or exclusively in the private sector. In our model the public/private interaction is two-fold. On the one hand, private practice might affect the performance of a physician in the public sector. On the other hand, if the private market recognizes and rewards ability it becomes costly for the Health Authority to retain highly skilled physicians within the public sector.

[5] The governments of Spain, Portugal and Italy, among others, have offered public physicians exclusive contracts that aim to ensure that signatories do not engage in private practice in exchange for salary supplementation or promotions.

[6] The restriction of private earnings of publicly employed physicians has been implemented in the UK and in France. In the UK, full-time NHS consultants, who are mostly senior specialists, are permitted to earn up to 10% of their gross income from private practice in addition to their NHS earnings. Those NHS doctors who work under a maximum part-time contract are allowed to practice privately without earning restrictions by giving up one eleventh of their NHS salary (European Observatory on Health Systems, 2004). Similarly, in France, public hospitals employ both full-time and part-time physicians who can also provide private services subject to the restriction that income from private fees is limited to 30% of physician total income (Rickman and McGuire, 1999).

[7] In Austria, Ireland and Italy physicians are encouraged to perform private services within government hospitals and the share of beds allocated to privately insured patients is legally defined. In Austria the share of beds allocated to privately insured patients must not exceed 25% of total beds (Stepan and Sommersguter-Reichmann, 2005). In Italy public hospitals are required to reserve between 6% and 12% of their beds for private patients (France, Taroni and Donatini, 2005). Similarly, in Ireland, 20% of beds in publicly funded hospitals are designated for private patients (Wiley, 2005).

We analyze regulations that deal with dual practice using two different health production functions in the public sector so as to illustrate various situations in different countries. First, we consider an environment where the production of health within the public sector depends mostly on the overall number of public physicians and not so much on their individual characteristics. We identify this situation with developed countries where the availability of advanced medical technology, existence of standardized treatment protocols and adherence to practice guidelines substantially reduces physician discretion. We also consider a health production function for which the personal characteristics of each physician play an important role in the provision of health care, a scenario that we believe more closely resembles what happens in less developed economies.

We focus on three kinds of interventions: banning dual practice, offering rewarding contracts to public physicians, and limiting dual practice, including both earnings limitations and limits to involvement in private activities.

Our model yields some interesting implications concerning regulation. First, if a policy of limiting dual practice is to be enforced, limiting physicians' earnings from dual practice is always worse than limiting their involvement. The reason is that a policy that constrains private income has a milder effect on the amount of dual practice performed, and therefore on its associated costs, as it only affects highly skilled physicians who must reduce private activities in order to satisfy their earning constraint. In contrast, a policy that limits involvement in private activities directly targets the intensity of dual practice and is therefore more effective in curbing losses in productivity.

While the above recommendation is general, our analysis suggests that in many respects optimal policies differ for developed and less developed economies. In developed countries the choice of regulatory intervention depends solely on the cost of the dual practice. For small costs no intervention is required, while for large costs the best intervention is to impose a limit on physician involvement in dual practice. Interestingly, we find that banning dual practice, even if it is enforceable, is never desirable. Even if dual practice imposes a significant burden on the public production of health, the Health Authority can alleviate these costs as dual practice reduces the salary needed to retain doctors working at public facilities. Finally, offering exclusive contracts to physicians who volunteer to work exclusively in the public sector is optimal only if a limiting policy faces enforceability problems.

In developing countries the results differ sharply, as it is the attractiveness of the private sector that determines the need for regulation. If the attractiveness of the private sector is high, then the government should never intervene, regardless of the cost of dual practice. In this case, restricting dual practice pushes highly skilled physicians into the private sector, and the Health Authority of a developing country cannot afford to lose its most able professionals. When the private sector is unattractive, however, the risk of losing physicians is low, and the best policy is either to limit dual practice (if the cost associated with dual practice is low) or to ban it (if the cost is high). Limiting policies in developing countries always emerge as the optimal instrument in situations in which the private sector is moderately attractive, i.e. not so low as to make banning feasible, and not so high as to draw a significant number of physicians away from the public sector. Exclusive contracts are never optimal in developing countries. The reason is that the physicians who accept the premium and become public-only providers tend to be the less productive. Given the importance that doctors' individual characteristics have for the production of public health in developing countries, paying such a premium is not worthwhile.

The theoretical literature on physician dual practice in mixed health care markets is not abundant and relatively recent.[8] There has been some research on physicians' incentives as dual providers. Rickman and McGuire (1999) concentrate on the implications of the fact that a doctor can offer both public and private services to the same patient and examine the optimal public reimbursement for doctors who are dual providers. Barros and Olivella (2002) and González (2005) analyze the physician's decision to "cream-skim" patients in a context with waiting lists in the public sector. While González (2005) shows that if doctors are dual providers, the most profitable patients will be referred to their private practices, Barros and Olivella (2002) find that if public treatment is rationed it is not necessarily the case that physicians end up treating the mildest cases from the waiting list in their private practice. Finally, Delfgaauw (2007) considers the implications of differences in physician altruism. He shows that allowing for private provision of health care in parallel to public provision is generally beneficial for patients, but allowing physicians to transfer patients from the public system to their private practices reduces these benefits, as it harms the poorest patients.

There are very few works that focus on the regulations that deal with dual practice. González (2004) presents a model in which a physician has an incentive to provide excessive quality in the public sector in order to raise prestige. In such a context, limiting private practice might not be desirable. She also shows that the use of exclusive contracts can be a valuable regulatory measure when governments cannot design appropriate incen-

[8] There are other papers in the health economics literature that have examined the interaction between public and private health care provision, but they do not consider job incentives of physicians working in both sectors. These include Barros and Martínez-Giralt (2002), which analyzes the effect of different reimbursement rules on quality and cost efficiency; Iversen (1997), which considers the effect of private health care provision on waiting lists in the public sector; Jofre-Bonet (2000), which studies the interaction between public and private providers when consumers differ in income; and Marchand and Schroyen (2005), which analyzes the desirability of mixed health care systems when distributional aspects matter.

tive contracts. Biglaiser and Ma (2007) also study the incentives of moonlighting, which can lead public-service physicians to refer their patients to their private practices. Using a model where some doctors are dedicated to the public system and behave honestly while others are utility maximizers, they show that limiting private practice revenues through price ceilings reduces the adverse behavioral reactions of public sector physicians and can improve public service quality. Finally, using a model in which physicians divide their labour between public and (if allowed) private sectors, Brekke and Sørgard (2007) suggest that allowing physician dual practice 'crowds out' public provision, and results in lower overall health care provision. Thus, a ban on dual practice can be an efficient policy when private sector competition is weak and public and private provisions are sufficiently close substitutes. All these papers analyze specific policies in different settings. Therefore, to the best of our knowledge, ours is the first work that provides a uniform theoretical framework through which the desirability of different regulations that deal with dual practice can be determined and compared.

The structure of the paper is as follows. Section 2 presents our model. Section 3 introduces two simple regimes: a laissez-faire scenario where dual practice is allowed without regulation, and the opposite extreme, where dual practice is forbidden. Section 4 concentrates on rewarding policies for physicians that work for the public sector exclusively, while Section 5 analyzes limiting policies. Section 6 characterizes the optimal policy mix for the regulation of dual practice and elaborates on the main policy implications of the preceding analysis. Finally, the last section offers some concluding remarks. All of the proofs are in the Appendix.

2 The Model

There are two sets of agents involved in the model: a set of physicians and the regulator or Health Authority (HA hereafter). The HA aims to provide public health care but is also concerned about its costs. The quality (or the level) of publicly provided care depends on which physicians work in the public system and on whether these physicians are involved in dual practice or not. We assume that the HA designs the rules for performing dual practice and, given the basic regime (dual practice allowed or not), the physicians choose among the different options available to them. Accordingly our model has two stages, and we solve the game by backwards induction.

Let us now detail for each of the players their objective functions, decision variables and all the parameters that are relevant in the model.

2.1 The Physicians

Physicians have different abilities a distributed uniformly on the interval $[0, \bar{a}]$.[9] The total amount of physicians has mass $\bar{a}$. Physicians can work solely in the public sector, work for the private sector or work in both sectors as dual providers. If they work for the public sector they receive a net wage w, in exchange for a fixed working time (that we denote by g, $g > 0$). To keep the model tractable we rule out the possibility that a physician can decide the extent of his working time in the public sector, g. In other words, we disregard the possibility that a physician chooses to work part-time at public facilities. If physicians work in the private sector (either as dual providers, D, or as exclusively private physicians, Pv) the profits they obtain Π are the revenues from the patients treated, net of the costs. Physicians' private revenues depend on the private market price per consultation ($p, p \geq 0$) and the number of patients treated. We assume that the number of patients treated in private practice follows a Cobb-Douglas type production function, $2h\sqrt{a\gamma}$, that depends on the ability of the physician, a, and his level of involvement (for instance, working hours) in private practice (γ, $\gamma \geq 0$), as well as on a total factor productivity $2h$ (with $h > 0$).[10] The cost of working in the private sector depends on the level of involvement in private activities and on whether physicians work also or not within the public sector ($G \in \{g, 0\}$). The cost function is given by $C(\gamma, G) = c_G\gamma$,[11] where:

$$c_G = \begin{cases} 2 & \text{if } G = g \\ 1 & \text{if } G = 0. \end{cases}$$

Consequently, profits from private practice have the following form:[12]

$$\Pi(a, \gamma, G) \equiv p2h\sqrt{a\gamma} - c_G\gamma. \tag{1}$$

[9] Note that denoting the lowest ability by $a = 0$ is only a normalization. In our model, all doctors have been trained and are able to perform as certified physicians.

[10] The factor of productivity is written as $2h$ since it simplifies the expression of the derivatives without loss of generality.

[11] A more natural modelization would be to introduce a strictly convex cost of working in the private sector. However, this leads to too complex expressions that prevent us from getting closed results. The discontinuity in the marginal costs adopted in our model allows us to introduce convexity in the model without adding more complexity.

[12] An alternative way of modeling physician private profits is to think of physicians as having a fixed time budget and deciding the fraction of time to allocate between work and leisure. Consider, for instance, $\Pi(a, \gamma, G) \equiv p2h\sqrt{a\gamma} + \varepsilon_G(T - G - \gamma)$, where T is total time available, γ measures again the time allocated to private practice and ε_G represents the value the physician allocates to leisure. Under the assumption that the physician appreciates more leisure when he has less free time (standard decreasing marginal utility of leisure), this alternative modelization is equivalent to ours. The details are available from the authors upon request. We thank an anonymous referee for pointing out this possibility.

Considering marginal cost $c_0 = 1$ for $G = 0$ is just a normalization. Assuming $c_g = 2$ when $G = g$ is done in order to simplify the exposition by reducing the number of cases under study. The essential element is that being involved in private practice is more costly for those physicians that are also working in the public sector. This reflects the idea that there is an inherent cost (in terms of effort) associated with dual practice (due, for instance, to the fact that these extra hours are done in addition to the normal dedication g in the public sector, the need to commute from one facility to the other, necessity to coordinate two patient schedules, potentially different working protocols, etc.). If we interpreted our model in terms of time constraints (see footnote 12), the extra cost associated with dual practice would come from the fact that simultaneously working in the two sectors leaves physicians less time for leisure.

2.2 The Health Authority

The other agent involved in the model is the HA. To define the *Health Authority*'s objective function, we take the view that the HA is only concerned about the level of heath care provided by the public system. In other words, we assume that the HA does not include the private provision of health in its objective function.[13] We assume that the performance of a physician in the public sector depends on his ability and is given by the function $F(a)$. If the physician is a dual supplier, however, this has an impact on his public sector performance. Formally, a dual provider's performance in the public sector is given by $\frac{1}{1+\delta\gamma}F(a)$, where δ measures the marginal impact of the private activity on public sector performance. Note that this functional form allows for several situations. If $\delta = 0$ public and private activities are independent, and dual practice does not affect the performance of the physician in the public sector. A loss associated with dual practice (related, for instance, to the fact that physicians divert time and attention from hard-to-control tasks, or to the emergence of conflicts of interest such as induced demand, etc.) is represented by positive values of δ. This functional form also accommodates situations in which complementarities exist between the two sectors, corresponding to a negative δ. In what follows, however, our discussion will concentrate on $\delta > 0$, since we are interested in analyzing situations where the regulator is concerned about the negative implications of doctors' involvement in dual practice.[14] Note that if $\delta > 0$ a dual provider's performance

[13] This assumption reflects that the performance of private health providers is out of the scope of intervention of the HA. This can be sustained on the grounds that the HA is an agency of the government whose mandate is to guarantee an adequate performance of the public health sector. In this line, one can argue that ensuring an efficient provision of public health services is part of the government policy agenda and voters will evaluate the attainment of this goal.

[14] We will briefly discuss the results when $\delta < 0$ in the next section, to illustrate why this case is not of particular interest for our analysis.

in the public sector is decreasing and convex in γ.

Let us define as $pub \subset [0, \bar{a}]$ the set of all doctors working exclusively for the public sector, and as $D \subset [0, \bar{a}]$ the set of all doctors involved in dual practice. We denote by $|pub|$ and $|D|$ the size (number of physicians) of the sets pub and D respectively. Then, we write the HA's objective function as:

$$\max_{w} \mathcal{W} = \int_{a \in pub} F(a)\, da + \int_{a \in D} \frac{1}{1+\delta\gamma} F(a)\, da - (1+\lambda)\, w(|pub| + |D|).$$

The first term measures the health provided at public facilities. The last term represents the wage costs: how many physicians work (exclusively or partially) in the public sector times the salary. The parameter λ reflects the marginal cost of public funds, and can be conceived as the relative importance that costs have as compared to health revenues.

The HA decides on the wage w, which indirectly determines the physician's decision to allocate services. Assuming that the wage in the public sector is independent of the physician's ability is a simplifying assumption trying to capture the idea that physician payments are more performance related in the private sector than in the public. Moreover, it accounts for the fact that most countries pay public physicians in the same service and/or category the same wage. Note that, without loss of generality, we assume that the HA does not introduce any constraint on the number of physicians that will be hired in the public sector since when it is interested in reducing participation it is sufficient to reduce the wage, which allows it to save costs.

