Working papers series

WP ECON 11.01

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

Paula González (U. Pablo de Olavide)
Inés Macho-Stadler (U. Autónoma de Barcelona)

JEL Classification numbers: I11, I18, L51, H51.
Keywords: Dual practice, optimal contracts, physicians' incentives, regulations.

Department of Economics

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

Paula González† and Inés Macho-Stadler‡

February 11, 2011

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 David Pérez-Castrillo, Nicolás Porteiro and seminar participants at 7th European Conference on Health Economics, 30th Jornadas de la Asociación Española de Economía de la Salud and Universitat de Barcelona. 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. Email: Ines.Macho@aub.es

1 Introduction

In many developed and developing countries it is common practice for physicians to work simultaneously in public hospitals and private facilities. 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, negatively impacting service provision in the public sector.[1] On the whole, there is no consensus on the net effects of dual practice in the health sector and there is no unique and simple answer as to whether and how this practice should be regulated.

This lack of consensus is reflected by the fact that there is wide cross-country heterogeneity in government responses to dual practice.[2] While some governments ban it altogether,[3] others regulate or restrict dual practice with different regulatory instruments. The measures implemented include offering higher salaries or other work benefits to physicians in exchange for their working exclusively in the public sector,[4] limiting the income physicians can earn through dual practice,[5] and limiting dual practice through government specification of the maximum involvement in private activities.[6] In addition, most of these regulations have been implemented only in developed countries, while in devel-

[1] See Eggleston and Bir (2006) for a thorough discussion on these issues.

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

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

[4] 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.

[5] 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).

[6] 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).

oping 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 the answers to 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.

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 ban dual practice (if the cost associated with dual practice is high) or leave it alone (if the cost is low). Limiting policies in developing countries emerge as the optimal instrument only for 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.[7] There has been some research on physicians' incentives as dual providers.

[7]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 Martinez-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

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 incentive 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,

(2005), which analyzes the desirability of mixed health care systems when distributional aspects matter.

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

We consider a Health Authority (HA hereafter) that aims to provide public health care but is also concerned about its costs. In order to keep the set-up tractable we abstract from patients and concentrate on the amount of health generated in the public sector. 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.

2.1 The physician's decision

There is a set of physicians with different ability a distributed uniformly on the interval $[0, \bar{a}]$.[8] The total amount of physicians has mass $\bar{a}$. Physicians can work solely in the public sector, work for the private sector or or work in both sectors as dual providers. If they work for the public sector they receive the wage w. In addition to the wage from the public sector w, a physician who is involved in dual practice receives profits from this practice. These profits are equal to a revenue ($\Pi^D(\gamma, a)$) that depends on the physician ability, a, and on the amount of dual practice he performs, measured by $\gamma \geq 0$. We assume $\Pi^D(\gamma, a)$ is increasing in a and γ and concave in γ.

When involved in dual practice, and in cases where the HA does not impose any restriction, the physician chooses the intensity of his dual practice γ in order to maximize his profits. If we denote by $\gamma^*(a)$ the optimal involvement in dual practice then $\Pi^D(a) \equiv \Pi^D(\gamma^*(a), a)$.

Finally, the physician can choose to practice solely in the private sector. In this case he receives the revenue $\Pi^{Pv}(a) \geq 0$. We assume that $\rho(a) \equiv \Pi^{Pv}(a) - \Pi^D(\gamma, a) > 0$, i.e., the amount of private profits earned by a physician who is a dual practitioner is always strictly smaller than that attained by leaving the public sector altogether.

[8] 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.

Formally, the physician's utility, depending on the type of practice, is

$$\begin{aligned} U^{pub} &= w \\ U^{D} &= w+\Pi^{D}(\gamma,a) \\ U^{Pv} &= \Pi^{Pv}(a) \end{aligned}$$

Now we can study the physician's decision as a function of his ability and the wage offered in the public sector. In what follows, we will assume particular functional forms but our first result can be easily stated in general:

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

$$\begin{array}{ll} \text{if } a\in\left[0,\tilde{a}^{D}(w)\right] & \text{he chooses dual practice} \\ \text{if } a\in\left[\tilde{a}^{D}(w),\bar{a}\right] & \text{he chooses to work only in the private sector,} \end{array}$$

with $\tilde{a}^{D}=\rho^{-1}(w)$.

b) If dual practice is not allowed, and

$$\begin{array}{ll} \text{if } a\in\left[0,\tilde{a}^{Pv}(w)\right] & \text{he chooses public practice} \\ \text{if } a\in\left[\tilde{a}^{Pv}(w),\bar{a}\right] & \text{he chooses to work only in the private sector,} \end{array}$$

with $\tilde{a}^{Pv}=\left(\Pi^{Pv}\right)^{-1}(w)$

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. Note that when dual practice is allowed no physician decides to work solely in the public sector.[9] When dual practice is forbidden, the population of physicians working for the public sector decreases (since $\Pi^{Pv}(a)>\rho(a)$ it is straightforward that, for a given w, $\tilde{a}^{Pv}(w)<\tilde{a}^{D}(w)$). 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.[10]

[9] This is due to the fact that we do not consider a fixed cost for engaging in dual practice, and this assumption makes it profitable for all physicians to be moonlighters. This assumption can be relaxed without altering the message of the paper, but at the expense of sacrificing expositional clarity.

[10] We assume the upper bound $\bar{a}$ is sufficiently large so as to avoid corner solutions (situations in which no physician decides to work solely in the private sector).

Since we will introduce the physician decision in a more complex game and we want to compare different regulatory regimes, in order to derive explicit results we will henceforth consider that the profits from dual and private practice are given by the functional forms $\Pi^D(\gamma, a) = (k\gamma a - \gamma^2)^{1/2}$ and $\Pi^{Pv}(a) = ka$ respectively. The parameter $k > 0$ serves as a proxy for the attractiveness of the private sector. This parameter may be speciality-specific, which allows us to discuss the different behaviours of physicians engaged in primary, secondary, and tertiary care, and in different specialities. Parameter k can also be seen as a measure of the physician's need for extra revenue or as the financial motivation of the physician.

These functional forms satisfy all the hypotheses we have mentioned above regarding dual and private practice benefits, and are simple but flexible enough to have interesting results. For these functional forms, given ability a, the optimal physician involvement in dual practice $\gamma^*(a)$ is

$$\gamma^*(a) = \tfrac{ka}{2} \quad \text{for any } a \in [0, \bar{a}],$$

which implies that at the optimal level of involvement in dual practice the physician has profits $\Pi^D(a) = \frac{ka}{2}$.[11] These profits illustrate the fact that since more able physicians can obtain higher income by working in the private sector, they will be more devoted to dual practice and will succeed in obtaining a higher income as a result (not only because they are more involved in private activity but also because the market values them more). The advantage of using specific functional forms, especially the ones we are considering, is that we can obtain explicit solutions for the thresholds defined in Lemma 1, which are

$$\tilde{a}^D = \frac{2w}{k} \qquad \text{and} \qquad \tilde{a}^{Pv} = \frac{w}{k}$$

both increasing in the wage received through public practice and decreasing in the attractiveness of the private sector. More able physicians tend to be more involved, or only involved, in private practice. Thresholds can also be read in terms of physicians with the same ability but different parameter k. If we follow our interpretation of k as speciality-specific, the properties of these thresholds are in accordance with some stylized facts since more doctors will be involved in the private sector as k 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 higher financial motivations will be

[11] This implies that $\rho(a) = k\frac{a}{2}$.

more prone to dual practice either because they suffer from financial constraints or because public sector salaries are low. This is 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.[12]

2.2 The Health Authority decision

To define the HA'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. 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 that will be increasing in the amount of private practice he performs (γ). 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 dual practice on public sector performance. Note that this functional form allows for several situations. 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. If $\delta = 0$ public and private activities are independent. 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.[13] This way, the cost of dual practice is increasing in γ (the marginal impact on performance increases as the physician becomes more involved in dual practice) and convex.

