WORKING PAPERS

N° 1595

November 2024

"Incentivizing Physicians' Diagnostic Effort and Test with Moral Hazard and Adverse Selection"

David Bardey, Philippe De Donder and Marie-Louise Leroux

Toulouse School of Economics

Incentivizing Physicians' Diagnostic Effort and Test with Moral Hazard and Adverse Selection*

David Bardey†, Philippe De Donder‡, Marie-Louise Leroux§

November 21, 2024

Abstract

We study a situation where physicians differing in their degree of altruism exert a diagnostic effort before deciding whether to test patients to determine the most appropriate treatment. The diagnostic effort generates an imperfect private signal of the patient's type, while the test is perfect. At the *laissez-faire*, physicians exert insufficient diagnostic effort and rely excessively on testing. We show that the first-best allocation (where the degree of altruism is observable) can be decentralized by a payment scheme composed of i) a pay-for-performance (P4P) part based on the number of correctly treated patients to ensure the provision of the optimal diagnostic effort, and of ii) a capitation part to ensure both the optimal testing decision and the participation of physicians. When physicians differ in their (non-observable) degree of altruism, the optimal contract is pooling rather than separating, an instance of non-responsiveness. Its uniform P4P component induces more altruistic physicians to exert a larger diagnostic effort while, to incentivize the second-best optimal testing decision, its capitation component must be contingent on the test cost.

Keywords: diagnostic risk, personalized medicine, non-responsiveness, capitation payment, pay-for-performance, hidden action and hidden information.

JEL codes: D82, D86, I18.

*We thank participants and especially discussants to the following conferences and seminars: 2023 European Health Economic Workshop (Augsburg), 2023 Barcelona Summer Forum, 13^{th} Conference on Economic Design (Girona), 22nd Journées Louis-André Gérard-Varet (Marseille), 1st CEPR Health Economics Conference (TSE, Toulouse, 2024), 45th Journés des Economistes Français de la Santé (Bordeaux), Digital Health Workshop (Toulouse, 2024), 63rd Congrès de la Société Canadienne de Sciences Economiques (Montreal), GATE Lyon, Le Mans University, Public University of Navarra, PUC seminar, 2024 Canadian Public Economist Group conference (Hamilton). We also thank D. Alary, F. Barigozzi, M. Cassou, G. Dionne, I. Jelovac, M. Kifmann, B. Montmartin, L. Siciliani, A. Terrieau for their comments and suggestions. Financial support from the ANR (Programme d'Investissement d'Avenir ANR-17-EURE-0010), from the Chaire "Marché des risques et création de valeurs, fondation du risque/Scor", from Fonds de Recherche du Québec-Société et Culture (FRQSC, grant number: 2024-SE3-328700) is gratefully acknowledged.

†Universidad de Los Andes and Toulouse School of Economics, Email: d.bardey@uniandes.edu.co

‡Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France. Email: philippe.dedonder@tse-fr.eu

§Département des Sciences Economiques, ESG-UQAM, Montréal, Canada; CESifo, Munich, Germany; CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium. E-mail: leroux.marie-louise@uqam.ca

1 Introduction

The healthcare sector represents a substantial portion of many developed countries economies. In 2022, for example, the OECD average for health spending was approximately 9.6% of GDP. Notably, the United States leads OECD nations with the highest share, at over 16% of GDP, while countries like Canada, Germany, France, and the United Kingdom consistently spend around 10-12% of their GDP on health. These figures, obtained from OECD(2023), include both public and private spending on health services, pharmaceuticals, and long-term care. Moreover, the indirect economic impact of health spending is also important, encompassing lost earnings, reduced leisure time, and diminished home production, which further amplify the sector's overall influence.

Given the importance of this health sector, it is crucial to consider physicians as major economic players whose varied medical decisions, even when treating patients with observationally similar characteristics, can significantly affect healthcare outcomes. This variability not only impacts treatment quality but also contributes to a central concern in healthcare: diagnostic errors. According to the World Health Organization, diagnostic errors (defined as missed, incorrect, delayed or miscommunicated diagnoses) account for 16% of preventable patient harm, representing a leading cause of medical malpractice cases in the United States.[1] Understanding and addressing the economic causes of these errors is essential for improving patient care and reducing the associated financial burdens.

In that respect, the rise of precision medicine, defined as the creation of treatments highly effective for specific patient subgroups, may help reduce diagnostic errors and improve health outcomes. This approach requires diagnostic tests to identify if a patient will benefit from personalized therapies. Examples include targeting specific molecular markers for optimal cancer treatment, molecular profiling of microbes to distinguish between bacteria, fungi, or viruses, employing diverse antibiotic families against various bacteria types or to combat resistance, and leveraging pharmacogenomics for tailored drug prescriptions and dosages. The future of

[1] `https://www.who.int/campaigns/world-patient-safety-day/world-patient-safety-day-2024`. Accessed November 11, 2024.

precision medicine looks promising, especially with advancements in AI technologies, as explored by Mullainathan and Obermeyer (2017, 2019).

The value of precision medicine relies not only on the availability of targeted treatments and diagnostic tests but also on physicians' incentives to prescribe these tests efficiently. While some healthcare systems see an overuse of diagnostic tests which do not necessarily improve patient outcomes,[2] underuse of diagnostic tests, particularly for detecting pathogens like bacteria, fungi, or viruses, can negatively impact patients' health. Proper test utilization is crucial to ensure that precision medicine treatments are applied appropriately. To quote Currie *et al.* (2024), "these new tools can be over-used, under-used, and can lead to harmful consequences for patients when used inappropriately. Understanding how humans can interact with the tools to produce better outcomes is a first order question" (p.37).