As mentioned in the Introduction, this model can be used to understand how the implications of dual practice might differ for developing and more developed countries, and also to assess how the relative merits of different regulations depend on the type of economy. For this purpose, we consider two alternative technologies $F(a)$ for the production of health in the public sector.[15] Developed countries benefit from widespread use of advanced technologies and test-based diagnoses, as well as rigorous training processes, standardized treatments and protocols, and strict adherence to practice guidelines. Moreover, the large size of public facilities facilitates the referral of patients to specialists and the formation of teams of physicians who share information and discuss especially difficult cases. All these features point towards a lower degree of physician discretion and hence

[15] Although throughout the paper we consider the access to different technologies as the identifying feature of developed and developing countries, we thank an anonymous referee for raising two issues that are also relevant for our discussion. First, within a given health care system, different specialties may have access to different technologies, what would call for a more speciality-specific dual practice regulation. Second, leaving aside technological aspects, there are other distinguishing features among countries. For instance, developing countries are often characterized by inefficient fiscal systems and, therefore, by a higher shadow cost of raising public funds than their developed counterparts. We discuss these (and other related) issues in Subsection 6.2.

reduced impact of individual physician characteristics on the quality of care delivered at public facilities. We model this by assuming a health production technology of the form $F(a) = \varphi$. In contrast, in developing countries the lower degree of specialization among physicians, their obligation to cope with illnesses outside their area of expertise, the lack of infrastructures and modern technologies that support diagnosis, and the lack of formalized medical protocols all make the actual quality of care more dependent on individual physician characteristics. For this reason, we consider a health production technology of the form $F(a) = fa$.[16]

In order to simplify the analysis, by avoiding degenerate cases, we make the following assumption.

Assumption 1 *We assume that: (i) $\bar{a}$ is high enough, i.e., $\bar{a}(1+\lambda)k^2 \geq \varphi$, and (ii) $f > (1+\lambda)k^2$.*

Assumption 1 part (i) guarantees that for the constant public health production function $F(a) = \varphi$ the upper bound $\bar{a}$ is sufficiently large so as to avoid (corner) situations in which the HA induces no physician to work solely in the private sector. Assumption 1 part (ii) guarantees that for the health production function $F(a) = fa$ the public productivity parameter f is large enough so that it always pays to sustain the public sector.

Now we have all the tools to study the impact of different policy options to regulate dual practice. As mentioned in the Introduction, we observe wide variations in how governments tackle the issue of dual practice. In the following sections we analyze several policies currently in force in some health care systems. We, first, consider only the choice allowing versus prohibiting dual practice. We then study more sophisticated regulations such as the desirability of allowing dual practice while offering work benefits to physicians in exchange for their working exclusively in the public sector, limiting the income physicians can earn through dual job holding, and limiting the degree of involvement of public physicians in private activities.

3 Laissez-faire versus Banning

The first possible policy option is to ban dual practice altogether. This scenario might represent a situation where the HA cannot use sophisticated regulations and is restricted to using simple all-or-nothing instruments.

[16] Similar arguments often appear when comparing urban and rural practitioners. For instance, Rabinowitz and Paynter (2002) higlights that rural physicians retain more clinical independence in their practice and, at the same time, they may experience professional isolation, with less access to colleagues and medical resources.

Using backwards induction, we study first the physician's decision. The physician makes two choices: where he will work and the time, if any, he will devote to private practice. Notice that the term ph appears together in (1). Hence, to simplify the presentation, in what follows we will use the notation $k \equiv ph$ and we will refer to k as a measure of the relative profitability (attractiveness) of the private sector. We will also label with the superscript Pb a physician who works in exclusivity for the public health service, with D a dual practitioner, and with Pv a physician working only in private practice.

A physician who practices solely in the private sector chooses the intensity of his private practice γ^{Pv} in order to maximize his profits. Analogously, in cases where the HA does not impose any restriction, a dual provider chooses his optimal level of involvement in private practice γ^{D}. Maximizing

$$\Pi\left(a, \gamma, G\right) \equiv 2k\sqrt{a\gamma} - c_G\gamma,$$

it is straightforward to compute that the optimal involvement in private practice for a dual provider and an only-private provider are $\gamma^{D}\left(a\right) = \frac{k^2 a}{4}$ and $\gamma^{Pv}\left(a\right) = k^2 a$, respectively. The optimal dedication, in turn, yields the following profits from private practice

$$\Pi^{D}\left(\gamma^{D}\left(a\right) = \frac{k^2 a}{4}\right) = \frac{k^2 a}{2} \quad \text{and} \quad \Pi^{Pv} = k^2 a.$$

From the previous analysis it follows that the physician's utility, depending on the type of practice, is:

$$\begin{aligned} U^{Pb} &= w \\ U^{D} &= w + \Pi^{D} \\ U^{Pv} &= k^2 a. \end{aligned} \tag{2}$$

We can now characterize the physician's decision as a function of his ability and the wage offered in the public sector. This is done by comparing the utilities in (2), taking into account that $\Pi^{D} = \frac{k^2}{2}a$, if there is no restriction on dual practice, and $\Pi^{D} = 0$ if dual practice is forbidden.

Lemma 1 *For a given salary w, the optimal decision of a physician, as a function of his ability a, is as follows:*

a) If dual practice is allowed, and

if $a \in \left[0, \frac{2w}{k^2}\right]$	*he chooses dual practice*
if $a \in \left(\frac{2w}{k^2}, \bar{a}\right]$	*he chooses to work only in the private sector*

b) If dual practice is not allowed, and

if $a \in \left[0, \frac{w}{k^2}\right]$	*he chooses public practice*
if $a \in \left(\frac{w}{k^2}, \bar{a}\right]$	*he chooses to work only in the private sector*

Lemma 1 presents the optimal strategy for physicians allocating time to the different types of practice. The more able ones tend to be more involved in the private sector since their ability allows them to get a higher return. The less able tend to combine both public and private activities if dual practice is allowed, or work only in public practice when this is not the case. When dual practice is forbidden, the population of physicians working for the public sector for a given salary decreases (since, for any w, $\frac{2w}{k^2} > \frac{w}{k^2}$). In addition, when the public and private sectors do not share physicians, higher private sector earnings are expected to attract more highly skilled physicians, leaving those of lesser ability in the public sector.

Thresholds can also be read in terms of physicians with the same ability but a different parameter k. The term k serves as a proxy for the attractiveness of the private sector for a particular specialitation and, consequently, the higher is k the less physicians will work solely in the public sector (both if dual practice is allowed and if it is forbidden). The fact that k is proportional to the price-per-consultation p, which may be considered as speciality-specific, allows us to discuss the different behaviors of physicians engaged in primary, secondary, and tertiary care, and in different specialities. In this regard, the properties of these thresholds are in accordance with some stylized facts since more doctors will be involved in the private sector as p increases. For example, Gruen et al. (2002), using data from a survey in Bangladesh, found that primary-care physicians were willing to give up dual practice in exchange for a higher salary but doctors engaged in secondary and tertiary care were far more reluctant to do so. This might reflect the higher attractiveness of the private sector for more specialized physicians. An alternative interpretation would relate the parameter k to the financial motivation of the physicians. In this case, our results suggest that physicians with stronger financial motivations will be more prone to dual practice either because they suffer from financial constraints or because public sector salaries are low. This is also in accordance with stylized facts that report that young physicians (whose salary is smaller and often have to pay off educational loans) tend to be substantially involved in dual practice. It also accords with the "brain drain," i.e. the desire to migrate to countries where physicians' pay is higher.[17]

When banning is the only policy available, the HA either lets physicians freely decide whether and how to be dual providers or forbids dual practice and lets physicians choose only between public or private provision. Given the physicians behavior, if dual practice

[17] See, for instance, Mainiero and Woodfield (2008) for an account of the evidence of moonlighting among radiology residents in the United States, and Mayta-Tristán et al. (2008) for a warning of the risk of brain drain of physicians in Peru.

is allowed, the problem that the HA faces to fix the wage in the public sector w^{LF} is

$$\max_{w} \mathcal{W}^{LF} = \int_{0}^{\frac{2w}{k^2}} \frac{F(a)}{1+\delta\frac{k^2 a}{4}} da - (1+\lambda) w \frac{2w}{k^2}. \tag{3}$$

If there is a ban on dual practice, and assuming that this policy is enforced, the problem that determines the optimal wage w^{B} is

$$\max_{w} \mathcal{W}^{B} = \int_{0}^{\frac{w}{k^2}} F(a)\, da - (1+\lambda) w \frac{w}{k^2}. \tag{4}$$

We focus first on the HA's choice for developed countries.

Proposition 1 *In **developed economies,** if the HA can only ban dual practice, there exists a $\bar{\delta}_1$, with $\bar{\delta}_1 \approx \frac{5.988(1+\lambda)}{\varphi}$, such that the best intervention is as follows,*

i) *If $\delta \leq \bar{\delta}_1$ not to regulate dual practice and set a wage level*

$$w^{LF} = \frac{\sqrt{1+\delta\frac{\varphi}{1+\lambda}}-1}{\delta}.$$

ii) *If $\delta > \bar{\delta}_1$ ban dual practice and set a wage level*

$$w^{B} = \frac{\varphi}{2(1+\lambda)}.$$

The results in Proposition 1 are predictable and, using Lemma 1, imply (respectively) the cut-offs

$$a^{LF} = \frac{2\left(\sqrt{1+\delta\frac{\varphi}{1+\lambda}}-1\right)}{\delta k^2} \qquad \text{and} \qquad a^{B} = \frac{\varphi}{2(1+\lambda)k^2}.$$

From Proposition 1, it is easy to check that for any combination of parameters, $w^{LF} < w^{B}$. This implies that dual practice might be desirable because it allows the HA to reduce the wage needed to retain physicians working in the public sector. This is in agreement with one of the traditional arguments in the literature in favor of allowing multiple job holdings, namely that the cost of attracting a worker is smaller when the primary job offers a wage and the possibility of extra income via dual practice (Holmström and Milgrom, 1991). However we also have to take into account the potential costs of dual practice, and we conclude that when this cost is sufficiently high (δ large), it does not pay to allow dual practice. Hence, for those specialities where (other things equal) the loss is high the HA will decide to ban dual practice.

To complete the analysis, we briefly discuss the comparative statics of the results in Proposition 1. As one might expect, if the HA puts increased weight on public health provision (lower λ), or health production technology becomes more efficient (higher φ), then a higher salary will be paid to public physicians and, hence, a larger number of practitioners will work for the public sector. Conversely, a larger cost of dual practice (higher δ) results in smaller wages and less physicians hired in the public sector when dual practice is allowed.

Note that the results presented in Proposition 1 are also valid (and well-defined) for negative values of δ (as long as $-\delta < \frac{1+\lambda}{2\varphi}$). For $\delta \leq 0$, the laissez-faire regime is always superior.[18] Since this superiority result is maintained throughout the paper, we will not discuss it further. The remaining analysis focuses on the case $\delta > 0$.

Let us now turn to the case of developing countries:

Proposition 2 *In **developing economies,** when the HA can only ban dual practice, there exists a $\bar{\delta}_2(k)$ such that the best intervention is as follows,*

i) *If $k \in \left(\sqrt{\frac{f}{2(1+\lambda)}}, \sqrt{\frac{f}{1+\lambda}}\right]$, irrespective of δ, or if $k \leq \sqrt{\frac{f}{2(1+\lambda)}}$ and $\delta \leq \bar{\delta}_2(k)$, not to regulate dual practice and set a wage level*

$$w^{LF} = \frac{2\left(f - (1+\lambda)\, k^2\right)}{\delta\,(1+\lambda)\, k^2}.$$

ii) *If $k \leq \sqrt{\frac{f}{2(1+\lambda)}}$ and $\delta > \bar{\delta}_2(k)$, ban dual practice and set a wage level*

$$w^B = \bar{a}k^2.$$

Using Lemma 1, the cut-offs when dual practice is allowed and forbidden are respectively

$$a^{LF} = \frac{4\left(f - (1+\lambda)\, k^2\right)}{\delta\,(1+\lambda)\, k^4} \qquad \text{and} \qquad a^B = \bar{a}.$$

We see how in developing economies the attractiveness of the private sector (k) plays a key role. Only when the private sector is relatively unattractive (and on top of that there is a high value of δ) does the HA find it optimal to ban dual practice. Otherwise the best it can do is to cope with its negative implications. The reason is that a high k implies that banning dual practice will encourage physicians to leave the public sector. Thus, the public health sector will suffer from a severe brain drain of the most able physicians.[19]

[18] For the particular case $\delta = 0$, the HA is not concerned about dual practice and it will set the same wage, $w = \frac{\varphi}{2(1+\lambda)}$ in both regimes. Since by allowing dual practice the HA is able to attract more doctors, and hence to provide more health, regulation will be never in the HA's interest.

[19] There is evidence that bans on dual practice in developing countries lead to a significant drain of physicians from public to private practice as well as a migration of physicians to other countries with better work conditions. See Globerman and Vining (1998) and Peters et al. (2002) for experiences in South Africa and India respectively.

Since the capacity of the public sector to produce health is directly linked to the ability of the public physicians, losing the most able professionals is something the HA cannot afford. Note that in the case where for a given speciality the private sector is extremely attractive, it will not be optimal for the HA even to maintain that specialty in the public sector. The minimum wage that a physician would require (k) in that case would exceed the marginal value of his contribution to the public sector $\left(\sqrt{\frac{f}{1+\lambda}}\right)$.[20] If one accepts that k may depend on the level of health care provision, the previous result indicates that in developing economies it might be optimal in some cases to provide only primary health care in the public sector.

Further analysis shows that, as the cost faced by the HA to allocate public funds to the health service decreases (lower λ), health production technology becomes more efficient (higher f), or the cost of dual practice goes down (lower δ), then salaries in the public sector rise, and an increasing number of practitioners work for the public sector under the Laissez-Faire regime.[21]

4 Rewarding Policies

Let us now consider the policy of paying (on top of a salary w^E) a premium $\Delta \geq 0$ to physicians who decide to work exclusively for the public sector (this bonus can also be interpreted in terms of better career opportunities). In this section we investigate the conditions under which this kind of policy, which is currently implemented in several health systems (e.g. those of Spain, Portugal, and Italy), can be an optimal regulatory tool.

In this setting, the physician's utility, depending on the type of practice, is

$$\begin{gathered} U^{pub} = w^E + \Delta \\ U^D = w^E + \tfrac{k^2 a}{2} \\ U^{Pv} = k^2 a. \end{gathered}$$

The following lemma summarizes the physician's decision as a function of his ability and the contracts offered in the public sector.