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:

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

[12]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.

[13]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.

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 $\lambda > 0$ represents the relative weight of the health provision as compared to the costs concern. It is easy to see that λ and $F(a)$ always go together. We have chosen to keep both variables in the model so as to discuss more easily cases where the HA has a higher concern about health care provision and cases where the HA has access to more productive health care technology.

The HA decides on the wage w, which indirectly determines the physician's decision to allocate services. 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. 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 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 $F(a) = fa$.[14]

Now we have all the tools to study the impact of different policy options to regulate dual practice. We observe wide variations in how governments tackle the issue of dual practice. While some governments fully prohibit this practice, others regulate or restrict dual job holding with different intensities and regulatory instruments. In the following

[14]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.

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. If this is the only intervention available, the alternatives of the HA are either to let physicians freely decide whether and how to be dual providers or to forbid dual practice and let physicians choose only between public or private provision.

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

$$\max_{w} \mathcal{W}^{LF} = \lambda \int_0^{\frac{2w}{k}} \frac{F(a)}{1+\delta \frac{ka}{2}} da - w\frac{2w}{k}.$$

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

$$\max_{w} \mathcal{W}^{B} = \lambda \int_0^{\frac{w}{k}} F(a)\, da - w\frac{w}{k}.$$

We focus first on developed countries. In this case, after computing the optimal wages in both regimes and comparing their associated HA's welfare, we conclude:

Proposition 1 *In* ***developed economies,*** *if the HA can only ban dual practice, there exists a $\bar{\delta}_1 \in \left(\frac{2}{\lambda\varphi}, \frac{4}{\lambda\varphi}\right)$ 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*

$$w^{LF} = \frac{\sqrt{1+2\delta\lambda\varphi}-1}{2\delta}.$$

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

$$w^{B} = \frac{\lambda\varphi}{2}.$$

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

$$a^{LF} \equiv \tilde{a}^{D}(w^{LF}) = \frac{\sqrt{1+2\delta\lambda\varphi}-1}{\delta k} \qquad \text{and} \qquad a^{B} \equiv \tilde{a}^{Pv}(w^{B}) = \frac{\lambda\varphi}{2k}.$$

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 (Holmstrom 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.

The complete analysis of the comparative statics of the results in Proposition 1 is presented in Table 1, which summarizes the sign of the derivatives of (w^{LF}, a^{LF}) and (w^{B}, a^{B}) with respect to the parameters λ, δ, k and φ.

	λ	δ	k	φ
w^{LF}	$+$	$-$	0	$+$
a^{LF}	$+$	$-$	$-$	$+$
	λ	δ	k	φ
w^{B}	$+$	0	0	$+$
a^{B}	$+$	0	$-$	$+$

Table 1: Comparative statics $F(a) = \varphi$

As one might expect, if the HA puts increased weight on public health provision (higher 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. It is also interesting to note that k has no effect on the salary paid in the public sector; it only affects the size of the population of physicians attracted to the public health system. The fact that k does not affect the salary in developed countries, where the public health production function is $F(a) = \varphi$, is related to the fact that in this scenario the marginal revenue required to keep one more doctor in the public health system and the marginal cost of doing so are both linear in $\frac{1}{k}$, and hence k does not affect the optimal wage. Thus, changes in k only affect the number of physicians that are hired.

Note that the results presented in Proposition 1 are also valid (and well-defined) for negative values of δ (as long as $-\delta < \frac{1}{2\lambda\varphi}$). For $\delta \leq 0$, the laissez-faire regime is always superior.[15] Since this superiority result is maintained throughout the paper, we will not

[15] For the particular case $\delta = 0$, the HA is not concerned about dual practice and it will set the same

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, the best intervention is as follows,*

i) *If $k > \lambda f$, then the public sector is unsustainable for every value of δ.*

ii) *If $k \in \left(\frac{\lambda f}{2}, \lambda f\right]$, not to regulate dual practice (for any level of δ) and set a wage level*

$$w^{LF} = \frac{\lambda f - k}{\delta k}.$$

iii) *If $k \leq \frac{\lambda f}{2}$, then there exists a threshold $\bar{\delta}_2 > 0$ such that,*

a) *If $\delta \leq \bar{\delta}_2$, not to regulate dual practice and set a wage level*

$$w^{LF} = \frac{\lambda f - k}{\delta k}.$$

b) *If $\delta > \bar{\delta}_2$, ban dual practice and set a wage level*

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

The results for $k \leq \frac{\lambda f}{2}$ imply (using Lemma 1) that the cut-offs when dual practice is allowed and forbidden are respectively

$$a^{LF} \equiv \tilde{a}^D(w^{LF}) = \frac{2(\lambda f - k)}{\delta k^2} \qquad \text{and} \qquad a^B \equiv \tilde{a}^{Pv}(w^B) = \bar{a}.$$

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 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.[16] 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

wage, $w = \frac{\lambda\varphi}{2}$ in both regimes. Since by allowing dual practice the HA is able to attract more doctors, and hence to provide more health, regulation will never be in the HA's interest.

[16] 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.

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 (λf).[17] 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 of the comparative statics for the production technology defined by $F(a) = fa$, yields the effects summarized in Table 2. Again, as the weight placed by the HA on public health provision increases (higher 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.[18]

	λ	δ	k	f
w^{LF}	$+$	$-$	$-$	$+$
a^{LF}	$+$	$-$	$-$	$+$
	λ	δ	k	f
w^{B}	0	0	$+$	0
a^{B}	0	0	0	0

Table 2: Comparative statics $F(a) = fa$

As we have already observed, the attractiveness of the private sector (k) reduces the number of physicians involved in public provision. Also, a more attractive private sector reduces the wage paid in the public sector if dual practice is allowed; wages increase only if dual practice is banned (i.e., when the public sector decides to hire all physicians).

4 Rewarding Policies

Let us now consider the policy of paying (on top of a salary w^E) a premium Δ to physicians who decide to work exclusively for the public sector.[19] 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.

[17] In what follows, we disregard the case 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). Therefore, we assume $k < \lambda f$.

[18] 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.

[19] This bonus can also be interpreted in terms of better career opportunities.

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

$$\begin{aligned} U^{pub} &= w^E + \Delta \\ U^D &= w^E + \frac{ka}{2} \\ U^{Pv} &= ka \end{aligned}$$

Now we can study the physician's decision as a function of his ability and the contracts offered in the public sector. Without loss of generality, we can restrict our attention to situations where $\Delta \geq 0$ because if the premium is zero then we have the case in which no doctor is working exclusively 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*

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

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

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

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

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 $(\Delta > w^E)$.[20] 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, more physicians tend to be involved in dual practice.

We wish to highlight that this regulatory environment is sufficiently rich so as to encompass the laissez-faire and banning regimes examined in the previous section, as the following remark shows.

[20]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.