The literature on physicians' payment schemes has extensively explored the effects of models like fee-for-service, capitation, salary, and payment-for-performance (P4P) on the quantity and quality of medical care, using a positive approach to understand their impacts. Another branch of the literature adopts a normative approach to design optimal payment models. We review both strands in Section 2. Although both approaches have been studied, few works examine these remuneration schemes in the context of diagnostic tests. This manuscript addresses that gap by exploring the properties of payment schemes when diagnostic tools are available.

We develop a model involving two patients' types (A and B) and two treatments (D and P). Consider for instance patients who suffer from bacterial infections of the urinary tract. Type-B patients are infected by Escherichia coli (E. coli), the most common cause of urinary tract infections (UTIs), while type-A patients suffer from Klebsiella pneumoniae, another bacteria responsible for UTIs. These two bacteria show different patterns of antibiotic resistance. For instance, E. coli is susceptible to ciprofloxacin, while Klebsiella could be resistant to it but sensitive to another antibiotic, such as cefuroxime. Initially, the patient's type is unknown to

[2] As pointed out by Currie *et al.*(2024), there is a large literature on the overuse of imaging technology– see Felder and Kifmann (2024) for MRIs for instance. Kowalski (2003) likewise documents overuse of mammography in Canada. Currie *et al.* (2024, section 4.4) surveys the recent empirical literature on the link between financial incentives and under-/overuse of medical technology, revealing that these patterns are often driven by supply- rather than demand-side considerations.

both the patient and the physician, with the physician only aware of the population distribution of types. In the absence of information about a patient's type, the optimal approach is to treat everyone with the less expensive, default treatment D – ciprofloxacin in our example. A type-A (*i.e.* Klebsiella infection) patient should rather be treated with the personalized treatment P (cefuroxime). A diagnostic test, such as a urine culture with an antibiotic sensitivity test (antibiogram) is able to identify the responsible bacterium in order to determine the most effective antibiotic.

Physicians can employ two methods to determine a patient's type. The first is traditional diagnostic effort, such as detailed consultations (lengthy auscultation, temperature reading, hunt for other symptoms) or teasing out of relevant family medical history, which is costly to the physician and produces signals with accuracy increasing in effort. The second method consists in using a diagnostic test, which has higher accuracy and, for simplicity, is assumed to perfectly identify the patient's type.

However, diagnostic tests come with costs for the patient. These costs can be monetary when subject to copayments, raising out-of-pocket expenses, or non-monetary, such as treatment delays, invasiveness, or potential side effects. Because these costs are rather more related to tests' features than patients' characteristics, we will consider regulations that may depend on the diagnostic tests' cost supported by patients.

Lastly, we assume that correctly matching the treatment to the patient's type (*i.e.*, applying treatment D for type-B patients and treatment P for type-A) results in fewer required visits or actions from the physician. Although our results would hold without this assumption, it reflects the practical observation that when physicians can make more accurate diagnoses, the need for follow-up visits or adjustments tends to decrease. This highlights the value of diagnostic precision in reducing the overall healthcare burden for both physicians and patients.

We adopt standard medical guidelines by analyzing a scenario where physicians first decide how much effort to exert in diagnosing a patient and then determine whether to proceed with a diagnostic test based on the signal they receive after exerting that effort. The decision to test follows the physicians' evaluation of the initial diagnostic signal, which may guide them in

identifying whether further diagnostic testing is warranted for precise treatment allocation.

In our model, physicians are assumed to be partially altruistic.[3] We first characterize the first-best allocation where a social planner observes the physicians' altruism degree. The planner determines the physicians' effort level and then decides on the use of diagnostic tests based on the signal received, while internalizing the cost borne by patients. Three cases arise depending on the level of diagnostic test costs. For low costs, testing all patients is optimal, rendering physician effort unnecessary. For intermediate costs levels, tests are only prescribed to patients with signal A, aligning with common medical practice.[4] When the cost of a diagnostic test is high, it is not prescribed to any patient, regardless of the signal received. Additionally, physicians' effort and diagnostic tests are strategic substitutes. As more patients undergo the test, physicians tend to reduce their effort, since the test provides more reliable information about the patient's condition. This makes exerting costly effort less appealing, as the test will accurately identify the patient's type, diminishing the need for the physicians' diagnostic efforts.

We then look at how to decentralize this (first-best) allocation, anticipating that physicians will choose both the effort level and whom to test based on the signal they observe and the compensation they receive from the social planner. In a *laissez-faire* scenario, physicians tend to exert insufficient diagnostic effort and overprescribe diagnostic tests. This occurs because physicians do not fully account for the costs borne by patients. We show that when the physicians' effort is observable and contractible, the social planner can decentralize the first-best allocation with a transfer combining a P4P component that varies based on the physician's testing decision, and a fixed capitation payment that remains the same across different cases, whether all patients are tested, only those with signal A are tested, or no tests are conducted. When effort is not observable or contractible, the social planner can still implement the first-best allocation by linking payments to the number of well-treated patients. In this situation, capitation payments

[3] Altruism may be an inherent trait of the doctor or can be viewed more broadly as a simplified representation of reputational concerns or apprehension about potential malpractice lawsuits.

[4] When treating bacterial infections, diagnostic tests are not always necessary for mild cases (*e.g.*, fever), where standard antibiotics like ciprofloxacin for UTIs are effective. However, for patients with more severe symptoms, performing a diagnostic test to identify the specific bacteria type is warranted, as in the case of the Klebsiella infection treated by cefuroxime described above.

are adjusted based on the physicians' aggregate testing decisions. Although the balance between P4P and capitation payments changes, the total payment to physicians remains the same as in the case where effort is observable.