Lemma 2 *Given (w^E, Δ, k), when dual practice is not restricted, the optimal decision of a physician as a function of his ability a is as follows*

[20] Note that the case with $k > \sqrt{\frac{f}{1+\lambda}}$, where the HA is confronted with a high k for all types of health care (implying that no physician would work in the public sector), is left-out by Assumption 1 (ii).

[21] Under the *Banning* regime, w^B and a^B do not depend on λ and f, but notice that these parameters affect the threshold separating the different regions in Proposition 2.

- *When* $\Delta > w^E$, *then*

$$\begin{array}{ll} \textit{if } a \in \left[0, \frac{w^E+\Delta}{k^2}\right] & \textit{he chooses to work only in the public sector} \\ \textit{if } a \in \left(\frac{w^E+\Delta}{k^2}, \bar{a}\right] & \textit{he chooses to work only in the private sector} \end{array}$$

- *When* $\Delta \leq w^E$, *then,*

$$\begin{array}{ll} \textit{if } a \in \left[0, \frac{2\Delta}{k^2}\right] & \textit{he chooses to work only in the public sector} \\ \textit{if } a \in \left(\frac{2\Delta}{k^2}, \frac{2w^E}{k^2}\right] & \textit{he chooses dual practice} \\ \textit{if } a \in \left(\frac{2w^E}{k^2}, \bar{a}\right] & \textit{he chooses to work only in the private sector} \end{array}$$

Lemma 2 presents the optimal strategy for physician allocating time to different types of practice when exclusive contracts are enforced. The more skilled physicians tend to be more involved in the private sector as their ability allows them to have a higher return. The less skilled tend to be fully involved in public practice. It can be seen that by setting Δ the HA can induce a situation in which no physician chooses to be a dual provider ($\Delta > w^E$).[22] Note also that if $\Delta = 0$ (there is no extra wage for exclusivity in the public sector) then a physician will never work exclusively in the public sector. As k -which summarizes the profitability of private practice- increases, less physicians tend to work exclusively in the public sector.

This regulatory environment encompasses the laissez-faire ($\Delta = 0$ and $w^E = w^{LF}$) and banning ($w^E = 0$ and $\Delta = w^B$) regimes examined in the previous section. What needs further analysis are the conditions under which it actually pays for the HA to offer a real exclusive contract that induces some physicians to work solely in the public sector. As the following proposition shows this depends crucially on the type of health care system.

Proposition 3 *In* ***developed economies,*** *when the* HA *can offer an exclusive contract the best intervention is,*

i) *If* $\delta \leq \frac{4(1+\lambda)}{\varphi}$ *not to regulate dual practice, fix* $\Delta = 0$*, and set a wage level*

$$w^E = w^{LF} = \frac{\sqrt{1+\delta\frac{\varphi}{1+\lambda}} - 1}{\delta}$$

ii) *If* $\delta \in \left(\frac{4(1+\lambda)}{\varphi}, \frac{8(1+\lambda)}{\varphi}\right]$ *to offer an exclusive contract with*

$$\Delta = \frac{\varphi\delta - 4(1+\lambda)}{2\delta(1+\lambda)} < w^E = w^{LF} = \frac{\sqrt{1+\delta\frac{\varphi}{1+\lambda}} - 1}{\delta}$$

[22] The fact that the bonus can exceed the baseline wage is a feature of the model, but the fact that it has to be large enough in order to effectively deter dual practice is not.

iii) *If $\delta > \frac{8(1+\lambda)}{\varphi}$ to ban dual practice, fix $w^E = 0$, and set an exclusivity premium*

$$\Delta = w^B = \frac{\varphi}{2(1+\lambda)}$$

We see how, in developed economies, whether it pays or not to allow dual practice depends on its costs. If δ is low, it does not pay to try to reduce the incentives of the physicians to work as dual suppliers. Exclusivity premiums are not paid and all physicians working in the public sector are dual providers. As δ increases, it is more and more profitable to pay an exclusivity premium in order to deter some physicians from being dual providers. In that case, some physicians decide to work exclusively in the public sector, some are dual providers and the remaining work solely in the private sector. Finally, if δ is sufficiently high, then it is in the HA's interest to pay a premium so high that it deters all physicians from dual practice (which is equivalent to banning dual practice).

If we compare these results with those in Proposition 1, we find that the threshold of the productivity loss beyond which it is optimal for the HA to ban dual practice is strictly lower in the laissez-faire scenario $\left(\bar{\delta}_1 \approx \frac{5.988(1+\lambda)}{\varphi}\right)$ than when exclusive contracts are available $\left(\frac{8(1+\lambda)}{\varphi}\right)$. This illustrates how exclusive contracts offer greater flexibility for the HA to mitigate the loss of productivity associated with dual practice, which makes the HA less interested in banning dual practice when rewarding policies are available.

However, as we now detail, the results for developing countries contrast sharply with those just described.

Proposition 4 *In developing economies the HA never finds it optimal to offer an exclusive contract to physicians. Instead, the decision is between no regulation and banning dual practice altogether, as characterized in Proposition 2.*

In developing countries a rewarding policy such as an exclusivity premium intended to induce some physicians to work solely in the public sector is never an optimal intervention. The reason is that such a policy would attract only the less able physicians (those with lower prospects of private earnings). This also happens in developed economies, but the characteristics of the health care systems in developing countries make the provision of care much more dependent on physician ability. For this reason, it never pays to offer an extra premium as it only attracts those physicians with the smallest capacity to contribute to health care production.

5 Limiting Policies

In this section we consider scenarios in which the HA restricts dual practice. This is modelled as a constraint fixed by the HA that limits physician involvement in dual practice. We consider two possible restrictions: in the first, physician involvement in the private sector is subject to a maximum of $\bar{\gamma} \geq 0$; in the second one, the earnings of the public physician in his private practice are limited to a maximum amount $\bar{\Pi}^D$. Then, given these cut-offs ($\bar{\gamma}$ or $\bar{\Pi}^D$), physicians choose their level of involvement γ.

We characterize physician behavior when the option to engage in dual practice is subject to limitation. Recall that, in the absence of any regulation, the optimal involvement in dual practice is $\gamma^D(a) = \frac{k^2 a}{4}$. Focusing first on an involvement constraint $\bar{\gamma} \geq 0$, we find:

Lemma 3 *When there is a policy that limits the maximum involvement in dual practice to $\bar{\gamma}$, the physician's amount of dual practice is*

$$\begin{array}{ll} \gamma^*(a) = \bar{\gamma} & \text{if } a > \frac{4\bar{\gamma}}{k^2}, \text{ and then } U^D = w + 2k(a\bar{\gamma})^{1/2} - 2\bar{\gamma} \\ \gamma^*(a) = \frac{k^2 a}{4} & \text{if } a \leq \frac{4\bar{\gamma}}{k^2}, \text{ and then } U^D = w + \frac{k^2 a}{2}. \end{array}$$

Consequently for a given $(\bar{\gamma}, w)$ the physician's optimal choice is:

- *If $\bar{\gamma} \geq \frac{w}{2}$, the limiting policy is ineffective and the physician's decision coincides with that in Lemma 1, part a).*
- *If $\bar{\gamma} < \frac{w}{2}$, the limiting policy is effective and*

$$\begin{array}{ll} \text{if } a \in \left[0, \frac{4\bar{\gamma}}{k^2}\right] & \text{the physician chooses dual practice and } \gamma^*(a) = \gamma^D(a) = \frac{k^2 a}{4} \\ \text{if } a \in \left(\frac{4\bar{\gamma}}{k^2}, \frac{w + 2\sqrt{\bar{\gamma}(w-\bar{\gamma})}}{k^2}\right] & \text{the physician chooses dual practice and } \gamma^*(a) = \bar{\gamma} \\ \text{if } a \in \left(\frac{w + 2\sqrt{\bar{\gamma}(w-\bar{\gamma})}}{k^2}, \bar{a}\right] & \text{the physician chooses to work only in the private sector.} \end{array}$$

The second limiting policy constrains the revenue that the physician can obtain from his dual practice to a maximum of $\bar{\Pi}^D$. In this case, given the cut-off $\bar{\Pi}^D$, the physician may choose any level of dual practice γ, provided the private revenues are such that $2k\sqrt{a\gamma} - 2\gamma \leq \bar{\Pi}^D$. Let us denote by $\hat{\gamma}\left(a, \bar{\Pi}^D\right)$ the level of dual practice that a physician with ability a will choose when he is subject to a limiting policy $\bar{\Pi}^D$.

Lemma 4 *When there is a policy that limits the maximum private earnings obtained in dual practice to $\bar{\Pi}^D$, the physician's amount of dual practice is*

$$\begin{array}{ll} \gamma^*(a) = \hat{\gamma}\left(a, \bar{\Pi}^D\right) < \frac{k^2 a}{4} & \text{if } a > \frac{2\bar{\Pi}^D}{k^2}, \text{ and then } U^D = w + \bar{\Pi}^D \\ \gamma^*(a) = \frac{k^2 a}{4} & \text{if } a \leq \frac{2\bar{\Pi}^D}{k^2}, \text{ and then } U^D = w + \frac{k^2 a}{2}. \end{array}$$

Accordingly, given $\left(\bar{\Pi}^D, w\right)$ the physician's optimal choice is:

- *If* $\bar{\Pi}^D \geq w$, *the limiting policy is ineffective and the physician's decision coincides with that in Lemma 1, part a).*

- *If* $\bar{\Pi}^D < w$, *the limiting policy is effective and*

$$\begin{array}{ll} \textit{if } a \in \left[0, \frac{2\bar{\Pi}^D}{k^2}\right] & \textit{the physician chooses dual practice and } \gamma^*(a) = \gamma^D(a) = \frac{k^2 a}{4} \\ \textit{if } a \in \left(\frac{2\bar{\Pi}^D}{k^2}, \frac{w_{\bar{\pi}}+\bar{\Pi}^D}{k^2}\right] & \textit{the physician chooses dual practice and } \gamma^*(a) = \hat{\gamma}\left(a, \bar{\Pi}^D\right) < \frac{k^2 a}{4} \\ \textit{if } a \in \left(\frac{w_{\bar{\pi}}+\bar{\Pi}^D}{k^2}, \bar{a}\right] & \textit{the physician chooses to work only in the private sector.} \end{array}$$

When earnings limitations are effective, dual practice depends (in a negative way) on the physician ability a and on the profitability parameter k: $\hat{\gamma}\left(a, \bar{\Pi}^D\right) = \frac{k^2 a}{2} - \frac{1}{2}\left(\bar{\Pi}^D + \sqrt{k^2 a\left(k^2 a - 2\bar{\Pi}^D\right)}\right) > 0$, which is a decreasing function of a. This means that the more able physicians, as well as those working in more profitable specialties, are constrained to be less involved in private practice if they work at all for the public sector. In other words, all the physicians above a certain level of ability or profitability will have the same utility. Hence, more able doctors in more profitable disciplines will be more tempted to work exclusively for the private sector.

Let us compare now these two types of limiting policies.

Proposition 5 *Both for* ***developing and developed economies****, a policy of limiting involvement in private practice* ***always dominates*** *a policy of limiting earnings from dual practice.*

The intuition for this result has to do with how the two policies affect different types of physicians. Overall, profit limitations have a milder effect on the amount of dual practice performed by physicians. Under a policy that limits profits to $\bar{\Pi}^D$, the more able physicians, those with $a > \frac{2\bar{\Pi}^D}{k^2}$, will be forced to allocate significantly less time to private practice in order to satisfy their earning constraint. Meanwhile dual-practicing physicians with a relatively low ability are not constrained by this policy because even if they engage in a high amount of dual practice their earnings are relatively low. In contrast, policies that limit involvement directly target the intensity of dual practice and are therefore more effective in limiting its costs. It is important to highlight that the dominance of involvement limits over income limits is fairly general: it does not depend on the particular characteristics of the health care system under consideration and therefore applies to both developing and more developed economies.[23]

[23] In fact this result can be extended to a more general model without the need to resort to particular functional forms. Moreover, the result is also robust to assuming the public sector payment to be dependent on physician ability. The details are available from the authors upon request.

The next step is to characterize the shape of the optimal limiting policy for the two alternative health care systems under consideration. From Lemma 3 we see that if the limit is too soft ($\bar{\gamma} \geq \frac{w}{2}$), then the policy is ineffective as the maximum-involvement constraint is not binding for any of the physicians that actually work for the public sector, and we are trivially back to the laissez-faire scenario. Therefore, the HA solves

$$\max_{w,\bar{\gamma}} \mathcal{W}^{\bar{\gamma}} = \int_0^{\frac{4\bar{\gamma}}{k^2}} \frac{F(a)}{1+\frac{\delta k^2 a}{4}} da + \int_{\frac{4\bar{\gamma}}{k^2}}^{\frac{w+2\sqrt{\bar{\gamma}(w-\bar{\gamma})}}{k^2}} \frac{F(a)}{1+\delta\bar{\gamma}} da - (1+\lambda)\, w \left(\frac{w+2\sqrt{\bar{\gamma}\,(w-\bar{\gamma})}}{k^2} \right) \tag{5}$$

$$s.t. \quad w \geq 0, \; \frac{w+2\sqrt{\bar{\gamma}\,(w-\bar{\gamma})}}{k^2} \leq \bar{a}, \; 0 \leq \bar{\gamma} \leq \frac{w}{2}.$$

Let us first consider a developed economy,

Proposition 6 *In* ***developed economies,*** *when the HA can limit physician involvement in private practice,*

- *It is never optimal to fully ban dual practice, i.e., $\bar{\gamma} = 0$ is never a solution.*
- *If $\delta \leq \frac{2(1+\lambda)}{\varphi}$ the best the HA can do is not to limit dual practice.*
- *If $\delta > \frac{2(1+\lambda)}{\varphi}$ there exists an optimal limit to the amount of dual practice.*

Two main insights emerge from Proposition 6. First, no matter how large the cost of dual practice, it is never in the best interest of the HA to ban it. The policy to limit dual practice is sufficiently rich so as to cope with different degrees of productivity loss. Secondly, there are values of the productivity loss ($\delta < \frac{2(1+\lambda)}{\varphi}$) for which it is in the best interest of the HA not to limit dual practice at all. The reason is that any limiting policy will reduce the profitability of dual practice and thus incline physicians towards working exclusively in the private sector. If the HA wants to keep those workers in the public sector it has to compensate them by paying a higher salary. For this reason, only when the cost of dual practice is sufficiently large does the HA find it profitable to incur the extra cost (higher wages) of imposing a limit on dual practice. In the proof of this result it is also clear that the decision to restrict dual practice does not depend on k, although k will affect the number of physicians eventually hired in the public sector.