Remark 1 *For any optimal contract w^{LF} or w^B it is possible to find a duple $\left(w^E, \Delta\right)$ that generates the same outcome.*

Remark 1 ensures that any outcome that could be achieved in the previous section can be replicated within this richer context. 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 $\delta \leq \frac{2}{\lambda\varphi}$ not to regulate dual practice, fix $\Delta = 0$, and set a wage level*

$$w^E = w^{LF} = \frac{\sqrt{1 + 2\delta\lambda\varphi} - 1}{2\delta}$$

ii) *If $\delta \in \left(\frac{2}{\lambda\varphi}, \frac{4}{\lambda\varphi}\right]$ to offer an exclusive contract with*

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

iii) *If $\delta > \frac{4}{\lambda\varphi}$ to ban dual practice, fix $w^E = 0$, and set an exclusivity premium*

$$\Delta = w^B = \frac{\lambda\varphi}{2}$$

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 see how exclusive contracts offer greater flexibility for the HA to mitigate the loss of productivity associated with dual practice. Using Proposition 1 we can conclude that in developed economies 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 ($\bar{\delta}_1$) than when exclusive contracts are available ($\frac{4}{\lambda\varphi}$). This makes the HA less interested in banning dual practice when such rewarding policies are available.

However, as we now detail, the results for developing countries contrasts 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.

First we characterize physician behaviour when the option to engage in dual practice is subject to limitation. We consider the two possible limitations, one after the other. Focusing on involvement constraints we find:

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

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

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

- *If $\bar{\gamma} \geq w$, the limiting policy is ineffective and*

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

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

$$\begin{array}{ll} \text{if } a \in \left[0, \frac{2\bar{\gamma}}{k}\right] & \text{the physician chooses dual practice and } \gamma^*(a) = \frac{ka}{2} \\ \text{if } a \in \left[\frac{2\bar{\gamma}}{k}, \frac{2w+\bar{\gamma}+\sqrt{\bar{\gamma}(4w-3\bar{\gamma})}}{2k}\right] & \text{the physician chooses dual practice and } \gamma^*(a) = \bar{\gamma} \\ \text{if } a \in \left[\frac{2w+\bar{\gamma}+\sqrt{\bar{\gamma}(4w-3\bar{\gamma})}}{2k}, \bar{a}\right] & \text{the physician chooses to work only in the private sector.} \end{array}$$

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 $(k\gamma a - \gamma^2)^{1/2} \leq \bar{\Pi}^D$.

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

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

Accordingly, with this, 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*

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

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

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

When earnings limitations are effective, dual practice depends (in a negative way) on the physician ability a and on the profitability parameter k: $\hat{\gamma}(a, \bar{\pi}) = \frac{ak - \sqrt{a^2k^2 - 4\bar{\pi}^2}}{2}$. 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 $a > \frac{2\bar{\Pi}^D}{k}$, 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.

The proof in the Appendix shows that for any possible policy $\left(\bar{\Pi}^D, w\right)$ it is possible to construct a policy of the form $(\bar{\gamma}, w)$ that incurs the same costs (i.e., pays the same wages and hires the same amount of physicians) while inducing a lower amount of dual practice (thus reducing losses of productivity). 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.[21]

Once we have shown that a policy that limits involvement in private practice is always preferable to one that limits physician earnings, 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 ($\bar{\gamma} \geq w$), 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. In this case, we are trivially back to the laissez-faire scenario.

The more interesting case is when $\bar{\gamma} < w$, and the policy actually affects physician behaviour. In this case, the HA solves

$$\max_{w,\bar{\gamma}} \mathcal{W}^{\bar{\gamma}} = \lambda \left(\int_0^{\frac{2\bar{\gamma}}{k}} \frac{F(a)}{1+\frac{\delta k a}{2}} da + \int_{\frac{2\bar{\gamma}}{k}}^{\frac{2w+\bar{\gamma}+\sqrt{\bar{\gamma}(4w-3\bar{\gamma})}}{2k}} \frac{F(a)}{1+\delta\bar{\gamma}} da \right) - w \left(\frac{2w+\bar{\gamma}+\sqrt{\bar{\gamma}\left(4w-3\bar{\gamma}\right)}}{2k} \right),$$

subject to the constraints that $w \geq 0$, and $0 \leq \bar{\gamma} \leq w$.

As before, the characteristics of the health care system determine the results. Let us first consider a developed economy,

[21] In fact this result can be extended to a more general model without the need to resort to particular functional forms. The details are available from the authors upon request.

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 $\delta < \frac{2}{\lambda\varphi}$ the best the HA can do is not to limit dual practice.*
- *If $\delta > \frac{2}{\lambda\varphi}$ there exists an optimal limit to the amount of dual practice.*

Two main insights emerge from this proposition. 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 ($\delta < \frac{2}{\lambda\varphi}$) 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.*
- *There exist intermediate values of k for which 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 ban dual practice (if δ is high) or not to limit dual practice at all (if δ is low).*

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 such professionals, due to aforementioned characteristics of health care provision in developing countries, would severely undermine the HA's capacity to provide health. When the private sector is relatively unattractive a limiting policy is also not optimal, for the opposite reason: in this case, because it is relatively cheap to retain physicians at public facilities, when the cost of dual practice is large enough the HA is better off banning rather than limiting dual practice. Thus banning dual practice can emerge in developing economies as the best intervention. Limits are optimal in developing countries only when the attractiveness of the private sector is moderate. 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.

As shown in the proof of Proposition 7, as the value of δ increases, the limit imposed by the HA on dual practice will be more stringent in order to mitigate its negative consequences.

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 that can be extracted from the analysis.

6.1 The Health Authority's Choice

First we provide an overview of the different policy options for both developing and developed economies. Combining the propositions discussed in previous sections we obtain the following result:

Proposition 8 *The optimal decision of the HA is*

- *In **developed economies**,*
  - *If $\delta \leq \frac{2}{\lambda\varphi}$ not to regulate dual practice.*
  - *If $\delta > \frac{2}{\lambda\varphi}$ 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 Laissez-Faire and Banning) also dominates exclusive contracts. Thus in developed countries it follows that 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, $\delta \in \left(\frac{2}{\lambda\varphi}, \frac{4}{\lambda\varphi}\right]$, 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. Hence the optimal policy coincides with the results stated in Proposition 7. Accordingly, for these countries, policy recommendations for dealing with dual practice are quite different. In developing economies, 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 (i.e., sufficiently low values of k) can result in the optimality of banning dual practice altogether. The optimality of a limiting policy is confined to intermediate values of k, i.e., not so low as to make banning affordable, and not so high as to trigger a brain-drain.

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 wage 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.

Turning to developed countries, we find that in our model k affects neither the choice of optimal regulation nor the kind of contract offered, though it does affect the overall number of physicians working within the public sector. In the case of 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 analysis 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.[22] If enforceability is an issue, then the design of the optimal policy to regulate dual practice should incorporate this enforcement dimension.

Although we have not included enforceability concerns in our analysis, we can make a few observations regarding this important issue. First, we speculate that it may be easier to control earnings than involvement. Perhaps this is why some countries seem

[22] With regard to this issue, Eggleston and Bir (2006) argue that the social trade-off between the benefits and costs of dual practice crucially depend on the quality of a country's contracting institutions.

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 (Austria, Ireland, Italy, etc.), 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).

Finally, it is worth reiterating that our model shows that in developing countries, even in a best-case scenario where enforcement is not an issue, limiting policies are optimal only for very few parameter configurations (k intermediate). If enforceability problems are introduced, the attractiveness of these policies would be reduced even more.