In the remainder of the manuscript, we consider the more realistic scenario where the physicians' degree of altruism is their private information. We show that the optimal contract in this context is a pooling contract offered to all physicians regardless of their altruism level. This outcome is driven by the property of non-responsiveness. Indeed, at the *laissez faire*, low-altruism physicians exert less effort than high-altruism ones, with both levels falling short of the optimal effort. Incentivizing low-altruism physicians to increase more their effort is then optimal. However, their lower concern for patients' welfare necessitates higher compensation, leading to a misalignment between the social planner's objectives and physicians' preferences. Consequently, a separating contract that differentiates between physicians' types becomes unattainable.

The pooling contract's P4P component induces second-best optimal effort levels and is determined by the average level of altruism among physicians, causing those who are less (resp., more) altruistic than average to exert less (resp., more) diagnostic effort than what would be socially optimal under the first-best allocation. When the capitation part of the pooling contract is set to satisfy the physicians' participation constraints, it leaves rents to less altruistic physicians. Those rents are increasing with the diagnostic effort levels, and since diagnostic efforts and testing decisions are substitutes, they induce physicians to under-utilize the tests. The social planner can correct this bias by increasing capitation levels when more patients are tested. We show that the second-best allocation can be reached provided that transfers can be conditioned on both the fraction of patients tested (all, none, or some) and the value of the test cost.

The manuscript is organized as follows. The next section reviews the related literature. Section 3 describes the model. In Section 4, we derive the first-best allocation, while in Section 5, we analyze the physicians' problem. Section 6 shows how to decentralize the first-best allocation, first when effort is observable and contractible, and second, when it is not. In Section 7, we relax the assumption that physicians' altruism is observable and study the optimal second-best

contract. Section 8 concludes.

2 Related Literature

Our paper lies at the intersection of two branches of the literature: one where physicians differ in altruism and choose treatments, but without diagnostic tests, and one where physicians make use of diagnostic tests but do not exhibit various altruism degrees.

Many health economics articles have studied the behavior of physicians differing in altruism when choosing treatments for their patients, without access to a diagnostic test.[5] While early contributions assume that the degree of altruism is public information,[6] more recent papers consider that doctors are heterogeneous with respect to their altruism degree which is their private information. For instance, while the seminal paper by Jack (2005) studies non-contractible quality, Choné and Ma (2011) assume that health care quantities are contractible but that physicians also have private information about their patient's illness severity before accepting a contract. Both papers show that it is impossible to decentralize the first-best allocation in these cases. Liu and Ma (2013) show that the first-best allocation can be decentralized with asymmetric information only if physicians can commit to a treatment plan before accepting a payment contract.

In the second branch of the related literature, devoted to the study of incentives for both diagnostic tests and treatment choices, most papers assume the existence of moral hazard (hidden action –diagnostic effort– and hidden information –signal from diagnostic effort), but do not consider adverse selection related to physicians' heterogeneity.

The seminal paper by Garcia Mariñoso and Jelovac (2003) introduces a setting where an income-maximizing physician first makes a costly diagnostic effort, receives an imperfect signal

[5] Currie *et al.* (2024) review the recent empirical economic literature on physician decision-making, highlighting how factors such as diagnostic skills, beliefs, patient demographics, incentives, training, experience, and external interventions (like guidelines and decision tools) influence doctors' treatment choices. They stress that doctors "care about patient welfare, but also about their own welfare which makes them imperfect agents. (p.36)" As is common in the health economics literature surveyed here, we label the relative weight put on the patient's utility (as opposed to these other considerations) the degree of altruism of the doctor.

[6] See for instance Allard, Jelovac and Leger (2011), Chalkley and Malcomson (1998), Ellis and McGuire (1986, 1990) and Rochaix (1989).

and then decides whether to treat the patient or send her to a specialist for advanced treatment. The specialist's advanced treatment works with certainty, but is costlier than the physician's treatment. As in Liu and Ma (2013), the first- and second-best allocations depend on the nature of the physician's participation constraint: while the first-best allocation is achieved when the physician's participation constraint is satisfied *ex ante*, the social planner has to leave informational rents to physicians when their participation constraint has to be satisfied in each and every state of the world. In both cases, the optimal contract offered to physicians is composed of a capitation part, plus a bonus if they correctly treat patients themselves (as the planner observes the return visit of any badly treated patient), and an additional fee if they refer patients to a specialist.

Beenk and Kifmann (2024) build on the same setting in which physicians exert a costly effort and study the optimal payment contracts for two subsequent tests in a treatment choice problem. The first test is costly to the physician and generates an imperfect private signal while the second test is costly to the payer but perfect. The payer can only observe whether the second test is taken and its result. The profit-maximizing physician decides whether to use either of both tests. The optimal contract inducing a selective use of the second test has the same flavor as in Garcia Mariñoso and Jelovac (2003) and includes a capitation payment for performing the diagnosis (first test), payments conditional on applying a successful treatment, and a fee for ordering the second test.[7]

Our setting also borrows several elements from Adida and Dai (2024) where an imperfectly altruistic physician decides first her diagnostic effort level and then whether to test the patient for a severe disease. As in our model, the effort generates a symmetrical imperfect signal, while the test is perfect. While our focus is normative (*i.e.* how to decentralize first- and second-best allocations allowing for different physicians' payment schemes), they focus on the impact of a fee-for-service payment on the incentives for doctors to exert effort and to test their patients.

[7]Pignataro (2024) analyzes an adverse selection model where (egoistic) physicians differ in their unobservable diagnostic ability. As in Beenk and Kifmann (2024), doctors first choose whether to exert a costly but imperfect diagnostic (binary) effort and then whether to order a (perfect) genetic test. As in our paper, Pignataro (2024) shows that diagnostic effort and test decision are substitutes. See also Brandt and Cassou (2024) for an application to prospective payments in hospitals, resulting in optimal cross-subsidizations within care pathways.