Now let us consider developing countries, for which the results are substantially different.

Proposition 7 *In developing economies, when the HA can limit physician involvement in private practice*

- *For high values of k the best the HA can do is not to limit dual practice.*

- *For intermediate values of k there is an optimal limit to the involvement in private practice.*
- *For low values of k the best the HA can do is either to limit dual practice (if δ is low) or to ban it (if δ is high).*

The results for developing economies sharply differ from those in the previous scenario. The attractiveness of the private sector (measured by k) turns out to be the key variable when characterizing the optimal policy. When the private alternative is very attractive, the establishment of limits to dual practice is never optimal. In this case, setting a limiting policy would make it very expensive to keep highly skilled physicians and the loss of such professionals would severely undermine the HA's capacity to provide health. When the private sector is relatively unattractive, a limiting policy might be also not optimal for the opposite reason: in this case, it is relatively cheap to retain physicians at public facilities and, thus, when the cost of dual practice δ is sufficiently high, the HA is better off banning rather than limiting dual practice. Thus banning dual practice can emerge in developing economies as the best intervention.[24] Finally, when the attractiveness of the private sector is moderate limits are always optimal in developing countries. The reason is two-fold. On the one hand, setting limits can help to reduce the loss of efficiency associated with dual practice without the risk of brain-drain, i.e. losing high-ability physicians to the private sector. But, on the other hand, keeping physicians in the public sector is not cheap enough to justify banning dual practice altogether.

6 The Optimal Policy-Mix to Regulate Dual Practice

In this section we combine previous results to present a comprehensive picture of the policy options available to the HA for the regulation of dual practice, and we offer some policy implications.

6.1 The Health Authority's Choice

Combining the propositions discussed in previous sections we obtain the following result:

Proposition 8 *The optimal decision of the HA is*

- *In **developed economies,***

[24] Our intuition is that the conclusion that for low values of k banning dominates if δ is high does not depend on the class of convex cost of effort functions we have considered. Considering a strictly convex quadratic disutility function is, however, much more cumbersome.

  - *If* $\delta \leq \frac{2(1+\lambda)}{\varphi}$ *not to regulate dual practice.*
  - *If* $\delta > \frac{2(1+\lambda)}{\varphi}$ *to impose a limit (but never a ban) on the physician involvement in dual practice.*

- *In* ***developing economies*** *the results in Proposition 7 directly apply.*

In developed countries we have shown that it suffices to concentrate on the decision of whether to limit physicians' involvement in the private sector. Note that this policy (whose extreme cases are analyzed in Section 3) also dominates exclusive contracts. Thus, the choice of optimal intervention depends on the cost of the dual practice. When this cost is low, the best policy is to leave dual practice unregulated. When the cost of dual practice is sufficiently severe, the best policy is to limit physicians' capacity to engage in dual practice. While the intensity of the productivity losses caused by dual practice will determine the stringency of this limit, banning dual practice is never worthwhile. An important insight that emerges from this comparison is the suboptimality of rewarding policies as a way to handle the negative consequences of dual practice. Although Proposition 3 states that for intermediate values of δ, $\delta \in \left(\frac{4(1+\lambda)}{\varphi}, \frac{8(1+\lambda)}{\varphi}\right]$, exclusive contracts are preferable to the extremes of laissez-faire and banning, for these values of δ it is even better to limit physicians' capacity to engage in dual practice.

For developing countries the comparison of the different regulatory policies is easier, as exclusive contracts are never optimal. The most important variable for determining the best policy mix turns out to be the attractiveness of the private sector. If the private sector is very attractive (i.e., k is high), then regardless of the cost of dual practice the HA should not impose any regulation. The reason is that any intervention would trigger a severe brain-drain of the most skilled professionals to the private sector and, because of the degree to which health provision in such countries depends on individual characteristics of physicians (due to less stringent practice protocols, etc.), this drain would severely damage the public provision of health care. In reverse, the same argument can be used to explain why a relatively unattractive private sector can result in the optimality of banning dual practice altogether. A limiting policy is optimal for intermediate values of k, i.e., not so low as to make banning affordable, and not so high as to trigger a brain-drain. Likewise, for low values of k setting a limit to the involvement in dual practice is the best policy when the cost of dual practice is relatively low.

Figures 1 and 2 illustrate the optimal policy choice for the two different scenarios considered, based on the results in Propositions 7 and 8.

[Insert Figures 1 and 2 here]

6.2 Policy Implications

We now describe the most important policy guidelines that can be extracted from our work and suggest some possible extensions of the analysis.

The relevance of the private sector attractiveness

One of the key variables for our results is private sector attractiveness (k). It seems clear that in practice this variable presents a wide and probably multi-dimensional heterogeneity. Specifically, we think it is important to consider (i) differences between developing and developed countries, and (ii) differences within countries (among specialities, or between primary physicians and specialists).

Regarding the first source of heterogeneity, we expect that the value of k will be high in developing countries where there are substantial revenue differences between public and private sectors and also between specialities. This fact points not only to a general tendency for physicians to be inclined toward heavier involvement in private provision, but also to the risk of brain-drain, i.e. loss of the most highly skilled medical professionals. This problem is exacerbated when limits are imposed on dual practice as a way to obtain extra revenue. With regard to this problem, our model predicts that in countries where the private alternative is very attractive the argument against regulating dual practice is correspondingly strong: if dual practice is regulated, the recruitment and retention of highly skilled physicians in the public sector becomes prohibitively expensive.

If, on the contrary, we consider k as speciality-specific, we observe that for specialties with a large private sector attractiveness, the health authority will choose to hire few physicians and provide a small level of public health. In other words, a high level of k points toward the crowding out of public provision by increased private provision. This effect is reinforced in our set-up by the fact that we have not imposed a lower bound on the amount of public health that should be guaranteed. Our setting, nevertheless, could be adapted to encompass circumstances in which there are specialities with large k which are deemed indispensable for the public sector (such as anesthesiologists, for instance) and, hence, whose level of production cannot be substantially reduced. In this case, our analysis would suggest that such essential specialities should receive higher salaries and softer regulations regarding dual practice.

Enforceability of policies

Our analysis makes the best-case assumption that policies are enforceable at zero cost, and hence ignores enforcement issues that can be important to practical policy application. However, we admit that the implementation of such regulations is seldom an easy task, especially in developing countries where the institutional and contracting environments are often weak.[25]

[25] With regard to this issue, Eggleston and Bir (2006) argue that the social trade-off between the benefits

Although we have not included enforceability concerns in our analysis, we can make a few observations regarding this important issue. First, one might argue that it may be easier to control earnings than involvement. This may offer an explanation to why some countries seem compelled to use this regulatory tool despite our finding that limits on involvement are, ceteris paribus, more efficient. Secondly, in the same vein, encouraging public physicians to perform private practice in public facilities may facilitate the monitoring of actual involvement in dual practice and thus aid in the enforcement of limiting policies. This is consistent with the pattern of several European countries, as described in the Introduction. Thirdly, regarding rewarding policies, these may be easy to enforce, or at least easier than any limiting policy. Thus, in more developed countries we can rationalize the use of exclusive contracts to induce some physicians to give up dual practice as a second best choice (when other kinds of policies are difficult to enforce).

The cost of raising public funds

In our model the HA maximizes net profits, i.e. the value associated with the production of health minus the wages paid to the physicians. The wage costs are weighted by λ, that represents the marginal cost of allocating more resources to public health provision. In this respect, if one expects that during an economic recession λ is higher due to more stringent budget constraints, our model provides an argument in favour of non-regulation both in developed and developing countries. Since any regulation makes the hiring of practitioners more expensive, whenever the budget is tight it is clear that the best policy is not to control dual practice.

Also, considering different values for λ provides an alternative way to compare developed and developing economies. In this respect, Auriol and Picard (2009) report that for developed countries λ has been estimated to be low, around 0.3 (Snow and Warren, 1996), while in developing countries it is three times higher, 0.9 (World Bank, 1998), and it can be much higher in countries heavily indebted. If we use these differences in λ as the basis for the comparison between developed and less developed economies, then our results predict that in developed countries it is more likely that dual practice should be limited, while in developing economies it should remain unregulated.

The cost of dual practice

The results in this paper depend on the cost of dual practice in terms of public performance. Theoretical analyses on the effects of dual practice on public health provision are scarce and show that this practice might bring about both positive and negative effects. It appears, however, that the arguments about the negative consequences of physician dual practice dominate the literature. Ferrinho et al. (2004), for instance, provide some evidence of the consequences of dual practice in terms of extracting more income from

and costs of dual practice crucially depend on the quality of a country's contracting institutions.

patients, reducing time served in public health posts, conflicts of interest, and even corruption, which suggests that the cost of dual practice, which is summarized by δ in our model, is positive. However, the real cost of dual practice for health systems is an empirical issue and there are no reliable studies that estimate this cost. Still, one would expect the value of δ be higher, due to weaker monitoring, mild self-regulation, etc., in developing countries. Interestingly, our model shows that while large values of δ point to the use of limits in more developed economies, this is not necessarily the case in developing countries, where the attractiveness of the private sector is crucial and may point to no intervention as the best option.

Health production technology

One may reasonably argue that the average productivity of the health care system in a developing country is lower than that of a developed country (that is why a developed country has chosen the technology φ, and corresponds to having $\varphi > \frac{f}{2}\bar{a}$). This difference suggests a new argument in favour of limiting dual practice in developed countries while de-regulating it in developing ones. This argument follows from our findings that in both economies lower technology implies less interest on the part of the HA to regulate dual practice.

Finally, as previously mentioned, within a given health care system different specialties may have access to different health technologies. Restating our results in this regard would suggest that specialities that are less technology-intensive and where the discretionality of doctors is higher are, ceteris paribus, worse candidates for regulation.

7 Conclusions

Dual practice is a complex phenomenon occurring in the public health systems of many developed and developing countries. In this paper we have considered some of the important factors that determine the optimal regulation for this practice and discussed different policy options. We have analyzed the optimal regulation under different hypotheses concerning the public health production function (as a way of describing the situations of different countries) and various policy instruments. The desirability of these instruments depends on the government ability to control physician dual practice but, more importantly, on the specific characteristics of the health sector in question.

In a very simple set-up our analysis has provided several interesting insights regarding the optimal regulation of dual practice. First, we have found that forbidding dual practice is seldom optimal, as it usually expels valuable professionals—indeed, the most valuable, if the private market rewards quality—from the public system. In this sense, dual practice can serve to the budgetary expenses needed to retain high-skilled physicians working in public facilities. Secondly, focusing on limiting policies, we have shown that limiting income is always less effective than limiting involvement. The reason is that the former policy has a milder effect on the amount of dual practice performed, as it only affects the high skilled physicians that are compelled to reduce private involvement in order to satisfy their earning constraint. Finally, our analysis has suggested that policy recommendations are different for more developed and developing economies, thus offering theoretical support for the desirability of different regulations in different economic environments. In developed countries the key factor is the potential negative effect of dual practice on public performance: when this effect is low the best option is not to intervene; when it is sufficiently high the best option is to impose a limit on physician involvement. Rewarding policies, i.e. those that pay an extra amount to physicians who give up their private practice, are only desirable when limitations are difficult to enforce. For developing countries, the design of the optimal policy is more complex as it also depends on the attractiveness of the private sector. When this attractiveness is very high the best option is not to intervene and thereby avoid an exodus of highly skilled physicians from the public sector. When it takes an intermediate value, then limits on the involvement are desirable. Finally, if the potential gains from private practice are low, the optimal intervention is either to limit dual practice (if the associated costs are low) or to ban it (if such costs are high). Rewarding contracts are never optimal in these countries as those physicians that would accept them are the ones with the smallest capacity to contribute to the production of health.

In our view, the main contribution of our paper is to identify the multiple factors that should guide the choice of the optimal regulatory policies on dual practice. Certainly, in order to evaluate the performance of countries in regard to dual practice and be able to make informed country-specific recommendations, more theoretical and empirical work is needed. Still, we believe that this work can enrich the discussion on dual practice and contribute to the development of a better policy making process.

References

[1] Auriol, E. and Picard, P.M. (2009). Infrastructure and Public Utilities Privatization in Developing Countries. *The World Bank Economic Review* 23 (1), 77-100.

[2] Barros, P.P. and Martínez-Giralt, X. (2002). Public and Private Provision of Health Care. *Journal of Economics and Management Strategy* 11(1), 109-133.

[3] Barros, P.P. and Olivella, P. (2005). Waiting Lists and Patient Selection. *Journal of Economics and Management Strategy* 14, 623-646.

[4] Biglaiser, G. and Ma, C.T. (2007). Moonlighting: Public Service and Private Practice. *Rand Journal of Economics* 38 (Winter), 1113-1133.

[5] Brekke, K.R. and Sørgard, L. (2007). Public versus private health care in a national health service. *Health Economics* 16 (6), 579-601.

[6] Delfgaauw, J. (2007). Dedicated Doctors: Public and Private Provision of Heath Care with Altruistic Physicians. Tinbergen Institute Discussion Papers 07-010/1.

[7] Eggleston, K. and Bir, A. (2006). Physician dual practice. *Health Policy* 78, 157-166.

[8] European Observatory on Health Systems and Policies. (2004). Snapshots of Health Systems: The state of affairs in 16 countries in summer 2004. Copenhagen, WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

[9] Ferrinho, P., Lerberghe, W.V., Fronteira, I., Hipólito, F. and Biscaia, A. (2004). "Dual practice in the health sector: review of the evidence". *Human Resources for Health* 2, 1-17.

[10] Flood, C.M. and Archibald, T. (2001). The illegality of private health care in Canada. *Canadian Medical Association Journal* 164, 825-830.

[11] France, G., Taroni, F. and Donatini, A. (2005). The Italian health-care system. *Health Economics* 14, 187-202.

[12] García-Prado, A. and González, P. (2007). Policy and regulatory responses to dual practice in the health sector. *Health Policy* 84, 142–152.