Budget Constraints

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 parameter λ measures the importance of public health provision into the government function. Note that there might be non-essential specialities (such as dermatologists, speech therapists or dieticians, among others) for which the HA may assign a low value of λ. Our model predicts that in developing countries if this value is sufficiently low (in particular, $\lambda < \frac{k}{f}$), these specialities should not be included in the coverage of public plans.

The value of λ can also be interpreted as the relative importance that revenue has as compared to wage costs or, in other words, as a budgetary concern. When considered in this light, we expect that during an economic recession λ should be lower due to more stringent budget constraints. Here our model provides an argument in favour of non-regulation in both 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.

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. In addition, many health care systems around the world have adopted some form of dual practice regulation. However, the real cost of dual practice for health systems is an empirical issue. There are no reliable studies that estimate this cost, which is summarized by δ in our model. 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).[23] 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.

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

[23] Formally, this amounts to assuming $\varphi > \frac{f}{2}\bar{a}$

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 ban dual practice (if the associated costs are high) or not to intervene (if such costs are low). 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.

Certainly, more theoretical and empirical work in this line of research is needed. Still, we believe that this work can inform the discussion of dual practice and contribute to the development of a better policy making process.

References

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

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

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

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

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

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

[7] 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.

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

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

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

[11] 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.

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

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

[14] 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.

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

[16] 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.

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

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

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

[20] 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.

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

[22] 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.

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

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

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

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

A Appendix

A.1 Proof of Lemma 1.

From the physician's utility under the different choices $U^{pub} = w$; $U^{D} = w + \Pi^{D}(\gamma, a)$; $U^{Pv} = \Pi^{Pv}(a)$, it is easy to conclude that no one decides to work only for the public sector if dual practice is allowed, and the physician ability that determines to go exclusively to private practice is the one presented in the lemma. ■

A.2 Proof of Proposition 1.

The HA has to solve two independent optimization problems. First, if there is no regulation, the HA solves

$$\max_{w} \mathcal{W}^{LF} = \lambda \int_{0}^{\frac{2w}{k}} \frac{\varphi}{1 + \delta \frac{ka}{2}} da - w \frac{2w}{k},$$

subject to the constraint that $w \geq 0$. Solving the f.o.c. we obtain the following candidate to optimum

$$w^{LF} = \frac{\sqrt{1 + 2\delta\lambda\varphi} - 1}{2\delta} > 0. \tag{1}$$

Moreover, it can be checked that the objective function is concave and, hence, that the s.o.c. is fulfilled. Evaluating the objective function in the optimal level of wage we have

$$\mathcal{W}^{LF}\left(w^{LF}\right) = \lambda \int_{0}^{\frac{1}{k\delta}\left(\sqrt{2\lambda\varphi\delta+1}-1\right)} \frac{\varphi}{\frac{1}{2}ak\delta + 1} da - \frac{1}{2k\delta^{2}} \left(\sqrt{2\varphi\lambda\delta + 1} - 1\right)^{2}. \tag{2}$$

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

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

If the HA decides to ban dual practice, the optimization program it solves is,

$$\max_{w} \mathcal{W}^{B} = \lambda \int_{0}^{\frac{w}{k}} \varphi da - w \frac{w}{k},$$

subject to the constraint that $w \geq 0$. Solving the f.o.c. we obtain the following candidate to optimum

$$w^{B} = \frac{\lambda\varphi}{2} > 0. \tag{3}$$

Moreover, it can be checked that the objective function is concave and, hence, that the s.o.c. is fulfilled. Evaluating the objective function in the optimal level of wage we have

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

that does not depend on δ. 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 $\delta < \bar{\delta}_1$, $\mathcal{W}^{LF}\left(w^{LF}\right) > \mathcal{W}^B\left(w^B\right)$ 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 $\bar{\delta}_1 \in \left(\frac{2}{\lambda\varphi}, \frac{4}{\lambda\varphi}\right)$. Let us write $\delta = \frac{x}{\lambda\varphi}$. Then

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

that we have to compare with

$$\mathcal{W}^B = \frac{\lambda^2\varphi^2}{4k}$$

Hence,

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

For $x = 4$ we have

$$\frac{2}{4}\left(\ln\frac{\sqrt{9}+1}{2}\right) - \frac{1}{2\,(4)^2}\left(\sqrt{9}-1\right)^2 = 0.221\,57 < \frac{1}{4}$$

and for $x = 2$ we have

$$\frac{2}{2}\left(\ln\frac{(\sqrt{5}+1)}{2}\right) - \frac{1}{2\,(2)^2}\left(\sqrt{5}-1\right)^2 = 0.290\,23 > \frac{1}{4}$$

For completeness, we can show that $\bar{\delta}_1$ is close to $\frac{3}{\lambda\varphi}$ since for $x = 3$

$$\frac{2}{3}\left(\ln\frac{\sqrt{6+1}+1}{2}\right) - \frac{1}{2\,(3)^2}\left(\sqrt{6+1}-1\right)^2 = 0.249\,8 < \frac{1}{4}$$

This completes the proof. $\blacksquare$

A.3 Proof of Proposition 2.

The HA has to solve two independent optimization problems. First, if there is no regulation, the HA determines w^{LF} by solving

$$\max_{w} \mathcal{W}^{LF} = \lambda \int_0^{\frac{2w}{k}} \frac{fa}{1+\delta\frac{ka}{2}} da - w\frac{2w}{k},$$

subject to the constraint that $w \geq 0$. Solving the f.o.c. we obtain the following candidate to optimum

$$w^{LF} = \frac{\lambda f - k}{\delta k}. \tag{5}$$

This wage level is positive only if $k < \lambda f$. Therefore, the candidate to solution is:

$$w^{LF} = \begin{cases} \frac{\lambda f - k}{\delta k} & \text{if } k < \lambda f \\ 0 & \text{if } k \geq \lambda f \end{cases}$$

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 laissez-faire contract is:

$$\mathcal{W}^{LF}\left(w^{LF}\right) = \lambda \int_0^{\frac{2(\lambda f - k)}{\delta k^2}} \frac{fa}{1+\delta\frac{ka}{2}} da - \frac{2}{k}\left(\frac{\lambda f - k}{\delta k}\right)^2 \tag{6}$$

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

$$\begin{aligned} \lim_{\delta\to 0} \mathcal{W}^{LF}\left(w^{LF}\right) &= +\infty \\ \lim_{\delta\to +\infty} \mathcal{W}^{LF}\left(w^{LF}\right) &= 0 \end{aligned}$$

If the HA decides to ban dual practice, the optimization program it solves is,

$$\max_{w} \mathcal{W}^{B} = \lambda \int_0^{\frac{w}{k}} fada - w\frac{w}{k},$$

subject to the constraint that $w \geq 0$. This objective function is monotone in w. Hence, the solution is always on the boundaries of the support. Either $w^B = 0$, or w^B is such that $\tilde{a}^{Pv} = \bar{a}$, ($w^N = \bar{a}k$). Evaluating the objective function in these two candidates we have

$$\begin{gathered} w^B = 0 \Longrightarrow \mathcal{W}^B = 0 \\ w^B = \bar{a}k \Longrightarrow \mathcal{W}^B = \bar{a}^2\left(\frac{\lambda f}{2} - k\right) \end{gathered} \tag{7}$$

From here it follows that,

$$\mathcal{W}^B\left(w^B\right) = \begin{cases} \bar{a}^2\left(\frac{\lambda f}{2} - k\right) & \text{if } k < \frac{\lambda f}{2} \\ 0 & \text{if } k \geq \frac{\lambda f}{2} \end{cases}. \tag{8}$$

Comparing the value functions with and without ban we get,

- If $k > \lambda f$, then $\mathcal{W}^B\left(w^B\right) = \mathcal{W}^{LF}\left(w^{LF}\right) = 0$.
- If $k \in \left(\frac{\lambda f}{2}, \lambda f\right]$ then $\mathcal{W}^B\left(w^B\right) < \mathcal{W}^{LF}\left(w^{LF}\right)$
- If $k \leq \frac{\lambda f}{2}$ 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 $\mathcal{W}^B\left(w^B\right) > \mathcal{W}^{LF}\left(w^{LF}\right)$

This completes the proof. ■

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 Remark 1.