Finally, Felder and Kifmann (2024), like our paper, lies at the intersection of these two branches of the literature.[8] They develop a model where patients have varying prior probabilities p of needing major versus minor treatments. In this model, physicians differ in their levels of altruism and have the ability to observe each patient's p value privately and without cost. Additionally, unlike our approach, Felder and Kifmann do not incorporate physician effort; instead, they focus on the use of a costly and imperfect diagnostic test by physicians. In their model, the first-best allocation consists in prescribing, without testing, the treatment for mild (resp., severe) disease if the patient's probability p is below (resp., above) a low (resp., high) threshold, and to test (and follow the test recommendation for the treatment) patients with intermediate values of p. When altruism is not observable, the social planner offers a menu of contracts, where the physician's cost share is below its first-best level, inducing them all to overtreat some of their patients. This is done in order to decrease the physicians' rents obtained in equilibrium by all physicians except the most altruistic one.

Compared to the existing literature, and to borrow the distinction introduced by McGuire (2000) in his survey on physician agency, we are, to the best of our knowledge, the first paper to study a context with both moral hazard (with hidden action and hidden information) and adverse selection problem (on the degree of altruism of physicians) and to analyze both the first-best and second-best optimal incentive schemes for physicians who first exert a diagnostic effort and then decide whether to prescribe a diagnostic test.

3 The model

3.1 Patients and treatments

There are two types of patients, indexed by $i \in \{A, B\}$, with a proportion λ of type A and $1-\lambda$ of type B. As in Bardey *et al.* (2020), for simplicity, we fix $\lambda = 1/2$.[9] Nobody (neither the

[8] Ghamat *et al.* (2018) and Dai and Singh (2020) also introduce adverse selection in set-ups that consider diagnostic tests. While the adverse selection is on the physician's ability in Dai and Singh (2020), private information concerns the patients' characteristics in Ghamat *et al.* (2018).

[9] In Appendix 9.7, we relax this assumption and solve the model for a generic value of $\lambda < 1/2$. While the analysis becomes more intricate (for reasons we refer to in footnotes 10, 19 and 21), our results do not qualitatively change.

patient nor the physician) knows the patient's type at the beginning of the period.

There are two treatments available to patients, indexed by $j \in \{P, D\}$, where D stands for the "default" treatment while P stand for the "personalized" treatment, as we shall see. The utility that a patient of type i receives from a treatment j is denoted by U_i^j. It can be seen for instance as the medical value of the treatment, minus its cost for the patient. We shall use the following notation.

Definition 1 *(i)* $\Delta U_A \equiv U_A^P - U_A^D$, *(ii)* $\Delta U_B \equiv U_B^D - U_B^P$.

We make the following assumption.

Assumption 1 $\Delta U_B > \Delta U_A > 0$.

The assumption that ΔU_A and ΔU_B are both strictly positive reflects that a type-A patient should be treated with P, while a type-B patient should be treated with D. It can be the case that treatment P for type A patients (or D for type B) provides greater medical benefits than the alternative treatment, or that the higher medical service rendered by the other treatment is not worth its additional cost.

The assumption that $\Delta U_B > \Delta U_A$ ensures that treatment D is the default treatment, namely the one that should be provided to all agents in the absence of any information on their individual type. Comparing expected utility with D and with P, we obtain that D should be given by default if

$$\frac{U_B^D + U_A^D}{2} > \frac{U_B^P + U_A^P}{2} \iff \Delta U_A < \Delta U_B.$$

This condition is equivalent to assuming that the relative gain of treating B with the D treatment is, on average, higher than the relative loss of treating A types with D (instead of P). Otherwise, it would be better to prescribe the treatment P (instead of D) to everybody.

3.2 Doctors' effort and diagnostic tests

There are two (non-exclusive) ways for physicians to obtain more information on the true type of their patient: a diagnostic effort (or clinical assessment) measured for instance by the time

spent with the patient, the thoroughness of examining symptoms, and the investigation into the patient's medical and family history; and a diagnostic test. We explain the characteristics of the two technologies in turn.

The physician can exert an effort, ε which generates a signal about the patient's type. The signal $\sigma \in \{A, B\}$ has the following precision

$$\begin{aligned} \varepsilon &= \Pr(\sigma = B \,|\, i = B) \\ &= \Pr(\sigma = A \,|\, i = A) \\ &\in [1/2, 1]. \end{aligned}$$

So, the minimum amount of effort corresponds to $1/2$ (minimum time and energy spent on a patient) while the maximum is equal to 1. Table 1 shows the updated probabilities after having received a signal with precision ε.

Type $\rightarrow$ Signal $\downarrow$	B	A	Total
B	$\frac{\varepsilon}{2}$	$\frac{1-\varepsilon}{2}$ false neg.	$1/2$
A	$\frac{1-\varepsilon}{2}$ false pos.	$\frac{\varepsilon}{2}$	$1/2$
	$1/2$	$1/2$	

Table 1: Frequencies in population

A few comments are in order. First, the minimum effort level, $1/2$, gives no information to the physician, since there is a one half *ex post* probability for each type whatever the signal received. Second, the maximum effort level of one guarantees a perfectly informative signal. Third, increasing the effort level improves the quality of the signal in a symmetrical way, reducing by the same amount the probability of false positives and false negatives.[10] This effort has a cost to the physician, which is increasing and convex and denoted by $\psi(\varepsilon)$. We assume an Inada's

[10] As we show in Appendix 9.7, with a generic proportion λ of type-A patients, the observed proportion of A signals differs from λ and depends also on ε, with an over-representation of signals corresponding to the minority type. The Appendix discusses the implications of these findings for our analysis.

condition such that $\psi(1/2) = 0$, $\psi'(1/2) = 0$ and $\lim_{\varepsilon \to 1} \psi'(\varepsilon) = +\infty$, to prevent corner solutions in the choice of ε. Finally, while we consider in the sequel both the cases where the effort is observable or not to the planner, we assume throughout that the signal on the patient's type that the physician has received remains her private information.