[13] García-Prado, A. and González, P. (2011). Whom do physicians work for? An Analysis of Dual Practice in the Health Sector. *Journal of Health Politics, Policy and Law*, 36 (2), 265-294.

[14] Globerman, S. and Vining, A. (1998). A Policy Perspective on "Mixed" Health Care Financial Systems of Business and Economics. *Journal of Risk and Insurance* 65 (1), 57-80.

[15] González, P. (2004). Should Physicians' Dual Practice Be Limited? An Incentive Approach. *Health Economics* 13, 505-524.

[16] González, P. (2005). On a Policy of Transferring Public Patients to Private Practice. *Health Economics* 14, 513-527.

[17] Gruen, R., Anwar, R., Begum, T., Killingsworth, J. R. and Normand, C. (2002). Dual job holding practitioners in Bangladesh: An exploration. *Social Science and Medicine* 54, 267-279.

[18] Holmström, B. and Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization* 7 (Special Issue), 24-52.

[19] Iversen, T. (1997). The effect of a private sector on the waiting time of a national health service. *Journal of Health Economics* 16, 381-396.

[20] Jingqing, Y. (2006). The privatisation of professional knowledge in the public health sector in China. *Health Sociology Review* 15(1), 16-28.

[21] Jofre-Bonet, M. (2000). Health care: private and/or private provision. *European Journal of Political Economy* 16, 469-489.

[22] Mainiero, M.B. and Woodfield, C.A. (2008). Resident Moonlighting in Radiology. *Journal of the American College of Radiology* 5 (6), 766-769.

[23] Marchand, M. and Schroyen, F. (2005). Can a mixed health care system be desirable on equity grounds? *Scandinavian Journal of Economics* 107(1), 1-23.

[24] Mayta-Tristán, P., Dulanto-Pizzorni, A. and Miranda J. (2008). Low wages and brain drain: an alert from Peru. *The Lancet* 371, 1577.

[25] Peters, D.H., Yazbeck, R.R., Sharma, G.M., Ramana, V., Pritchett, L.H. and Wagstaff, A. (2002). Better health systems for India's poor: findings, analysis and options. Washington, World Bank.

[26] Rabinowitz, H.K. and Paynter, N.P. (2002). The Rural vs Urban Practice Decision. *Journal of the American Medical Association* 287(1), 113.

[27] Rickman, N. and McGuire, A. (1999). Regulating Providers' Reimbursement in a mixed market for health care. *Scottish Journal of Political Economy* 46, 53-71.

[28] Snow, A. and Warren, K. (1996). The Marginal Welfare Cost of Public Funds: Theory and Estimates. *Journal of Public Economics* 61, 289-305.

[29] Socha, K.Z. and Bech, M. (2011). Physician dual practice: A review of literature. *Health Policy* 102 (1), 1-7.

[30] Stepan, A. and Sommersguter-Reichmann, M. (2005). Monitoring political decision-making and its impact in Austria. *Health Economics* 14, S7-S23.

[31] Wiley, M.M. (2005) The Irish health system: development in strategy, structure, funding and delivery since 1980. *Health Economics* 14, 169-186.

[32] World Bank (1998). *World Development Indicators*. Washington D.C. World Bank.

A Appendix

A.1 Proof of Lemma 1.

Immediate from the comparison of U^{pub}; U^{D} and U^{Pv} when dual practice is allowed or banned.

A.2 Proof of Proposition 1.

The HA has to solve, and compare, two independent optimization problems.
Case laissez-faire: The HA solves Program (3) for $F(a)=\varphi$, subject to the constraints $w \geq 0$ and $w \leq \frac{\bar{a}k^2}{2}$. From the f.o.c. we obtain the following candidate to optimum

$$w^{LF} = \frac{\sqrt{1+\frac{\delta\varphi}{1+\lambda}}-1}{\delta} > 0 \text{ if } \bar{a} > \frac{2}{k^2\delta}\left(\sqrt{\frac{\delta\varphi}{\lambda+1}+1}-1\right) \tag{6}$$

$$w^{LF} = \frac{k^2\bar{a}}{2} \text{ otherwise.} \tag{7}$$

Note that $w^{LF}(\delta)$ is a decreasing function of δ, and $w^{LF}(\delta=0)=\frac{\bar{a}k^2}{2}$. Moreover, it can be checked that the objective function is concave and, hence, that for the interior solution the s.o.c. is fulfilled. Note that Assumption1 (i) implies $\bar{a} \geq \frac{2}{k^2\delta}\left(\sqrt{\frac{\delta\varphi}{\lambda+1}+1}-1\right)$, so we can concentrate on the interior solution. Evaluating the objective function in the optimal level of wage we have

$$\mathcal{W}^{LF}\left(w^{LF}\right) = \int_0^{\frac{2}{k^2\delta}\left(\sqrt{1+\frac{\delta\varphi}{1+\lambda}}-1\right)} \frac{\varphi}{\frac{1}{4}ak^2\delta+1}\,da - \frac{(1+\lambda)\,2}{k^2\delta^2}\left(\sqrt{1+\frac{\delta\varphi}{1+\lambda}}-1\right)^2 \tag{8}$$

The Envelope Theorem ensures that $\mathcal{W}^{LF}\left(w^{LF}\right)$ is decreasing in δ. If we evaluate in the two extreme values of δ we have

$$\begin{aligned}\lim_{\delta\to 0}\mathcal{W}^{LF}\left(w^{LF}\right) &= \frac{\varphi^2}{2\,(1+\lambda)\,k^2}\\ \lim_{\delta\to+\infty}\mathcal{W}^{LF}\left(w^{LF}\right) &= 0\end{aligned}$$

Case dual practice banned: The HA solves Program (4) for $F(a)=\varphi$, subject to the constraints $w \geq 0$ and $w \leq \bar{a}k^2$. Solving the f.o.c. we obtain the following candidate to optimum

$$\begin{aligned} w^B &= \frac{\varphi}{2(1+\lambda)} > 0 \text{ if } \frac{\varphi}{2(1+\lambda)k^2} \leq \bar{a}, & (9) \\ w^B &= \bar{a}k^2 \text{ otherwise.} & (10) \end{aligned}$$

Note that Assumption 1 (i) implies $\bar{a} \geq \frac{\varphi}{2(1+\lambda)k^2}$, so we can concentrate on the interior solution. Moreover, it can be checked that the objective function is concave and at the interior solution the s.o.c. is fulfilled. Evaluating the objective function in the optimal level of wage we have

$$\mathcal{W}^B\left(w^B\right) = \frac{\varphi^2}{4(1+\lambda)k^2}, \tag{11}$$

that does not depend on δ.

Comparison: By comparing $\mathcal{W}^{LF}\left(w^{LF}\right)$ and $\mathcal{W}^B\left(w^B\right)$, it follows directly that there exists a threshold $\bar{\delta}_1 > 0$ such that:

- If $0 < \delta < \bar{\delta}_1$, $\mathcal{W}^{LF}\left(w^{LF}\right) > \mathcal{W}^B\left(w^B\right)$ and, hence, the optimal decision is to allow dual practice.
- If $\delta > \bar{\delta}_1$, $\mathcal{W}^{LF}\left(w^{LF}\right) < \mathcal{W}^B\left(w^B\right)$ and, hence, the optimal decision is to ban dual practice.

We finally show that $\bar{\delta}_1 \simeq 5.988\frac{(1+\lambda)}{\varphi}$. Let us write $\delta = \frac{x(1+\lambda)}{\varphi}$. Then

$$\mathcal{W}^{LF}\left(\delta = \frac{x(1+\lambda)}{\varphi}\right) = \frac{\varphi^2}{(1+\lambda)k^2}\left[\frac{4}{x}\ln\left(\frac{1}{2}\sqrt{x+1}+\frac{1}{2}\right) - \frac{2}{x^2}\left(\sqrt{x+1}-1\right)^2\right]$$

that we have to compare with $\mathcal{W}^B\left(w^B\right)$ as defined in (11). Hence,

$$\mathcal{W}^{LF} < \mathcal{W}^B \Leftrightarrow \frac{4}{x}\ln\left(\frac{1}{2}\sqrt{x+1}+\frac{1}{2}\right) - \frac{2}{x^2}\left(\sqrt{x+1}-1\right)^2 < \frac{1}{4} \Leftrightarrow x > 5.988$$

A.3 Proof of Proposition 2.

The HA has to solve two independent optimization problems.

Case laissez-faire: The HA solves Program (3) for $F(a) = fa$, subject to the constraint that $w \geq 0$ and $w \leq \frac{\bar{a}k^2}{2}$. First note that for $\delta = 0$ the welfare function can be written as:

$$\max_w \mathcal{W}^{LF}(\delta = 0) = 2w^2\left(\frac{f}{k^2} - (1+\lambda)\right)$$

Hence, for $\delta = 0$ the solution is

$$w^{LF}(\delta = 0) = \begin{cases} \frac{\bar{a}k^2}{2} & \text{if } k < \sqrt{\frac{f}{1+\lambda}} \\ 0 & \text{if } k \geq \sqrt{\frac{f}{1+\lambda}} \end{cases}$$

For $\delta > 0$, solving the f.o.c. we obtain the following candidate to optimum

$$w^{LF} = \frac{2\left(f - (1+\lambda) k^2\right)}{\delta (1+\lambda) k^2}. \tag{12}$$

This wage level is positive only if $k < \sqrt{\frac{f}{1+\lambda}}$. Therefore, the candidate to solution is:

$$w^{LF} = \begin{cases} \min\left\{\frac{2\left(f-(1+\lambda)k^2\right)}{\delta(1+\lambda)k^2}, \frac{\bar{a}k^2}{2}\right\} & \text{if } k < \sqrt{\frac{f}{1+\lambda}} \\ 0 & \text{if } k \geq \sqrt{\frac{f}{1+\lambda}} \end{cases}$$

Assumption 1(ii) rules out the corner case with $k \geq \sqrt{\frac{f}{1+\lambda}}$. Moreover, it can be checked that, for the interior solution, the objective function is concave and, hence, that the s.o.c. is fulfilled. The value of the objective function at the optimal interior solution under laissez-faire contract is:

$$\mathcal{W}^{LF}\left(w^{LF}\right) = \int_0^{\frac{2\left(f-(1+\lambda)k^2\right)}{\delta(1+\lambda)k^4}} \frac{fa}{1+\delta\frac{k^2 a}{4}} da - (1+\lambda)\frac{2}{k^2}\left(\frac{f-(1+\lambda)k^2}{\delta(1+\lambda)k^2}\right)^2 \tag{13}$$

The Envelope Theorem ensures that $\mathcal{W}^{LF}\left(w^{LF}\right)$ is decreasing in δ. If we evaluate in the two extreme values of δ we have

$$\begin{aligned} \lim_{\delta\to 0} \mathcal{W}^{LF}\left(w^{LF}\right) &= \mathcal{W}^{LF}\left(w^{LF} = \frac{\bar{a}k^2}{2}\right) = \frac{\bar{a}^2}{2}\left(f - (1+\lambda)k^2\right) \\ \lim_{\delta\to +\infty} \mathcal{W}^{LF}\left(w^{LF}\right) &= 0 \end{aligned}$$

Case dual practice banned: The HA solves Program (4) for $F(a) = fa$, subject to the constraints that $w \geq 0$ and $w \leq \bar{a}k^2$. This objective function is monotone in w. Hence, the solution is always on the boundaries of the support. Either $w^B = \bar{a}k^2$ or $w^B = 0$ depending on whether $k \leq \sqrt{\frac{f}{2(1+\lambda)}}$ or not. Again, under Assumption 1 (ii), values of $k \geq \sqrt{\frac{f}{1+\lambda}}$ are not possible and we can concentrate on $k < \sqrt{\frac{f}{2(1+\lambda)}}$, which gives,

$$\mathcal{W}^B\left(w^B\right) = \begin{cases} \bar{a}^2\left(\frac{f}{2} - (1+\lambda)k^2\right) & \text{if } k \leq \sqrt{\frac{f}{2(1+\lambda)}} \\ 0 & \text{if } k \in \left(\sqrt{\frac{f}{2(1+\lambda)}}, \sqrt{\frac{f}{1+\lambda}}\right] \end{cases} \tag{14}$$

Comparison: Considering the value functions with and without banning we get,

- If $k \in \left(\sqrt{\frac{f}{2(1+\lambda)}}, \sqrt{\frac{f}{1+\lambda}}\right]$ then $\mathcal{W}^B\left(w^B\right) < \mathcal{W}^{LF}\left(w^{LF}\right)$
- If $k \leq \sqrt{\frac{f}{2(1+\lambda)}}$ then, there exists a threshold $\bar{\delta}_2 > 0$ such that,
  - If $\delta < \bar{\delta}_2$ then $\mathcal{W}^B\left(w^B\right) < \mathcal{W}^{LF}\left(w^{LF}\right)$
  - If $\delta > \bar{\delta}_2$ then $\mathcal{W}^B\left(w^B\right) > \mathcal{W}^{LF}\left(w^{LF}\right)$.

A.4 Proof of Lemma 2.

From the physician's utility under the different choices U^{pub}, U^{D} and U^{Pv}, we obtain the results presented in the lemma.

A.5 Proof of Proposition 3.