To replicate the laissez-faire scenario it suffices to set $\Delta = 0$ and $w^E = w^{LF}$. To replicate the situation when dual practice is banned, simply set $w^E = 0$ and $\Delta = w^B$. Then, all the results in Propositions 1 and 2 follow directly. ■

A.6 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 1). Therefore, in the region $\Delta > w^E$, the best contract yields a value function

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

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

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

subject to the constraints, $w^E \geq 0$, $0 \leq \Delta \leq w^E$.

The f.o.c.'s of this program are:

$$\frac{d\mathcal{W}^E}{dw} = \frac{2\lambda}{k}\left[\frac{\varphi}{1+\delta w} - \frac{2w}{\lambda}\right] = 0$$

$$\frac{d\mathcal{W}^E}{d\Delta} = \frac{2\lambda}{k}\left[\varphi\left(1 - \frac{1}{1+\delta\Delta}\right) - \frac{2\Delta}{\lambda}\right] = 0$$

From here it follows that $w^E = \frac{\sqrt{1+2\delta\lambda\varphi}-1}{2\delta} > 0$ (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 $\delta \leq \frac{2}{\lambda\varphi}$. Secondly, $\Delta^* = \frac{\lambda\varphi\delta-2}{2\delta}$ that fulfills the s.o.c. only if $\delta \geq \frac{2}{\lambda\varphi}$. Finally, when $\Delta^* = \frac{\lambda\varphi\delta-2}{2\delta}$ we also have to check the constraint $\Delta^* \leq w^*$:

$$\Delta^* \leq w^* \iff \delta \leq \frac{4}{\lambda\varphi}.$$

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

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

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

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

It can be shown that,

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

and this is equal to $\mathcal{W}^E\left(\Delta^* = \frac{\lambda\varphi}{2},\ w^* = 0\right)$. Therefore, we have shown that,

- For every $\delta < \frac{4}{\lambda\varphi}$, then the solution is

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

  since

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

- For every $\delta > \frac{4}{\lambda\varphi}$, then the solution is $w^E = 0$, and $\Delta = \frac{\lambda\varphi}{2}$.

This completes the proof. ■

A.7 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 1). Therefore, in the region $\Delta > w^E$, the best contract yields a value function

$$\mathcal{W}^B\left(w^B\right) = \begin{cases} \bar{a}^2\left(\frac{\lambda f}{2} - k\right) & \text{if } k < \frac{\lambda f}{2} \\ 0 & \text{if } k \geq \frac{\lambda f}{2} \end{cases}$$

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

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

subject to the constraints, $w \geq 0$, $0 \leq \Delta \leq w$.

The f.o.c.'s of this program are:

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

There are two candidates that satisfy the f.o.c. for each variable:

$$\begin{aligned} w^* &= 0 \quad \text{or} \\ w^* &= \frac{\lambda f - k}{k\delta} \geq 0 \Longleftrightarrow k \leq \lambda f \\ \Delta^* &= 0 \quad \text{or} \\ \Delta^* &= \frac{k}{\delta(\lambda f - k)} \geq 0 \Longleftrightarrow k \leq \lambda f \end{aligned}$$

When checking the s.o.c. we have that

$$\begin{aligned} \frac{d^2\mathcal{W}^E}{dw^2} &< 0 \text{ for } \begin{cases} w^* = 0 & \text{if } k \geq \lambda f \\ w^* = \frac{\lambda f - k}{k\delta} & \text{if } k \leq \lambda f \end{cases} \\ \frac{d^2\mathcal{W}^E}{d\Delta^2} &< 0 \text{ for } \begin{cases} \Delta^* = 0 & \text{if } k \leq \lambda f \\ \Delta^* = \frac{k}{\delta(\lambda f - k)} & \text{if } k \geq \lambda f \end{cases} \\ \frac{d^2\mathcal{W}^E}{d\Delta dw} &= 0. \end{aligned}$$

It is easy to check that $\Delta^* = \frac{k}{\delta(\lambda f - k)}$ cannot be a solution, as the s.o.c only holds for values of k such that $\Delta^* = \frac{k}{\delta(\lambda f - k)} < 0$. 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 *4*.

This completes the proof. ■

A.8 Proof of Lemmas 3 and 4.

They follow the same steps that the previous lemmas. ■

A.9 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 $\left(\bar{\pi}, w_{\bar{\pi}}\right)$. 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 $\bar{\gamma} > w_{\bar{\gamma}}$) is, by construction, as good as the original $\bar{\pi}$-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, this policy, generates a situation where:

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

Now, let us show that we can find a new duple for the policy that limits the involvement in private practice $(\bar{\gamma}, w_{\bar{\gamma}})$ that is more efficient.

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

$$\frac{2w + \bar{\gamma} + \sqrt{\bar{\gamma}\left(4w - 3\bar{\gamma}\right)}}{2k} = \frac{w + \bar{\Pi}^D}{k} \Longleftrightarrow \bar{\gamma} + \sqrt{\bar{\gamma}\left(4w - 3\bar{\gamma}\right)} = 2\bar{\Pi}^D$$

In words, this means that the physician with ability $a = \frac{w+\bar{\Pi}^D}{k}$ (denote this threshold $\hat{a}$) when limited through maximum earnings will perform an amount of dual practice exactly equal to $\bar{\gamma}$. Note that it is always possible to find such a value of $\bar{\gamma}$, it suffices to take into account that

$$\lim_{\bar{\gamma}\to 0}\left(\bar{\gamma}+\sqrt{\bar{\gamma}\left(4w-3\bar{\gamma}\right)}\right) = 0 < 2\bar{\Pi}^D$$
$$\lim_{\bar{\gamma}\to w}\left(\bar{\gamma}+\sqrt{\bar{\gamma}\left(4w-3\bar{\gamma}\right)}\right) = 2w > 2\bar{\Pi}^D \text{ (since we are in the region with } \bar{\Pi}^D < w_{\bar{\pi}}\text{).}$$

Done this way, both policies imply the same wages and the same number of physicians working in each sector. Now, let us compare the amount of dual practice that dual providers exert with each policy.

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

$$\begin{array}{ll} \text{if } a \in \left[0, \frac{2\bar{\gamma}}{k}\right] & \text{the physician chooses dual practice and } \gamma^*\left(a\right) = \frac{ka}{2} \\ \text{if } a \in \left[\frac{2\bar{\gamma}}{k}, \hat{a}\right] & \text{the physician chooses dual practice and } \gamma^*\left(a\right) = \bar{\gamma} \end{array}$$

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

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

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

$$\bar{\Pi}^D(\hat{\gamma}\left(a\right)) = \left(k\hat{\gamma}\left(a\right)a - \hat{\gamma}\left(a\right)^2\right)^{1/2} = \bar{\Pi}^D$$

Note that we have constructed $\bar{\gamma}$ in such a way that $\hat{\gamma}\left(\hat{a}\right) = \bar{\gamma}$. This, together with the fact that $\hat{\gamma}\left(a\right)$ is decreasing in a implies that for every $a \in \left[\frac{2\bar{\Pi}^D}{k}, \hat{a}\right)$ we have $\hat{\gamma}\left(\hat{a}\right) > \bar{\gamma}$.