The other technology available to doctors consists in prescribing a diagnostic test which reveals the type of the agent with 100% accuracy, but generating a cost z to patients. This cost may account for different factors: direct monetary expenses, opportunity costs, or a utility loss. For example, the utility cost may arise from an invasive test, exposition to potential harmful radiation, or from the delay in initiating treatment, as time is required for conducting, processing, and interpreting the diagnostic test.[11]

Along the paper, we assume that both the *fraction* of patients tested[12] and the test cost z are observable (and contractible) by the health authority, while the test result remains the doctor's private information.[13]

The timing of the game runs as follows. First, the health authority proposes a payment scheme to physicians, who either accept or reject it.[14] In the latter case, the game stops. Second, physicians choose a diagnostic effort level, $\varepsilon \in [1/2, 1]$, generating a private signal about the patient's type. Third, physicians decide whether to run a diagnostic test based on the signal received.[15] Fourth, physicians prescribe a treatment (D or P), and the payoffs are realized.

3.3 Payoffs

Although patients do not make decisions in this model, their welfare is crucial for defining both the social optimal outcomes and guiding physicians' choices. We add two elements to the utilities

[11] We address both types of costs (treatment and diagnostic test) symmetrically by assuming that the patient bears them both. This approach prevents our results from being influenced by any imbalance in who is responsible for the costs, ensuring that the conclusions are not driven by such an asymmetry.

[12] We do not require that the test decision be observable for each individual patient.

[13] If this were not the case, it would necessitate the introduction of collusion-proof contracts to prevent the patient and physician from colluding (see, for instance, Wu *et al.* [2021]). This approach would be impractical in our setup, where the patient has a passive role.

[14] We introduce explicitly the participation constraints in Section 5.

[15] This sequence follows WHO(2014)'s guidelines on diagnostic tests.

U_i^j defined above. First, we allow for the possibility that treating a patient adequately (*i.e.* with P if A and with D if B) allows the physician to restore the patient's health with fewer visits. We then define two levels for the number of visits required to treat a patient: e^M (where M stands for "matched treatment") if the patient is correctly treated (*i.e*, treatment D for true type B and P for true type A), and e^{NM} ("non-matched" treatment) otherwise, with $e^M \leqslant e^{NM}$. We define $\Delta e \equiv e^{NM} - e^M \geqslant 0$ as the gain in number of visits when a patient is correctly treated.[16] We further add the possibility that patients dislike seeing their doctor, with a linear cost $\gamma \geqslant 0$ per visit (either the opportunity cost of the time spent with the doctor or a utility cost).[17]

The utility of a patient of type $i = \{A, B\}$ prescribed treatment $j = \{D, P\}$ is then

$$\tilde{U}_i^j = U_i^j - \gamma e^m - lz,$$

where $l = \{0, 1\}$ according to whether a diagnostic test is prescribed or not, and where $m = \{M, NM\}$ according to whether the treatment matches the patient's type or not.

Doctors care both about their own income (*i.e.* any transfer T received from the planner), minus their cost of effort, $\psi(\varepsilon)$, and about their patient's welfare. More precisely, we assume that the doctor puts a weight of $\alpha \in [0, 1]$ on the patient's utility, so that her utility is given by

$$V = T - \psi(\varepsilon) + \alpha \tilde{U}_i^j. \tag{1}$$

We denote by α the degree of altruism of the doctor, and we call her imperfectly altruistic if $\alpha < 1$.

We now proceed as follows. Next section describes the socially optimal allocation. Section 5 studies the physicians' optimization problem, while Section 6 decentralizes the optimum. These sections assume that the degree of altruism of physicians is observable by the planner. In contrast, Section 7 relaxes this assumption.

[16] In our setting, as demonstrated below, a fee-for-service payment mechanism will never be optimal (*i.e.* it cannot decentralize the optimum). Such a mechanism would decrease welfare by incentivizing doctors to administer inappropriate treatments to patients in order to increase the number of visits.

[17] All our results carry through to the simplified scenario where both Δe and γ are set to zero.

4 The social optimum

We consider a utilitarian social planner who maximizes the total utility of individuals while excluding the altruistic component of the physician's utility.[18] We denote W as the objective of the social planner, which is influenced by the level of effort provided and the decision of whether to conduct a diagnostic test. We first determine the optimal effort level before addressing the testing decisions. We assume that the remuneration paid to the physician to guarantee her participation is seen as a pure transfer by the planner (*i.e.*, there is no cost of public funds), and thus plays no role in this section.

4.1 Optimal effort levels

We have to deal with three possible cases, where we denote the first-best optimal level of a variable with a star.

Case All: Test all patients (whatever the signal received). In such a case, welfare as a function of effort level ε_{All} is given by

$$W_{All}(\varepsilon_{All}) = -\psi(\varepsilon_{All}) + \frac{1}{2}U_A^P + \frac{1}{2}U_B^D - z - \gamma e^M.$$

Effort is useless (*i.e.* $\varepsilon^*_{All} = 1/2$ and $\psi(\varepsilon^*_{All}) = 0$), because it is costly to exert, while the test anyway will reveal the patient's type with certainty.

Case 0: No test is prescribed to anyone.