We need to distinguish two cases depending on whether $\Delta > w^E$ or $\Delta \leq w^E$. If we are in the case with $\Delta > w^E$ then no physician works as dual provider. In this case, trivially, the best contract is the optimal banning contract (as defined in Lemma 2). Therefore, in the region $\Delta > w^E$, the best contract yields a value function

$$\mathcal{W}^E\left(\Delta = \frac{\varphi}{2(1+\lambda)},\ w^E = 0\right) = \frac{\varphi^2}{4(1+\lambda)k^2}$$

We need to focus, therefore, on the case with $\Delta \leq w^E$. The objective function of the HA in this case is

$$\max_{w^E,\Delta} \mathcal{W}^E = \int_0^{\frac{2\Delta}{k^2}} \varphi da + \int_{\frac{2\Delta}{k^2}}^{\frac{2w^E}{k^2}} \frac{\varphi}{1+\delta\frac{k^2 a}{4}} da - (1+\lambda)\left(\frac{2\Delta}{k^2}\left(w^E+\Delta\right) + \left(\frac{2w^E}{k^2} - \frac{2\Delta}{k^2}\right) w^E\right),$$

subject to the constraints, $w^E \geq 0$, $w^E \leq \frac{\bar{a}k^2}{2}$ and $0 \leq \Delta \leq w^E$. We first ignore the constraint $w^E \leq \frac{\bar{a}k^2}{2}$ and check afterwards that it is satisfied by the solution. The f.o.c.'s of this program are:

$$\begin{aligned}
\frac{\partial \mathcal{W}^E}{\partial w} &= \frac{2(1+\lambda)}{k^2}\left[\frac{2\varphi}{(1+\lambda)(2+\delta w)} - 2w\right] = 0 \\
\frac{\partial \mathcal{W}^E}{\partial \Delta} &= \frac{2(1+\lambda)}{k^2}\left[\frac{\varphi}{(1+\lambda)}\left(1 - \frac{1}{2+\delta\Delta}\right) - 2\Delta\right] = 0
\end{aligned}$$

From here it follows that $w^E = \frac{\sqrt{1+\delta\frac{\varphi}{1+\lambda}}-1}{\delta} > 0$ (the s.o.c. is trivially fulfilled). Regarding Δ there are two candidates that verify the f.o.c. First, $\Delta^* = 0$ that fulfills the s.o.c. provided $\delta \leq \frac{4(1+\lambda)}{\varphi}$. Secondly, $\Delta^* = \frac{\varphi\delta - 4(1+\lambda)}{2\delta(1+\lambda)}$ that fulfills the s.o.c. only if $\delta \geq \frac{4(1+\lambda)}{\varphi}$. Finally, when $\Delta^* = \frac{\varphi\delta-4(1+\lambda)}{2\delta(1+\lambda)}$ we also have to check the constraint $\Delta^* \leq w^*$:

$$\Delta^* \leq w^* \iff \delta \leq \frac{8(1+\lambda)}{\varphi}.$$

Therefore, the solution in the region $\Delta \leq w^E$ is

$$\begin{aligned}
w^E &= \frac{\sqrt{1+\delta\frac{\varphi}{1+\lambda}}-1}{\delta} \\
\Delta &= \begin{cases} \frac{\varphi\delta-4(1+\lambda)}{2\delta(1+\lambda)} & \text{if } \delta \in \left[\frac{4(1+\lambda)}{\varphi}, \frac{8(1+\lambda)}{\varphi}\right] \\ 0 & \text{if } \delta \leq \frac{4(1+\lambda)}{\varphi} \end{cases}
\end{aligned}$$

It only rests to evaluate the HA's objective function in the solution of this case and compare it with $\mathcal{W}^E\left(\Delta = \frac{\varphi}{2(1+\lambda)},\ w^E = 0\right)$.

The Envelope Theorem ensures that $\mathcal{W}^E\left(\Delta = \max\left\{\frac{\varphi\delta-4(1+\lambda)}{2\delta(1+\lambda)}, 0\right\},\ w^E = \frac{\sqrt{1+\delta\frac{\varphi}{1+\lambda}}-1}{\delta}\right)$ is decreasing in δ. Therefore the value of the objective function is bounded below by the value it would take for the upper bound of δ (i.e., $\delta = \frac{8(1+\lambda)}{\varphi}$).

It can be shown that,

$$\lim_{\delta\to\frac{8(1+\lambda)}{\varphi}} \mathcal{W}^E\left(\Delta^* = \max\left\{\frac{\varphi\delta - 4(1+\lambda)}{2\delta(1+\lambda)}, 0\right\},\ w^{E*} = \frac{\sqrt{1+\delta\frac{\varphi}{1+\lambda}}-1}{\delta}\right) = \frac{\varphi^2}{4(1+\lambda)k^2},$$

and this is equal to $\mathcal{W}^E\left(\Delta^* = \frac{\varphi}{2(1+\lambda)},\ w^* = 0\right)$. In addition, it is easy to check that the constraint $w^E \leq \frac{\bar{a}k^2}{2}$ is satisfied by the solution. Therefore, we have shown that,

- For every $\delta < \frac{8(1+\lambda)}{\varphi}$, the solution is

$$\begin{aligned} w^E &= \frac{\sqrt{1+\frac{\delta}{1+\lambda}\varphi}-1}{\delta} \\ \Delta &= \begin{cases} \frac{\varphi\delta-4(1+\lambda)}{2\delta(1+\lambda)} & \text{if } \delta \in \left[\frac{4(1+\lambda)}{\varphi}, \frac{8(1+\lambda)}{\varphi}\right] \\ 0 & \text{if } \delta \leq \frac{4(1+\lambda)}{\varphi} \end{cases} \end{aligned}$$

  since $\mathcal{W}^E\left(\Delta = \max\left\{\frac{\varphi\delta-4(1+\lambda)}{2\delta(1+\lambda)}, 0\right\},\ w^E = \frac{\sqrt{1+\delta\frac{\varphi}{1+\lambda}}-1}{\delta}\right) > \mathcal{W}^E\left(\Delta = \frac{\varphi}{2(1+\lambda)},\ w^E = 0\right)$.

- For every $\delta > \frac{8(1+\lambda)}{\varphi}$, the solution is $w^E = 0$, and $\Delta = \frac{\varphi}{2(1+\lambda)}$.

A.6 Proof of Proposition 4.

As in the previous proposition, we need to distinguish two cases depending on whether $\Delta > w^E$ or $\Delta \leq w^E$. If we are in the case with $\Delta > w^E$ then no physician works as dual provider. In this case, trivially, the best contract is the optimal banning contract (as defined in Lemma 2). Therefore, in the region $\Delta > w^E$, the best contract yields a value function $\mathcal{W}^B\left(w^B\right)$ defined in (14).

We need to focus, therefore, on the case with $\Delta \leq w^E$. The objective function of the HA in this case is

$$\max_{w,\Delta} \mathcal{W}^E = f\left[\int_0^{\frac{2\Delta}{k^2}} a da + \int_{\frac{2\Delta}{k^2}}^{\frac{2w^E}{k^2}} \frac{a}{1+\delta\frac{k^2 a}{4}} da\right] - \frac{2(1+\lambda)}{k^2}\left(\Delta^2 + w^{2E}\right),$$

subject to the constraints, $w^E \geq 0$, $w^E \leq \frac{\bar{a}k^2}{2}$ and $0 \leq \Delta \leq w^E$. We ignore first the constraint $w^E \leq \frac{\bar{a}k^2}{2}$ and check afterwards that it is satisfied by the solution. The f.o.c.'s of this program are:

$$\begin{aligned}\frac{\partial \mathcal{W}^E}{\partial w} &= \frac{4}{k^2} w \left(\frac{2f}{k^2 (2 + w\delta)} - (1+\lambda) \right) = 0 \\ \frac{\partial \mathcal{W}^E}{\partial \Delta} &= \frac{4}{k^2} \Delta \left(f \left(1 - \frac{2}{(2 + w\Delta)} \right) - (1+\lambda) \right) = 0\end{aligned}$$

Assumption 1(ii) implies $k < \sqrt{\frac{f}{(1+\lambda)}}$ and there are two candidates that satisfy the f.o.c. for each variable:

$$\begin{aligned}w^* &= 0 \quad \text{or} \quad w^* = \frac{2\left(f - (1+\lambda) k^2\right)}{(1+\lambda) k^2 \delta} \\ \Delta^* &= 0 \quad \text{or} \quad \Delta^* = \frac{2(1+\lambda) k^2}{\delta \left(f - (1+\lambda) k^2\right)}\end{aligned}$$

When checking the s.o.c. it is easy to check that $\Delta^* = \frac{2(1+\lambda)k^2}{\delta(f-(1+\lambda)k^2)}$ cannot be a solution, as the s.o.c only holds for values of k such that $\Delta^* = \frac{2(1+\lambda)k^2}{\delta(f-(1+\lambda)k^2)} < 0$. Thus, there does not exist a solution with $\Delta^* > 0$ and $\Delta \leq w^E$. The optimal contract, therefore, will be the one in Proposition *2*. This also shows that constraint $w^E \leq \frac{\bar{a}k^2}{2}$ is satisfied by the solution.

A.7 Proof of Lemmas 3 and 4.

From the physician's utility under the different choices U^{pub}, U^D (that depends on the limiting policy) and U^{Pv}, we obtain the results presented in the lemma.

A.8 Proof of Proposition 5.

We first define each policy by a pair: $\left(\bar{\Pi}^D, w_{\bar{\pi}}\right)$ and $\left(\bar{\gamma}, w_{\bar{\gamma}}\right)$. Each policy, in turn, will determines a series of thresholds (as defined in Lemmas 3 and 4) that characterize the behavior of the physicians.

To do the proof, we show that for any possible earnings limitation, we can find a policy of limiting the involvement in private practice that is more efficient (it provides more health at the same costs).

Consider any policy of limiting private earnings $\left(\bar{\Pi}^D, w_{\bar{\pi}}\right)$. This contract can give rise to different scenarios. Let us study them independently:

Non-binding Policy: $(\bar{\Pi}^D \geq w_{\bar{\pi}})$ Consider that the limit to earnings is so high that it is not binding for any of the physicians that actually work for the public sector. In other words, the first physician that would be affected by the policy is one that already chooses to work solely in the private sector. In this case, as Lemma 4 states, the policy is irrelevant. Thus, any policy of limiting the involvement in private practice with the same salaries $w_{\bar{\gamma}} = w_{\bar{\pi}}$, and with a $\bar{\gamma}$ so high that is not binding for any physician (i.e., with $\bar{\gamma} > \frac{w_{\bar{\gamma}}}{2}$) is, by construction, as good as the original $\bar{\Pi}^D$-policy.

Binding policy: $(\bar{\Pi}^D < w_{\bar{\pi}})$ The limit is such that some physicians are unconstrained dual providers, while others are affected by the policy. Formally, following Lemma 4, under $\left(\bar{\Pi}^D, w_{\bar{\pi}}\right)$ the physicians with $a \leq \frac{w_{\bar{\pi}}+\bar{\Pi}^D}{k^2}$ are involved in dual practice. We will propose a policy that limits the involvement in private practice $(\bar{\gamma}, w_{\bar{\gamma}})$ that is more efficient than $\left(\bar{\Pi}^D, w_{\bar{\pi}}\right)$.

Consider a policy that sets $w_{\bar{\gamma}} = w_{_{\bar{\pi}}} = w$ and $\bar{\gamma}$ such that the physician that is indifferent between being a dual provider or leaving the public sector is the same under the two policies. From Lemmas 3 and 4, this value of $\bar{\gamma}$ is such that

$$\frac{w + 2\sqrt{\bar{\gamma}\left(w - \bar{\gamma}\right)}}{k^2} = \frac{w + \bar{\Pi}^D}{k^2} \Longleftrightarrow 2\sqrt{\bar{\gamma}\left(w - \bar{\gamma}\right)} = \bar{\Pi}^D$$

In words, this means that the physician with ability $a = \frac{w+\bar{\Pi}^D}{k^2}$ (denote this threshold $\hat{a}$) when limited through maximum earnings $\bar{\Pi}^D$ will perform an amount of dual practice exactly equal to $\bar{\gamma}$. Function $F(\bar{\gamma}) = 2\sqrt{\bar{\gamma}\left(w - \bar{\gamma}\right)}$ gets value zero at $\bar{\gamma} = 0$ and $\bar{\gamma} = w$, and has a maximum at $\bar{\gamma} = \frac{w}{2}$, where $F(\bar{\gamma} = \frac{w}{2}) = w > \bar{\Pi}^D$. Hence, $F(\bar{\gamma}) = \bar{\Pi}^D$ is defined by the pair $\bar{\gamma} = \frac{1}{2}\left(w \pm \sqrt{w^2 - \left(\bar{\Pi}^D\right)^2}\right)$, which are always well defined since we are in the region with $\bar{\Pi}^D < w_{\bar{\pi}}$. Nevertheless, only the negative root is compatible with the restriction for a policy of limiting the involvement to be binding for some physicians, i.e., $\bar{\gamma} \leq \frac{w}{2}$. Therefore, we set $\bar{\gamma} = \frac{1}{2}\left(w - \sqrt{w^2 - \left(\bar{\Pi}^D\right)^2}\right) \in \left(0, \frac{w}{2}\right)$.

Now, let us compare the amount of dual practice that dual providers exert with each policy.

Under the $\bar{\gamma}$-policy

if $a \in \left[0, \frac{4\bar{\gamma}}{k^2}\right]$	the physician chooses dual practice and $\gamma^*\left(a\right) = \gamma^D\left(a\right) = \frac{k^2 a}{4}$
if $a \in \left(\frac{4\bar{\gamma}}{k^2}, \hat{a}\right]$	the physician chooses dual practice and $\gamma^*\left(a\right) = \bar{\gamma}$

Under the $\bar{\Pi}^D$-policy

if $a \in \left[0, \frac{2\bar{\Pi}^D}{k^2}\right]$	the physician chooses dual practice and $\gamma^*\left(a\right) = \frac{k^2 a}{4}$
if $a \in \left(\frac{2\bar{\Pi}^D}{k^2}, \hat{a}\right]$	the physician chooses dual practice and $\gamma^*\left(a\right) = \hat{\gamma}\left(a, \bar{\Pi}^D\right) < \frac{k^2 a}{4}$

Those physicians in $a \in \left(\frac{2\bar{\Pi}^D}{k^2}, \hat{a}\right]$ do an amount of dual practice $\hat{\gamma}\left(a, \bar{\Pi}^D\right)$ such that

$$\Pi^D(\hat{\gamma}\left(a\right)) = 2k\sqrt{a\hat{\gamma}\left(a, \bar{\Pi}^D\right)} - 2\hat{\gamma}\left(a, \bar{\Pi}^D\right) = \bar{\Pi}^D$$

Note that we have constructed $\bar{\gamma}$ in such a way that $\hat{\gamma}\left(\hat{a},\bar{\Pi}^D\right)=\bar{\gamma}$. This, together with the fact that $\hat{\gamma}\left(a,\bar{\Pi}^D\right)$ is decreasing in a implies that for every $a\in\left[\frac{2\bar{\Pi}^D}{k^2},\hat{a}\right)$ we have $\hat{\gamma}\left(a,\bar{\Pi}^D\right)>\bar{\gamma}$. Finally, we need to check that $\frac{4\bar{\gamma}}{k^2}<\frac{2\bar{\Pi}^D}{k^2}$ or, analogously, that $2\bar{\gamma}<\bar{\Pi}^D$. This is true for any binding $\bar{\Pi}^D$-policy, since,

$$2\bar{\gamma}=\left(w-\sqrt{w^2-\bar{\Pi}^{D2}}\right)<\bar{\Pi}^D\Leftrightarrow\bar{\Pi}^D<w$$

With this, we have that the amount of dual practice performed by the physicians under the two policies is:

Level of ability	$\bar{\gamma}$-policy	$\bar{\Pi}^D$-policy
$a\in\left[0,\frac{4\bar{\gamma}}{k^2}\right)$	$\gamma^*(a)=\gamma^D(a)=\frac{k^2a}{4}$	$\gamma^*(a)=\gamma^D(a)=\frac{k^2a}{4}$
$a\in\left[\frac{4\bar{\gamma}}{k^2},\frac{2\bar{\Pi}^D}{k^2}\right)$	$\bar{\gamma}$	$\gamma^*(a)=\gamma^D(a)=\frac{k^2a}{4}>\bar{\gamma}$
$a\in\left[\frac{2\bar{\Pi}^D}{k^2},\hat{a}\right)$	$\bar{\gamma}$	$\hat{\gamma}\left(a,\bar{\Pi}^D\right)>\bar{\gamma}$
$a=\hat{a}$	$\bar{\gamma}$	$\hat{\gamma}\left(\hat{a},\bar{\Pi}^D\right)=\bar{\gamma}$

Under the $\bar{\gamma}$-policy, some physicians, all those in the range of abilities $a\in\left[\frac{4\bar{\gamma}}{k^2},\hat{a}\right)$, do less dual practice under the $\bar{\gamma}$-policy than under the $\bar{\Pi}^D$-policy. Therefore, the $\bar{\gamma}$-policy dominates as it implies paying the same wages, having the same amount of physicians working in the public sector, but a lower amount of dual practice, what causes a lower aggregate productivity loss.