Finally, it is easy to check that, since $\bar{\gamma}$ is such that

$$\frac{2w+\bar{\gamma}+\sqrt{\bar{\gamma}\left(4w-3\bar{\gamma}\right)}}{2k} = \frac{w+\bar{\Pi}^D}{k}$$

then $\bar{\gamma} < \bar{\Pi}^D$. 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	$\bar{\pi}$-policy
$a \in \left[0, \frac{2\bar{\gamma}}{k}\right)$	$\gamma^*\left(a\right) = \frac{ka}{2}$	$\gamma^*\left(a\right) = \frac{ka}{2}$
$a \in \left[\frac{2\bar{\gamma}}{k}, \frac{2\bar{\Pi}^D}{k}\right)$	$\bar{\gamma}$	$\gamma^*\left(a\right) = \frac{ka}{2} > \bar{\gamma}$
$a \in \left[\frac{2\bar{\Pi}^D}{k}, \hat{a}\right)$	$\bar{\gamma}$	$\hat{\gamma}\left(a\right) > \bar{\gamma}$
$a = \hat{a}$	$\bar{\gamma}$	$\hat{\gamma}\left(a\right) = \bar{\gamma}$

Under the $\bar{\gamma}$-policy, some physicians (all those in the range of abilities $a \in \left[\frac{2\bar{\gamma}}{k}, \hat{a}\right)$) do less dual practice under the $\bar{\gamma}$-policy than under the $\bar{\pi}$-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.

This completes the proof. ■

A.10 Proof of Proposition 6.

We focus on the case with $\bar{\gamma} \leq w$, and the policy is actually affecting physician behavior. In this case, the HA solves

$$\max_{w,\bar{\gamma}} \mathcal{W}^{\bar{\gamma}} = \lambda\varphi \left(\int_0^{\frac{2\bar{\gamma}}{k}} \frac{1}{1+\frac{\delta k a}{2}} da + \int_{\frac{2\bar{\gamma}}{k}}^{\frac{2w+\bar{\gamma}+\sqrt{\bar{\gamma}(4w-3\bar{\gamma})}}{2k}} \frac{1}{1+\delta\bar{\gamma}} da \right) - w \left(\frac{2w+\bar{\gamma}+\sqrt{\bar{\gamma}\,(4w-3\bar{\gamma})}}{2k} \right),$$

subject to the constraints that $w \geq 0$, and $0 \leq \bar{\gamma} \leq w$.

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 $\alpha \equiv \frac{\bar{\gamma}}{w} \in [0,1]$, where $\alpha = 0$ corresponds to $\bar{\gamma} = 0$ (banning dual practice) and $\alpha = 1$ corresponds to $\bar{\gamma} = w$ (laissez-faire). The objective function can be written as:

$$\mathcal{W}^{\alpha} = \frac{1}{k} \left(\lambda\varphi \left(\frac{2\ln(1+\alpha\delta w)}{\delta} + \frac{\frac{w\left(2+\alpha+\sqrt{\alpha(4-3\alpha)}\right)}{2} - 2\alpha w}{1+\alpha\delta w} \right) - w^2 \left(\frac{2+\alpha+\sqrt{\alpha\,(4-3\alpha)}}{2} \right) \right).$$

We now force a common factor $\frac{1}{\delta^2}$ to the whole function, this yields:

$$\mathcal{W}^{\alpha} = \frac{1}{\delta^2 k} \left(\lambda\varphi\delta \left(2\ln(1+\alpha\delta w) + \frac{\frac{\delta w\left(2+\alpha+\sqrt{\alpha(4-3\alpha)}\right)}{2} - 2\alpha\delta w}{1+\alpha\delta w} \right) - (\delta w)^2 \left(\frac{2+\alpha+\sqrt{\alpha\,(4-3\alpha)}}{2} \right) \right),$$

which shows that the solution to the program will be independent of the parameter k. We finally rename the combined parameter $\lambda\varphi\delta$ as x and δw as $\tilde{w}$. The optimization program can be rewritten as

$$\begin{aligned} \max_{\tilde{w},\alpha} W &= \left(x \left(2\ln(1+\alpha\tilde{w}) + \frac{\frac{\tilde{w}\left(2+\alpha+\sqrt{\alpha(4-3\alpha)}\right)}{2} - 2\alpha\tilde{w}}{1+\alpha\tilde{w}} \right) - \tilde{w}^2 \left(\frac{2+\alpha+\sqrt{\alpha\,(4-3\alpha)}}{2} \right) \right) \\ \text{s.t. } \tilde{w} &\geq 0 \text{ and } \alpha \in [0,1]. \end{aligned}$$

Note that 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.

We will analyze, in turn, each of the cases regarding the constraints. First, it is straightforward to dismiss $\tilde{w}^* = 0$ as a candidate to solution, since the objective function would take value zero. Secondly, corner cases $\alpha = 0$ and $\alpha = 1$ correspond to the ban and laissez-faire scenarios analyzed in Proposition 1. Thus, the optimal values of $\tilde{w}$ can be obtained directly from (3) and (1). The value of the objective function in each of the two cases, adapted from (4) and (2), is

$$\mathcal{W}^{\alpha=0} = \frac{1}{\delta^2 k}\frac{x^2}{4} \qquad \text{and} \qquad \mathcal{W}^{\alpha=1} = \frac{1}{\delta^2 k}\left(-1 - x + \sqrt{2x+1} + 2x\ln\left[\frac{1}{2}\left(1+\sqrt{1+2x}\right)\right]\right).$$

Finally, the optimal values for the interior solution $\alpha \in (0,1)$ and $\tilde{w} > 0$ are the solution of the system formed by the two first order conditions of the optimization program,

$$\frac{\partial W}{\partial \alpha} = -\frac{\tilde{w}\left(2-3\alpha+\sqrt{\alpha\left(4-3\alpha\right)}\right)\left(\tilde{w}\left(1+\alpha\tilde{w}\right)^2 + x\left(-1+\left(-\alpha+\sqrt{\alpha\left(4-3\alpha\right)}\right)\tilde{w}\right)\right)}{2\sqrt{\alpha\left(4-3\alpha\right)}(1+\alpha\tilde{w})^2} = 0$$

$$\frac{\partial W}{\partial \tilde{w}} = \frac{-2\left(2+\alpha+\sqrt{\alpha\left(4-3\alpha\right)}\right)\tilde{w}(1+\alpha\tilde{w})^2 + x(2+\alpha+\sqrt{\alpha\left(4-3\alpha\right)}+4\alpha^2\tilde{w})}{2(1+\alpha\tilde{w})^2} = 0.$$

For $\frac{\partial W}{\partial \alpha} = 0$, either the first parenthesis of the numerator or the second one have to be equal to zero. It is straightforward to see that the first parenthesis is zero if and only if $\alpha = 1$, which corresponds to the corner case analyzed above. Therefore, if $\alpha \in (0,1)$, the solution to the optimization program ($\alpha^*(x)$ and $\tilde{w}^*(x)$) is such that:

$$\tilde{w}\left(1+\alpha\tilde{w}\right)^2 + x\left(-1+\left(-\alpha+\sqrt{\alpha\left(4-3\alpha\right)}\right)\tilde{w}\right) = 0 \qquad (9)$$

$$-2\left(2+\alpha+\sqrt{\alpha\left(4-3\alpha\right)}\right)\tilde{w}(1+\alpha\tilde{w})^2 + x(2+\alpha+\sqrt{\alpha\left(4-3\alpha\right)}+4\alpha^2\tilde{w}) = 0. \qquad (10)$$

The complexity of the system 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.