In such a case, the welfare function is a function of the effort level ε_0,

$$W_0(\varepsilon_0) = -\psi(\varepsilon_0) + \frac{\varepsilon_0}{2}(U_B^D + U_A^P) + \frac{1-\varepsilon_0}{2}(U_A^D + U_B^P) - \gamma(\varepsilon_0 e^M + (1-\varepsilon_0)e^{NM}),$$

where the third term accounts for classification errors: false negatives occur when type A patients are incorrectly treated with treatment D due to being mistaken for type B patients, while false positives happen when type B patients are erroneously treated with treatment P because they

[18] Horizontal equity concerns make it undesirable to assign greater weight to patients fortunate enough to be treated by physicians with higher levels of altruism. Other papers proceed in the same way, such as Beenk and Kifmann (2024, Appendix F), Chalkley and Malcomson (1998), Liu and Ma (2013).

are mistaken for type A patients. The first-order condition for ε_0 is:

$$\psi'(\varepsilon_0^*) = \frac{\Delta U_B + \Delta U_A}{2} + \gamma \Delta e. \tag{2}$$

The intuition for the first term in the right-hand side is that a marginal increase in effort decreases by one half the number of both false positives (type B patients who would otherwise be mistakenly treated with P, with a per person gain of ΔU_B) and false negatives (type A patients who would otherwise be mistakenly treated with D, with a per person gain of ΔU_A). The intuition for the second term is that we gain Δe visits each time the doctor makes more effort (half of type B and half of type A).

The corresponding welfare level is given by

$$W_0(\varepsilon_0^*) = -\psi(\varepsilon_0^*) + \frac{\varepsilon_0^*}{2}(U_B^D + U_A^P) + \frac{1-\varepsilon_0^*}{2}(U_A^D + U_B^P) - \gamma(\varepsilon_0^* e^M + (1-\varepsilon_0^*)e^{NM}).$$

Note that if ε_0^* is sufficiently small, the previously discussed solution may produce lower welfare compared to a strategy where no effort is exerted, and all patients are treated with D, regardless of their signals. In the following, we exclude this possibility and assume that ε_0^* satisfies

$$W_0(\varepsilon_0^*) > \frac{1}{2}(U_B^D + U_A^D - \gamma(e^M + e^{NM})).$$

Case 1: Test prescribed only if signal A is received

When the test is prescribed only after observing a signal A, welfare as a function of effort level ε_1 becomes

$$W_1(\varepsilon_1) = -\psi(\varepsilon_1) + \frac{1}{2}U_B^D + \frac{\varepsilon_1}{2}U_A^P + \frac{1-\varepsilon_1}{2}U_A^D - \frac{z}{2} - \gamma[\frac{1+\varepsilon_1}{2}e^M + \frac{1-\varepsilon_1}{2}e^{NM}],$$

where the diagnostic test eliminates false positives, *i.e.*, patients of type B who sent a type-A signal. This is achieved at a cost z for half of the sample that sent an A-signal. Once the false positives are identified through testing, they are treated with the appropriate default treatment D.

The first-order condition for ε_1 is

$$\psi'(\varepsilon_1^*) = \frac{\Delta U_A}{2} + \frac{\gamma}{2}\Delta e, \tag{3}$$

with, compared to (2), a marginal gain only on false negatives (ΔU_A) and correspondingly only half the marginal gain in number of visits.[19]

This gives the optimal welfare level

$$W_1(\varepsilon_1^*) = -\psi(\varepsilon_1^*) + \frac{1}{2}U_B^D + \frac{\varepsilon_1^*}{2}U_A^P + \frac{(1-\varepsilon_1^*)}{2}U_A^D - \frac{z}{2} - \gamma[\frac{1+\varepsilon_1^*}{2}e^M + \frac{1-\varepsilon_1^*}{2}e^{NM}].$$

We then obtain the following proposition:

Proposition 1 *Effort and test are strategic substitutes:* $\varepsilon_{All}^* < \varepsilon_1^* < \varepsilon_0^*$.

Proof. Immediate comparison of both first-order conditions, acknowledging the convexity of the function $\psi(.)$. ■

The intuition runs as follows: as explained after equation (3), when a diagnostic test is run only for patients receiving an A-signal, the marginal benefit of the physician's diagnostic effort is lower compared to the scenario where no test is run at all. This is because in the test scenario, effort only reduces false negatives, while the test itself identifies the false positives. Consequently, physicians exert less effort when a test is used as a backup technology, allowing them to rely on the test rather than exerting diagnostic effort upfront.[20] Importantly, the optimal effort levels are independent of the test cost z.

Finally, we must rule out the case where it would be optimal to test after receiving the signal B (rather than A).

Lemma 1 *Under Assumption 1, testing only patients with a signal B is dominated by testing only patients who signal A, whatever the effort level.*

[19] In Appendix 9.7, we show that, for a generic value of $\lambda < 1/2$, the optimal level of effort when testing only A-signals, ε_1^*, depends on the test cost, z. This is a direct consequence of the result, mentioned in footnote 10, that the fraction of A-signals varies with the effort level.

[20] This relationship is confirmed empirically by Chu *et al.* (2024) who show that Emergency Departments doctors substitute testing for their time and attention.

Proof. The welfare level reached when only signal-B agents are tested is

$$\frac{1}{2}U_A^P + \frac{1-\varepsilon}{2}U_B^P + \frac{\varepsilon}{2}U_B^D - \psi(\varepsilon) - \frac{z}{2} - \gamma[\frac{1+\varepsilon}{2}e^M + \frac{1-\varepsilon}{2}e^{NM}] < W_1(\varepsilon),$$

for any ε if and only if $\Delta U_A < \Delta U_B$ is satisfied. ■

Testing only patients with a signal B (resp., A) allows to eliminate the false negatives (resp., positive), with a per patient marginal gain of ΔU_A (resp., ΔU_B). Assumption 1 (which guarantees that D, rather than P, is the default treatment) then ensures that testing only signal-A patients is socially preferred to testing only signal-B patients.