A.9 Proof of Proposition 6.

The HA solves Program (5), for $F(a)=\varphi$.

We make some manipulations on the objective function in order to work with a more compact optimization program. This is done without loss of generality. First, we do a change of variable and define $\alpha\equiv\frac{\bar{\gamma}}{w}\in[0,\frac{1}{2}]$, where $\alpha=0$ corresponds to $\bar{\gamma}=0$ (banning dual practice) and $\alpha=\frac{1}{2}$ corresponds to $\bar{\gamma}=\frac{w}{2}$ (laissez-faire). We also force a common factor $\frac{1+\lambda}{\delta^2k^2}$ to the whole function, this yields:

$$\mathcal{W}^{\alpha}{=}\frac{1+\lambda}{\delta^2k^2}\left(\frac{\varphi\delta}{1+\lambda}\left(4\ln(1+\alpha\delta w)+\frac{\delta w\left(1+2\sqrt{\alpha-\alpha^2}-4\alpha\right)}{1+\alpha\delta w}\right)-(\delta w)^2\left(1+2\sqrt{\alpha-\alpha^2}\right)\right),$$

which shows that the solution to the program will be independent of the parameter k. We finally rename the combined parameter $\frac{\delta\varphi}{1+\lambda}$ as xe, the product δw as $\tilde{w}$, and $A\equiv\delta k^2\bar{a}$, that by Assumption 1 (i) implies $A\geq x$. The previous optimization program is equivalent to solving:

$$\begin{aligned}\max_{\tilde{w},\alpha} W &= \left(x\left(4\ln(1+\alpha\tilde{w})+\frac{\tilde{w}\left(1+2\sqrt{\alpha-\alpha^2}-4\alpha\right)}{1+\alpha\tilde{w}}\right)-\tilde{w}^2\left(1+2\sqrt{\alpha-\alpha^2}\right)\right)\\ \text{s.t. } \tilde{w} &\geq 0,\ \tilde{w}\left(1+2\sqrt{\alpha-\alpha^2}\right)\leq A \text{ and } \alpha\in\left[0,\frac{1}{2}\right],\end{aligned}$$

This program is simpler (but equivalent) to the original one. The variable that determines the intensity of the limiting policy, α, is defined over a compact set and, moreover, there is only one parameter that is relevant for the optimization (x) instead of three (λ, φ, δ) in the original program.

To solve this program, using Kuhn-Tucker conditions, we have to consider the five candidates to solution. These candidates are:

(i) $\tilde{w}=0$, in which case $W=0$.

(ii) $\tilde{w}=\frac{A}{\left(1+2\sqrt{\alpha-\alpha^2}\right)}$, in which case W is a function of α parametrized by A:

$$W=x\left(4\ln\left(1+\alpha\frac{A}{\left(1+2\sqrt{\alpha-\alpha^2}\right)}\right)+\frac{A\left(1+2\sqrt{\alpha-\alpha^2}-4\alpha\right)}{\left(1+2\sqrt{\alpha-\alpha^2}\right)+\alpha A}\right)-\frac{A^2}{\left(1+2\sqrt{\alpha-\alpha^2}\right)}.$$

Note that when x is small as compared to A, this candidate leads to negative values of W. In fact, often, even when x is very close to A, under this candidate W is negative. The maximum in $\alpha(A)$ is difficult to find analytically but it can be computed numerically and compared to the other candidates to solution.

(iii) $\alpha=0$, which has associated $\tilde{w}(\alpha=0)=\frac{x}{2}$ if $\frac{x}{2}<A$ (which is always the case under Assumption 1 (i), $A\geq x$), and leads to $W^{\alpha=0}=\frac{x^2}{4}$.

(iv) $\alpha=\frac{1}{2}$, which has associated $\tilde{w}(\alpha=\frac{1}{2})=\sqrt{1+x}-1$, if $2\left(\sqrt{1+x}-1\right)<A$, which is always true under Assumption 1 (i) since it is implied by $A\geq x$, and yields $W^{\alpha=\frac{1}{2}}=4x\ln\left(\frac{1}{2}\sqrt{x+1}+\frac{1}{2}\right)-2\left(\sqrt{x+1}-1\right)^2$.

(v) The solution satisfying the two first order conditions:

$$\begin{aligned}\frac{\partial W}{\partial\alpha} &= \frac{\tilde{w}\left(x\left((1+\alpha\tilde{w})(1-2\alpha)-\tilde{w}\sqrt{\alpha-\alpha^2}\left(1-4\alpha+2\sqrt{\alpha-\alpha^2}\right)\right)-\tilde{w}(1+\alpha\tilde{w})^2(1-2\alpha)\right)}{\sqrt{\alpha-\alpha^2}(1+\alpha\tilde{w})^2}=0\\ \frac{\partial W}{\partial\tilde{w}} &= \frac{x\left(4\tilde{w}\alpha^2+2\sqrt{\alpha-\alpha^2}+1\right)-2\tilde{w}(1+\alpha\tilde{w})^2\left(1+2\sqrt{\alpha-\alpha^2}\right)}{(1+\alpha\tilde{w})^2}=0,\end{aligned}$$

which are equivalent to find $(\tilde{w},\alpha)$ from the system:

$$x\left((1+\alpha\tilde{w})(1-2\alpha)-\tilde{w}\sqrt{\alpha-\alpha^2}\left(1-4\alpha+2\sqrt{\alpha-\alpha^2}\right)\right)-\tilde{w}(1+\alpha\tilde{w})^2(1-2\alpha)=0 \tag{15}$$

$$x\left(4\tilde{w}\alpha^2+2\sqrt{\alpha-\alpha^2}+1\right)-2\tilde{w}(1+\alpha\tilde{w})^2\left(1+2\sqrt{\alpha-\alpha^2}\right)=0. \tag{16}$$

> The complexity of this system of equations prevents us from achieving an explicit algebraic solution. However, a numerical solution, and its corresponding welfare $W(\alpha^*(x), \tilde{w}^*(x))$, can be easily computed for each value of x. It can also be seen that for very low values of x, $x < 1.78$, there is no $(\alpha^*(x), \tilde{w}^*(x))$ satisfying the f.o.c.'s and the constraints.

Finally, to identify the solution to the problem we have to compare the different candidates in order to conclude whether the solution is interior or at one of the corners. First, we compare candidates (iii) and (iv) and we find that $W^{\alpha=0} > W^{\alpha=\frac{1}{2}}$ if and only if $x > 5.988$. It can also be seen that candidate (iii) with $\alpha = 0$ is never a solution since $\frac{\partial W}{\partial \alpha}\left(\alpha = 0, \tilde{w} = \frac{x}{2}\right) > 0$ for all x. Candidate (ii) is always dominated by (iv) or (v). Consider, finally, the comparison between candidate (iv), i.e. $\alpha = \frac{1}{2}$ and the interior solution (v). For low values of x there is no interior solution and, hence, the optimum is trivially candidate (iv), $\alpha = \frac{1}{2}$ and $\tilde{w} = \sqrt{1+x} - 1$. For higher values of x, let us define $\Delta W^{LF} \equiv W^{\alpha=\frac{1}{2}} - W^{\alpha}(\alpha^*(x), \tilde{w}^*(x))$. It can be shown numerically that ΔW^{LF} is decreasing in x. Moreover, $\Delta W^{LF} = 0$ if and only if $x = 2$, i.e., if $\delta = \frac{2(1+\lambda)}{\varphi}$. Thus, summarizing, for any $x < 2$, the solution is $\left(\alpha = \frac{1}{2}, \tilde{w} = \sqrt{1+x} - 1\right)$ (laissez-faire), while for $x > 2$, the interior candidate $(\alpha^*(x), \tilde{w}^*(x))$ provides the highest welfare.

A.10 Proof of Proposition 7.

The HA solves Program (5), for $F(a) = fa$.

Without loss of generality, following the same strategy as in the proof of Proposition 6, we make some manipulations on the objective function in order to work with a more compact optimization program. After solving the integrals, forcing a common factor equal to $\frac{f}{2\delta^2 k^4}$, using $\alpha \equiv \frac{\bar{\gamma}}{w} \in [0, \frac{1}{2}]$ (where $\alpha = 0$ corresponds to $\bar{\gamma} = 0$ –banning dual practice– and $\alpha = \frac{1}{2}$ corresponds to $\bar{\gamma} = \frac{w}{2}$ –laissez-faire–) and, finally, defining $K \equiv \frac{k^2(1+\lambda)}{f}$, $\tilde{w} \equiv \delta w$ and $A \equiv \delta k^2 \bar{a}$, the optimization program can be rewritten as*:*

$$\max_{\tilde{w},\alpha} W^{\alpha} = \Bigg[\left(32\left(\alpha\tilde{w} - \ln\left(1+\alpha\tilde{w}\right)\right) + \frac{\tilde{w}^2\left(\left(1+2\sqrt{\alpha-\alpha^2}\right)^2 - 16\alpha^2\right)}{(1+\alpha\tilde{w})}\right) \\ -2\tilde{w}^2 K\left(1+2\sqrt{\alpha-\alpha^2}\right)\Bigg]$$

$$\text{s.t. } \tilde{w} \geq 0,\ \tilde{w}\left(1+2\sqrt{\alpha-\alpha^2}\right) \leq A \text{ and } \alpha \in \left[0, \frac{1}{2}\right].$$

This program is simpler (but equivalent) to the original one. The variable that determines the intensity of the limiting policy, α, is defined over a compact set and, moreover, there is only one parameter that is relevant for the optimization (K) instead of four (λ, f, δ, k) in the original program. Note that we only consider cases with $f > (1+\lambda)k^2$ (see Assumption 1 (ii)), what restricts the space of K to $K \in (0, 1)$.

To solve this program, using Kuhn-Tucker, we have to consider the five candidates to solution. These candidates are:

(i) $\tilde{w} = 0$, in which case $W = 0$.

(ii) $\tilde{w} = \frac{A}{\left(1+2\sqrt{\alpha-\alpha^2}\right)}$, in which case W is a function of α parametrized by $A:$

$$W = 32\left(\tfrac{\alpha A}{1+2\sqrt{\alpha-\alpha^2}} - \ln\left(1+\tfrac{\alpha A}{1+2\sqrt{\alpha-\alpha^2}}\right)\right) +$$
$$+\frac{\left(\frac{A}{1+2\sqrt{\alpha-\alpha^2}}\right)^2}{\left(1+\frac{\alpha A}{1+2\sqrt{\alpha-\alpha^2}}\right)}\left(\left(1+2\sqrt{\alpha-\alpha^2}\right)^2 - 16\alpha^2\right) - 2K\left(\tfrac{A}{1+2\sqrt{\alpha-\alpha^2}}\right)^2\left(1+2\sqrt{\alpha-\alpha^2}\right).$$

It can be shown that for the two extremes in α

$$W(\alpha = 0, \tilde{w} = A; K, A) = A^2\left(1-2K\right)$$
$$W\left(\alpha = \frac{1}{2}, \tilde{w} = \frac{A}{2}; K, A\right) = 32\left(\frac{A}{4} - \ln\left(1+\frac{A}{4}\right)\right) - KA^2.$$

And for very low values of K, $W(\alpha = 0, \tilde{w} = A; K, A)$ is larger than $W\left(\alpha = \frac{1}{2}, \tilde{w} = \frac{A}{2}; K, A\right)$ for all A. Notice, however, that the solution in this region might also be interior in α. Finding analytically the maximum in α is difficult, but doing it numerically we can show that in this candidate, $\alpha(A)$ is decreasing in A and is strictly positive. It can also be easily checked that the welfare for this candidate is decreasing in A.

(iii) $\alpha = 0$, which has associated $\tilde{w} = A$ if $K < \frac{1}{2}$ (which is a particular case of candidate (ii)), that gives profits $W^{\alpha=0} = A^2\left(1-2K\right)$; while the optimal wage is $\tilde{w} = 0$ if $K \geq \frac{1}{2}$ (which is a particular case of candidate (i)) and leads to $W = 0$.