What is left to do to complete the proof is to compare the welfare attained at the interior solution with that at the two corner cases analyzed above. First, consider the comparison between $\alpha = 0$ and $\alpha^*(x)$. Let us define for any value of x, $\Delta\mathcal{W}^B \equiv \mathcal{W}^{\alpha=0} - \mathcal{W}^{\alpha}(\alpha, \tilde{w})$. Consider now the particular case with $\alpha = 0$ and $\tilde{w} = \tilde{w}\left(\alpha = 0\right) = \frac{x}{2}$. Trivially, $\Delta\mathcal{W}^B = 0$. We now show that it is possible to find values of $\alpha \in (0,1)$ such that $\mathcal{W}^{\alpha=0} < \mathcal{W}^{\alpha}(\alpha, \frac{x}{2})$. This follows directly from the fact that $\frac{\partial \Delta\mathcal{W}^B}{\partial\alpha}_{|\left(\alpha=0, \tilde{w}=\frac{x}{2}\right)} = -\infty$. This is sufficient to ensure that $\mathcal{W}^{\alpha=0} < \mathcal{W}^{\alpha}(\alpha^*(x),\ \tilde{w}^*(x))$ and, therefore, that setting $\alpha = 0$ is never a solution.

Consider, finally, the comparison between $\alpha = 1$ and $\alpha^*(x)$. Similarly as before, let us define for any value of x, $\Delta\mathcal{W}^{LF} \equiv \mathcal{W}^{\alpha=1} - \mathcal{W}^{\alpha}(\alpha^*(x), \tilde{w}^*(x))$. It can be shown that $\forall x > 0$, $\Delta\mathcal{W}^{LF}$ is decreasing in x. Moreover, $\Delta\mathcal{W}^{LF} = 0$ if and only if $x = 2$, i.e., if $\delta = \frac{2}{\lambda\varphi}$. Thus, for any $\delta > \frac{2}{\lambda\varphi}$, $\Delta\mathcal{W}^{LF} < 0$ meaning that the interior solution $(\alpha^*(x), \tilde{w}^*(x))$ provides higher welfare than the corner one with $\alpha = 1$ (laissez-faire). Conversely, for $\delta < \frac{2}{\lambda\varphi}$, the solution with $\alpha = 1$ and $\tilde{w} = \frac{1}{2}\left(\sqrt{2\lambda\varphi\delta+1} - 1\right)$ provides the highest welfare. And this completes the proof. $\blacksquare$

A.11 Proof of Proposition 7.

We focus on the case with $\bar{\gamma} \leq w$, and the policy is actually affecting physician behavior. In this case, the HA solves

$$\max_{w,\bar{\gamma}} \mathcal{W}^{\bar{\gamma}} = \lambda f \left(\int_0^{\frac{2\bar{\gamma}}{k}} \frac{a}{1+\frac{\delta k a}{2}} da + \int_{\frac{2\bar{\gamma}}{k}}^{\frac{2w+\bar{\gamma}+\sqrt{\bar{\gamma}(4w-3\bar{\gamma})}}{2k}} \frac{a}{1+\delta\bar{\gamma}} da \right) - w \left(\frac{2w+\bar{\gamma}+\sqrt{\bar{\gamma}(4w-3\bar{\gamma})}}{2k} \right),$$

subject to the constraints that $w \geq 0$, and $0 \leq \bar{\gamma} \leq w$.

Again, and without loss of generality, we make some manipulations on the objective function in order to work with a more compact optimization program. First, we do a change of variable and define $\alpha \equiv \frac{\bar{\gamma}}{w} \in [0,1]$, where $\alpha = 0$ corresponds to $\bar{\gamma} = 0$ (banning dual practice) and $\alpha = 1$ corresponds to $\bar{\gamma} = w$ (laissez-faire). The objective function can be written as:

$$\mathcal{W}^{\alpha} = \frac{\lambda f}{k^2} \left(\frac{w^2 \left(-16\alpha^2 + \left(2+\alpha+\sqrt{\alpha(4-3\alpha)} \right)^2 \right)}{8(1+\alpha\delta w)} + \frac{4(\alpha\delta w - \ln(1+\alpha\delta w))}{\delta^2} \right) - \frac{w^2 \left(2+\alpha+\sqrt{\alpha(4-3\alpha)} \right)}{2k},$$

Some algebraic manipulations allow us to rewrite the objective function as:

$$\mathcal{W}^{\alpha} = \frac{\lambda f}{8k^2\delta^2} \left[\left(\frac{w^2\delta^2 \left(-16\alpha^2 + \left(2+\alpha+\sqrt{\alpha(4-3\alpha)} \right)^2 \right)}{(1+\alpha\delta w)} + 32(\alpha\delta w - \ln(1+\alpha\delta w)) \right) - 4w^2\delta^2 \frac{k}{\lambda f} \left(2+\alpha+\sqrt{\alpha(4-3\alpha)} \right) \right],$$

We finally define $K \equiv \frac{k}{\lambda f}$ and $\tilde{w} \equiv \delta w$. Hence, the optimization program can be rewritten as*:*

$$\max_{\tilde{w},\alpha} W = \left[\left(\frac{\tilde{w}^2 \left(-16\alpha^2 + \left(2+\alpha+\sqrt{\alpha(4-3\alpha)} \right)^2 \right)}{(1+\alpha\tilde{w})} + 32(\alpha\tilde{w} - \ln(1+\alpha\tilde{w})) \right) - 4\tilde{w}^2 K \left(2+\alpha+\sqrt{\alpha(4-3\alpha)} \right) \right]$$

s.t. $\tilde{w} \geq 0$ and $\alpha \in [0,1]$.

This program is simpler (but equivalent) to the original one. The variable that determines the intensity of the limiting policy, α, is defined now 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 $k < \lambda f$ (see footnote 16), what restricts the space of K to $K \in (0,1)$.