4.2 Optimal diagnostic test decision

The planner must decide whether to test no one, only those with a signal A, or everyone, represented by cases All, 1 or 0 respectively. This decision is influenced by the cost of the test, z. We thus need to compare $W_{All}(1/2)$, $W_1(\varepsilon_1^*)$ and $W_0(\varepsilon_0^*)$ as a function of the value of z. More precisely, we focus on the more interesting scenario where, as the cost of the test increases, the optimal decision changes from testing everyone to only testing patients with an A signal, and ultimately to not testing anyone at all. In contrast, the abrupt shift from testing everyone to testing no one is less interesting and rarely observed in practice. Furthermore, all key insights derived in this analysis would remain applicable in the context of this simpler framework.

We define the threshold z_{All}^* as the level of the test cost below which it is optimal to make no effort and to test everyone in the population. This is the case when $W_{All}(1/2) \geqslant W_1(\varepsilon_1^*)$. This threshold level is obtained when the above equation holds with equality, that is when

$$z_{All}^* \equiv (1-\varepsilon_1^*)(\Delta U_A + \gamma \Delta e) + 2\psi(\varepsilon_1^*). \tag{4}$$

The intuition behind this formulation of the threshold is the following. The only gain (from a welfare perspective) from testing only patients with a A signal, rather than all patients, is that we only test half the population, saving half the cost of the test. This gain is balanced by three losses, corresponding to the three (positive) terms defining z_{All}^*: (i) a loss of correct treatments among the A-agents (those who send signal B), (ii) an increase in the number of visits for those

patients, and (iii) a cost of effort for the physician. The larger these three costs, the larger the value of z below which it is socially optimal to test everyone.

We now define z_1^* as the threshold level of test cost at which the social planner is indifferent between treating only patients signalling A and not treating anyone, $W_1(\varepsilon_1^*) = W_0(\varepsilon_0^*)$, so that:

$$z_1^* \equiv 2\left(\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*)\right) + (1-\varepsilon_0^*)\,\Delta U_B - (\varepsilon_0^* - \varepsilon_1^*)\,\Delta U_A + \gamma(1+\varepsilon_1^* - 2\varepsilon_0^*)\Delta e. \tag{5}$$

The intuition behind this formulation is similar to that of z_{All}^*. Going from testing signal A only to testing nobody allows to save on the test cost for half the population. This has four consequences corresponding to the four terms above. First, more effort is needed if nobody is tested, which is socially costly (first term). Second, not testing anyone results in *fewer correctly treated B- types* (because all B types are correctly treated if we test patients with a signal A, while this is not true in Case 0), hence a positive second term above. Third, not testing anyone results in *more correctly treated A- types*. The reasons are as follows: (i) testing A signals does not help treating correctly the true type A in Case 1, so that (ii) the proportion of type A correctly treated only depends on effort levels ε, and (iii) effort is larger in Case 0 than in Case 1. This in turn results in a negative third term. Fourth, because effects (ii) and (iii) are of opposite signs, the number of correctly treated patients may increase or decrease when moving from Case 1 to 0. This last term is positive if there is a lower proportion of correctly treated patients in Case 0 (low ε_0^*) than in Case 1 (high ε_1^*), and is twice more sensitive to ε_0^* than to ε_1^* because the correct treatment of both types B and A increases with ε_0, while only type-A patients are affected by ε_1.[21]

Note that in the following, we assume that $z_{All}^* < z_1^*$ as the opposite relationship would correspond to a situation where the planner should move abruptly from testing everyone to no-one as a test cost threshold is crossed, a situation we have excluded at the beginning of this section.

[21] Since, for a generic value of $\lambda < 1/2$, ε_1^* depends on the cost of the test z, we face a fixed point issue when computing z_1^*. This additional complication, analyzed in Appendix 9.7, makes the model significantly more complicated to solve without bringing any commensurate additional insight.

We now turn to the analysis of the setting where the physician (rather than the social planner) chooses both how much effort to exert and who to submit to a diagnostic test (Section 5), as a necessary prelude to the decentralization of the optimal allocation (Section 6).

5 The physicians' problem

The physician maximizes her utility (1) with respect to both her effort level and testing decision (*i.e.*, who to submit to a diagnostic test). More precisely, the physician's utility depends on whether she tests all patients (V_{All}), only those with a A signal (V_1) or nobody (V_0):

$$\begin{aligned}
V_{All} &= \frac{\alpha}{2}[U_A^P + U_B^D - 2z - 2\gamma e^M] + T_{All}(.) - \psi(\varepsilon_{All}), \\
V_1 &= \frac{\alpha}{2}\{U_B^D + \varepsilon_1 U_A^P + (1-\varepsilon_1)U_A^D - z - \gamma[(1+\varepsilon_1)e^M + (1-\varepsilon_1)e^{NM}]\} \\
&+ T_1(.) - \psi(\varepsilon_1), \\
V_0 &= \frac{\alpha}{2}\{\varepsilon_0[U_A^P + U_B^D] + (1-\varepsilon_0)[U_A^D + U_B^P] - 2\gamma(\varepsilon_0 e^M + (1-\varepsilon_0)e^{NM})\} \\
&+ T_0(.) - \psi(\varepsilon_0),
\end{aligned}$$

where $T_k(.)$ are payments received by the doctor in Case $k = \{All, 1, 0\}$ and are functions of effort levels assumed for the moment to be both observable and contractible.[22] This yields the following (equilibrium) levels of efforts:

$$\psi'(\varepsilon_{All}^{eq}) = T'_{All}(\varepsilon_{All}^{eq}), \tag{6}$$

$$\psi'(\varepsilon_0^{eq}) = \alpha\left(\frac{\Delta U_B + \Delta U_A}{2} + \gamma\Delta e\right) + T'_0(\varepsilon_0^{eq}), \tag{7}$$

$$\psi'(\varepsilon_1^{eq}) = \alpha\left(\frac{\Delta U_A}{2} + \gamma\frac{\Delta e}{2}\right) + T'_1(\varepsilon_1^{eq}). \tag{8}$$