(iv) $\alpha = \frac{1}{2}$, which has associated $\tilde{w} = 2\frac{1-K}{K}$ if $2\frac{1-K}{K} \leq \frac{A}{2}$, i.e., $A \geq \frac{4(1-K)}{K}$, with a welfare of

$$W\left(\alpha = \frac{1}{2}, \tilde{w} = 2\left(\frac{1-K}{K}\right)\right) = 32\left(\frac{1}{K} - 1\right) - 32\ln\left(\frac{1}{K}\right) - 16\frac{\left(1-K\right)^2}{K},$$

while for $A < \frac{4(1-K)}{K}$, the optimal wage is $\tilde{w} = \frac{A}{2}$ (which is a particular case of candidate (ii)) with an associated welfare of

$$W\left(\alpha = \frac{1}{2}, \tilde{w} = \frac{A}{2}\right) = 32\left(\frac{A}{4} - \ln\left(1+\frac{A}{4}\right)\right) - KA^2$$

(v) Finally, the candidates from the interior solution are the solution of the system formed by the two first order conditions of the optimization program,

$$\begin{aligned}\frac{\partial W}{\partial \alpha} &= \frac{\tilde{w}^2}{\sqrt{\alpha-\alpha^2}(1+\alpha\tilde{w})^2}\left[(1+\alpha\tilde{w})\,2\,(1-2\alpha)\left(1+2\sqrt{\alpha-\alpha^2}\right)\right. \\ &\left. -\tilde{w}\sqrt{\alpha-\alpha^2}\left(\left(1+2\sqrt{\alpha-\alpha^2}\right)^2-16\alpha^2\right)-2K\,(1-2\alpha)\,(1+\alpha\tilde{w})^2\right] \\ \frac{\partial W}{\partial \tilde{w}} &= \frac{\tilde{w}}{(1+\alpha\tilde{w})^2}\left[16\alpha^3\tilde{w}+\left(1+2\sqrt{\alpha-\alpha^2}\right)^2(2+\alpha\tilde{w})\right. \\ &\left. -4K\left(1+2\sqrt{\alpha-\alpha^2}\right)(1+\alpha\tilde{w})^2\right].\end{aligned}$$

If there exists an $\alpha \in \left(0, \frac{1}{2}\right)$ that is a candidate to solution of the optimization program then $\alpha^*(x)$ and $\tilde{w}^*(x)$) are such that:

$$\begin{gathered}(1+\alpha\tilde{w})\,2\,(1-2\alpha)\left(1+2\sqrt{\alpha-\alpha^2}\right)-\tilde{w}\sqrt{\alpha-\alpha^2}\left(\left(1+2\sqrt{\alpha-\alpha^2}\right)^2-16\alpha^2\right) \\ -2K\,(1-2\alpha)\,(1+\alpha\tilde{w})^2=0\end{gathered} \tag{17}$$

$$16\alpha^3\tilde{w}+\left(1+2\sqrt{\alpha-\alpha^2}\right)^2(2+\alpha\tilde{w})-4K\left(1+2\sqrt{\alpha-\alpha^2}\right)(1+\alpha\tilde{w})^2=0 \tag{18}$$

It is straightforward to see that $\alpha = \frac{1}{2}$ is always a solution to the first order condition $\frac{\partial W}{\partial \alpha} = 0$. It can also be seen that there does not always exist an interior solution. In particular, for $K = 0$ equation (18) does not hold as it is always positive. Analogously, it can be shown numerically that for $K = 1$ equation (18) does not hold as it is always negative. Thus, by continuity, we can ensure that for very high ($K \to 1$) and very low ($K \to 0$) values of K there is no interior solution for both variables. For such extreme values of K the solution will be at one of the corners. In addition, even if there exists a solution for $\alpha \in \left[0, \frac{1}{2}\right]$, since the system formed by (17) and (18) does not depend on A, for low levels of A, the pair $(\alpha^*(K),\ \tilde{w}^*(K))$ satisfying the system may not fulfill condition $\tilde{w} \leq \frac{A}{\left(1+2\sqrt{\alpha-\alpha^2}\right)}$. Again, in these cases the solution will be one at the corners of the domain.

From the previous discussion about the candidates to solution, it is easy to conclude that:

- For $A \leq \min\left\{\frac{4(1-K)}{K},\ \tilde{w}^*(K)\left(1+2\sqrt{\alpha^*(K)-\alpha^*(K)^2}\right)\right\}$, the only candidate to solution is candidate (ii). Hence the solution will be of this type.

- For $A \geq \tilde{w}^*(K)\left(1+2\sqrt{\alpha^*(K)-\alpha^*(K)^2}\right)$ the interior solution (v) is also a candidate. The complexity of the system prevents us from fully characterizing the interior candidate, but we can easily show that there exist intermediate values of K for which the system has a solution satisfying the constraints and it is the solution of the HA problem. To prove existence it suffices to take, for instance $K = 0.6$. For this particular value the system formed by (17) and (18) yields $\alpha^*(K = 0.6) \simeq 0.1276$

and $\tilde{w}^*(K = 0.6) \simeq 2.0348$. Note also that for $\alpha^*(K = 0.6) \simeq 0.1276$ and $\tilde{w}^*(K = 0.6) \simeq 2.0348$ to satisfy $\tilde{w} \leq \frac{A}{1+2\sqrt{\alpha-\alpha^2}}$, it has to be the case that $A \geq 3.3926$. It can be easily proven that an increase in δ will translate into a lower $\bar{\gamma}$. To show this point, note that K is the only parameter that affects α^* and $\tilde{w}^*$. This allows us to show that for the interior solution an increase in δ (which does not affect K) will not affect the solution of the problem. This, in turns, implies that w will decrease (to keep $\tilde{w}^*$ invariant) and hence $\bar{\gamma}$ will decrease (to keep α^* invariant).

To complete the proof it rests to compare the objective function evaluated at the different possible solutions. To have a better understanding of the solution it is useful to use the comparison between candidate (iii), $\alpha = 0$, and candidate (iv), $\alpha = \frac{1}{2}$:

- *If* $K \in \left(\frac{1}{2}, 1\right)$ then $W^{\alpha=1/2} > W^{\alpha=0}$. In addition, we know that for K close to one the interior candidate (v) does not exist. Hence for large values of K the solution to the problem is either of type (iv) or of type (ii).
- If $K \leq \frac{1}{2}$ and $A > \frac{4(1-K)}{K}$ then, whether $W^{\alpha=0} = A^2(1-2K)$ is higher or lower than $W^{\alpha=1/2} = 32\left(\frac{A}{4} - \ln\left(1+\frac{A}{4}\right)\right) - K(A)^2$ depends on A. Let us denote by $\tilde{A}(K)$ the level of A that satisfies

  $$\tilde{A}^2(1-K) = 32\left(\frac{\tilde{A}}{4} - \ln\left(1+\frac{\tilde{A}}{4}\right)\right)$$

  Note that $\tilde{A}(K)$ is an increasing function of K, with $\tilde{A}(K = \frac{1}{2}) = 6.4951$. Then, for $A < \tilde{A}(K)$, it holds that $W^{\alpha=1/2} > W^{\alpha=0}$, while for $A > \tilde{A}(K)$, we have that $W^{\alpha=1/2} < W^{\alpha=0}$. The best of them has to be compared with the other candidates to solution. As for K close to 0 there is no interior solution of type (v), if K is very small and A is large then candidate (iii) has to be compared with candidate (ii).
- If $K \leq \frac{1}{2}$ and $A \leq \frac{4(1-K)}{K}$ (a decreasing function of K), then (iii) and (iv) are particular cases of (ii) and are taken into account there. Since for K close to 0 there is no solution of type (iv), we have that for low values of K and A the solution of the problem is of type (ii).

  Final comparisons for each combination of parameters can be done numerically.

To show that it is possible to find intermediate values of K for which an interior candidate (v) is optimal it suffices to consider again $K = 0.6$. A direct computation shows that, when comparing the welfare at this value with the one obtained in candidates (ii) or (iv), $W^{\alpha}\left(\alpha^*(K = 0.6), \tilde{w}^*(K = 0.6)\right)$ is the highest one.

A.11 Proof of Proposition 8.

Developing countries: The proof follows from combining Propositions 2, 4 and 7.
Developed countries: To proof that for $\delta \leq \frac{2(1+\lambda)}{\varphi}$ the best is laissez-faire and that for $\delta \in \left[\frac{2(1+\lambda)}{\varphi}, \frac{4(1+\lambda)}{\varphi}\right]$ and $\delta > \frac{8(1+\lambda)}{\varphi}$ the best policy is to impose a limit $\bar{\gamma} > 0$ on the physician involvement in dual practice is direct from combining Propositions 1, 3 and 6.

Thus, it remains to show that for values of $\delta \in \left[\frac{4(1+\lambda)}{\varphi}, \frac{8(1+\lambda)}{\varphi}\right]$ a policy of limits dominates that with exclusive contracts.

From Proposition 3 we know that for $\delta \in \left[\frac{4(1+\lambda)}{\varphi}, \frac{8(1+\lambda)}{\varphi}\right]$ the optimal exclusive contract is given by $\Delta = \frac{\varphi\delta - 4(1+\lambda)}{2\delta(1+\lambda)}$ and $w^E = \frac{\sqrt{1+\frac{\delta\varphi}{1+\lambda}}-1}{\delta}$. Defining $x \equiv \frac{\delta\varphi}{1+\lambda}$ the associated HA's welfare is:

$$\begin{aligned}\mathcal{W}^{E*} &\equiv \mathcal{W}^{E}\left(\Delta = \frac{x-4}{2\delta},\ w^E = \frac{\sqrt{1+x}-1}{\delta}\right) = \\ &\frac{1+\lambda}{\delta^2 k^2}\left(\frac{x^2}{2} - 2x + 4\sqrt{x+1} - 12 + 4x\left(\ln\left(2\left(1+\sqrt{1+x}\right)\right) - \ln(x)\right)\right)\end{aligned}$$

Note that, as we are in the region with $\delta \in \left[\frac{4(1+\lambda)}{\varphi}, \frac{8(1+\lambda)}{\varphi}\right]$, this function is only defined for $x \in [4, 8]$.

We compare now $\mathcal{W}^{E*}$ with the welfare obtained under the optimal policy for $x \in [4, 8]$, as characterized in Proposition 6, given by $\mathcal{W}^{\alpha}(\alpha^*(x), \tilde{w}^*(x))$ with $\alpha^*(x)$ and $\tilde{w}^*(x)$ being the solution to the system (15) and (16). For this purpose, let us define, for any value of x, $\breve{\mathcal{W}} \equiv \mathcal{W}^{E*} - \mathcal{W}^{\alpha}(\alpha^*(x), \tilde{w}^*(x))$.

It can be shown numerically that $\forall x \in [4, 8]$, $\breve{\mathcal{W}}$ is decreasing in x. Moreover, we find that when $x = 4$, $\breve{\mathcal{W}} < 0$. Therefore, for any $x \in [4, 8]$, i.e., $\delta \in \left[\frac{4(1+\lambda)}{\varphi}, \frac{8(1+\lambda)}{\varphi}\right]$ it holds that $\mathcal{W}^{\alpha}(\alpha^*(x), \tilde{w}^*(x)) > \mathcal{W}^{E*}$.

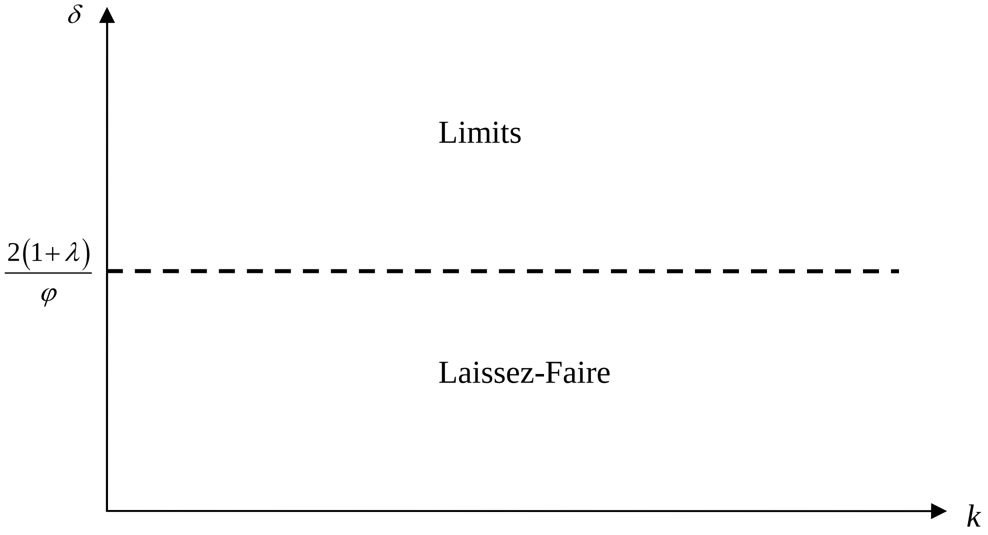

Figure 1: Optimal Regulation for Developed Countries

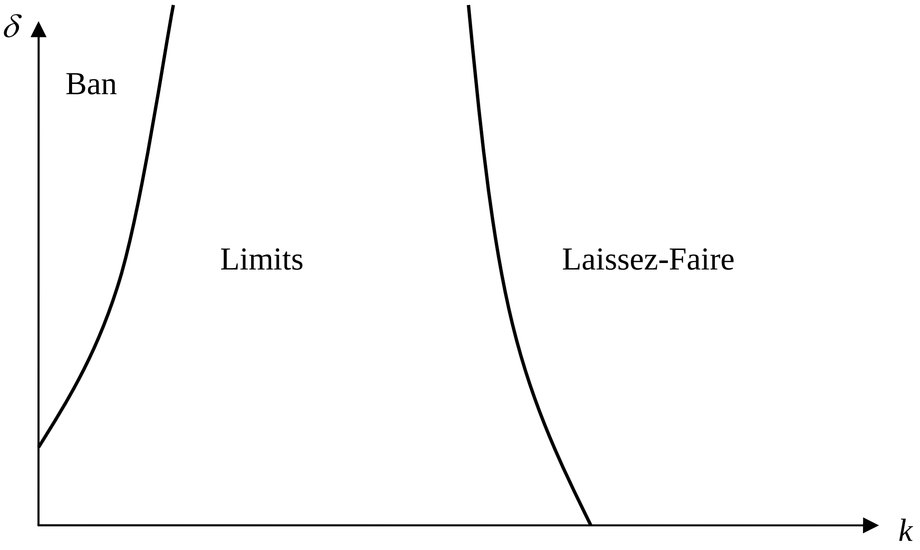

Figure 2: Optimal Regulation for Developing Countries for $\delta > 0$