We analyze, in turn, each of the cases regarding the constraints. First, it is straightforward to dismiss $\tilde{w}^* = 0$ as a candidate to solution, since the objective function would take value zero. Secondly, corner cases $\alpha = 0$ and $\alpha = 1$ correspond to the ban and laissez-faire scenarios analyzed in Proposition 2. Thus, the optimal values of $\tilde{w}$ can be obtained directly from (7) and (5). The value of the objective function in each of the two cases, adapted from (8) and (6), is

$$\begin{aligned}
\mathcal{W}^{\alpha=0} &= \max\left\{0, \lambda f \bar{a}^2 \left(\frac{1}{2} - K\right)\right\} \text{ and} \\
\mathcal{W}^{\alpha=1} &= \frac{\lambda f}{8k^2\delta^2}\left(-\frac{16(1-K)^2}{K} + 32\left(\frac{(1-K)}{K} - \ln\left[1 + \frac{(1-K)}{K}\right]\right)\right).
\end{aligned}$$

Finally, the optimal values for the interior solution $\alpha \in (0,1)$ and $\tilde{w} > 0$ are the solution of the system formed by the two first order conditions of the optimization program,

$$\begin{aligned}
\frac{\partial W}{\partial \alpha} = \frac{\left(2-3\alpha+\sqrt{\alpha(4-3\alpha)}\right)\tilde{w}^2}{\sqrt{\alpha(4-3\alpha)}(1+\alpha\tilde{w})^2}&\left[-4-2\sqrt{\alpha(4-3\alpha)}+4K+2\tilde{w}\sqrt{\alpha(4-3\alpha)}+\right. \\
&\left.+3\alpha\tilde{w}\sqrt{\alpha(4-3\alpha)}-\alpha^2\tilde{w}(4K\tilde{w}-5)+2\alpha(4K\tilde{w}-1)\right] \\
\frac{\partial W}{\partial \tilde{w}} = \tilde{w}&\left[-8K\left(2+\alpha+\sqrt{\alpha(4-3\alpha)}\right) - \frac{\alpha\tilde{w}\left(-16\alpha^2+\left(2+\alpha+\sqrt{\alpha(4-3\alpha)}\right)^2\right)}{(1+\alpha\tilde{w})^2}+\right. \\
&\left.+\frac{2\left(2+\alpha+\sqrt{\alpha(4-3\alpha)}\right)^2}{1+\alpha\tilde{w}}\right].
\end{aligned}$$

For $\frac{\partial W}{\partial \alpha} = 0$, either the parenthesis $\left(2-3\alpha+\sqrt{\alpha(4-3\alpha)}\right)$ or the term in brackets have to be equal to zero. It is straightforward to see that the parenthesis is zero if and only if $\alpha = 1$, which corresponds to the corner case analyzed above. Therefore, if $\alpha \in (0,1)$, the solution to the optimization program ($\alpha^*(K)$ and $\tilde{w}^*(K)$) is such that:

$$\begin{aligned}
&-4-2\sqrt{\alpha(4-3\alpha)}+4K+2\tilde{w}\sqrt{\alpha(4-3\alpha)}+ \\
&+3\alpha\tilde{w}\sqrt{\alpha(4-3\alpha)}-\alpha^2\tilde{w}(4K\tilde{w}-5)+2\alpha(4K\tilde{w}-1)=0
\end{aligned} \tag{11}$$

$$\begin{aligned}
&-8K\left(2+\alpha+\sqrt{\alpha(4-3\alpha)}\right) - \frac{\alpha\tilde{w}\left(-16\alpha^2+\left(2+\alpha+\sqrt{\alpha(4-3\alpha)}\right)^2\right)}{(1+\alpha\tilde{w})^2}+ \\
&+\frac{2\left(2+\alpha+\sqrt{\alpha(4-3\alpha)}\right)^2}{1+\alpha\tilde{w}} = 0.
\end{aligned} \tag{12}$$

First of all we can see that there does not always exist an interior solution. In particular, it is easy to check that for $K = 0$ equation (12) does not hold as it is always positive. Similarly, for $K = 1$ equation (12) 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 such extreme values of K the solution will be therefore either the one with $\alpha = 0$ or $\alpha = 1$.

The complexity of the system prevents us from fully characterizing the interior solution, but we can easily show that there exist intermediate values of K for which it exists. To prove existence it suffices to take, for instance $K = 0.54$. For this particular value the system formed by (11) and (12) yields $\alpha^*(K = 0.54) \simeq 0.224517$ and $\tilde{w}^*(K = 0.54) \simeq 1.26873$. What can be easily proven is that an increase in δ will translate in a decrease in $\bar{\gamma}$. 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. The comparison between $\alpha = 0$ and $\alpha = 1$ was done in Proposition **2** and hence the results there directly apply. Therefore,

- If $K \in \left(\frac{1}{2}, 1\right)$ then $\mathcal{W}^{\alpha=0} < \mathcal{W}^{\alpha=1}$.
- If $K \leq \frac{1}{2}$ then, there exists a threshold $\bar{\delta}_2 > 0$ such that,
  - If $\delta < \bar{\delta}_2$ then $\mathcal{W}^{\alpha=0} < \mathcal{W}^{\alpha=1}$.
  - If $\delta > \bar{\delta}_2$ then $\mathcal{W}^{\alpha=0} > \mathcal{W}^{\alpha=1}$.

Finally, to show that it is possible to find intermediate values of K for which an interior solution with $\alpha \in (0, 1)$ is optimal it suffices to consider again $K = 0.54$. A direct computation shows that for this value of K, it holds that $\mathcal{W}^{\alpha=1}_{|K=0.54} < \mathcal{W}^{\alpha}\left(\alpha^*(K = 0.54), \tilde{w}^*(K = 0.54)\right)$. Since $K = 0.54 > \frac{1}{2}$, this ensures that $\mathcal{W}^{\alpha}\left(\alpha^*(K = 0.54), \tilde{w}^*(K = 0.54)\right) > \mathcal{W}^{\alpha=1} > \mathcal{W}^{\alpha=0}$ and, therefore, that $\alpha^* \in (0, 1)$ is optimal. This completes the proof. ■

A.12 Proof of Proposition 8.

For developing countries the proof follows directly from combining Propositions 2, 4 and 7.

For developed countries, the proof that for $\delta \leq \frac{2}{\lambda\varphi}$ the best is not to regulate dual practice and that for $\delta > \frac{4}{\lambda\varphi}$ the best policy is to impose a limit 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 $\delta \in [\frac{2}{\lambda\varphi}, \frac{4}{\lambda\varphi}]$ a policy of limits dominates that with exclusive contracts.

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

$$\mathcal{W}^{E*} \equiv \mathcal{W}^E\left(\Delta = \frac{\lambda\varphi\delta - 2}{2\delta},\ w^E = \frac{\sqrt{1+2\delta\lambda\varphi}-1}{2\delta}\right) =$$
$$= \frac{-6 - 2\delta\lambda\varphi + (\delta\lambda\varphi)^2 + 2\sqrt{1+2\delta\lambda\varphi} - 4\delta\lambda\varphi\ln(\delta\lambda\varphi) + 4\delta\lambda\varphi\ln(1+\sqrt{1+2\delta\lambda\varphi})}{2\delta^2 k}$$

Defining $x \equiv \delta\lambda\varphi$ we can rewrite the welfare as:

$$\mathcal{W}^{E*} = \frac{-6 - 2x + (x)^2 + 2\sqrt{1+2x} - 4x\ln(x) + 4x\ln(1+\sqrt{1+2x})}{2\delta^2 k}.$$

Note that, as we are in the region with $\delta \in [\frac{2}{\lambda\varphi}, \frac{4}{\lambda\varphi}]$, this function is only defined for $x \in [2, 4]$.

We compare now $\mathcal{W}^{E*}$ with the welfare obtained under the optimal policy for $x \in [2, 4]$, 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 (9) and (10).

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 that $\forall x \in [2, 4]$, $\breve{\mathcal{W}}$ is decreasing in x. Moreover, we find that when $x = 2$, $\breve{\mathcal{W}} = 0$. Therefore, for any $x \in [2, 4]$, i.e., $\delta \in [\frac{2}{\lambda\varphi}, \frac{4}{\lambda\varphi}]$ it holds that $\mathcal{W}^\alpha(\alpha^*(x), \tilde{w}^*(x)) > \mathcal{W}^{E*}$. This completes the proof. ■