The left-hand side of the above equations measures the physician's marginal cost of effort, while the right-hand side represents its marginal private benefit. In equation (6), the right-hand side term only includes the variation in the transfer received as effort varies. In equations (7) and (8), it also accounts for the marginal social benefit of effort (equal to zero in Case *All*), multiplied by the physician's altruism parameter.

[22] Recall that we assume throughout the paper that the planner observes the fraction of patients' tested, and thus knows the Case (All, 0 or 1) in which physicians operate.

Comparing (2) to (7), and (3) to (8), we obtain that, at the *laissez faire*, less-than-perfectly altruistic physicians under-provide effort in Cases 0 and 1. Moreover, total differentiation of (7) and (8) shows that effort is increasing in altruism, so that the lower the altruism degree α, the more the physician under-provides effort.

We now compute the equilibrium partition of whether to test or not, namely the thresholds z_{All}^{eq} and z_1^{eq}. The first threshold z_{All}^{eq} is such that $V_{All}(z_{All}^{eq}) = V_1(z_{All}^{eq})$, and this condition yields

$$z_{All}^{eq} = (1 - \varepsilon_1^{eq})(\Delta U_A + \gamma \Delta e) + \frac{2}{\alpha}(T_{All} - T_1) + \frac{2}{\alpha}\psi(\varepsilon_1^{eq}). \tag{9}$$

Comparing (4) and (9), we see that the doctor over-emphasizes (compared to the social planner) both her cost of effort (last term in (9)) and the difference in transfers received in Cases *All* and 1 (second term). At the *laissez-faire*, the second term is nil and the cost of effort drives her to test everyone for larger values of z. The less altruistic the physician is, the greater the over-testing behavior. Note however that, once differentiated transfers are introduced (as we shall do in the next sections), this effect may be compensated by offering physicians a larger transfer in Case 1 than in Case *All*.

We proceed in the same way for z_1^{eq}, which is such that $V_0(z_1^{eq}) = V_1(z_1^{eq})$ and we obtain that:

$$z_1^{eq} = (1-\varepsilon_0^{eq})\Delta U_B - (\varepsilon_0^{eq} - \varepsilon_1^{eq})\Delta U_A + \gamma(1+\varepsilon_1^{eq} - 2\varepsilon_0^{eq})\Delta e + \frac{2}{\alpha}(T_1 - T_0) + \frac{2}{\alpha}(\psi(\varepsilon_0^{eq}) - \psi(\varepsilon_1^{eq})). \tag{10}$$

Comparing (5) and (10), we see that the doctor over-emphasizes (compared to the social planner) both the difference in effort costs and in transfers received in Cases 0 and 1. At the *laissez-faire*, her larger effort in Case 0 than in Case 1 induces her to test only signal-A patients for larger values of z. Similar to the situation for z_{All}^{eq}, less altruistic physicians are more prone to over-testing. Again, this could be counterbalanced by giving her a larger transfer in Case 0 than in Case 1.

We summarize those results in the following proposition.

Proposition 2 *At the* laissez-faire *allocation, physicians exert too little effort ($\varepsilon_k^{eq} < \varepsilon_k^*, \forall k \in \{0,1\}$) and rely too much on testing ($z_k^{eq} > z_k^*, \forall k \in \{All, 1\}$). Moreover, the under-provision of*

effort and the tendency to over-test decrease as the physician's altruism degree, represented by α*, increases.*

We now introduce formally the physicians' participation constraints

$$T_k(\varepsilon_k^{eq}) \geqslant \psi(\varepsilon_k^{eq}),\ \forall k \in \{All, 1, 0\}, \tag{11}$$

which require that the transfer received from the authority in any Case $k \in \{All, 0, 1\}$ has to compensate for the effort disutility. Note that the term $\alpha\tilde{U}_i^j$ is absent from the participation constraint, meaning that the planner cannot take advantage of the doctor's altruism to reduce the required transfer amount.[23]

We are now in a position to look at the decentralization of the first-best allocation.

6 First-best decentralization of the optimum (with observable altruism)

We proceed as in Section 4, starting by decentralizing the effort levels and then moving to the testing decisions. We first state the general formulas for decentralization, before looking at how to operationalize them as a function of whether effort is observable/contractible. Recall that we assume for the moment that the physician's altruism degree is observable.

6.1 General formulas

In order to make the optimal and the equilibrium levels of efforts coincide, we need to set:

$$T'_{All}(\varepsilon_{All}^{eq}) = 0, \tag{12}$$

$$T'_0(\varepsilon_0^{eq}) = (1-\alpha)\,[\frac{\Delta U_B + \Delta U_A}{2} + \gamma\Delta e], \tag{13}$$

$$T'_1(\varepsilon_1^{eq}) = (1-\alpha)\,[\frac{\Delta U_A}{2} + \frac{\gamma\Delta e}{2}]. \tag{14}$$

The intuition behind these formulas is straightforward.[24] In both cases the transfer is such that, at the margin, it complements the altruistic part of the doctor's utility to induce her to

[23] These participation constraints can be interpreted as limited-liability constraints, as in, among others, Liu and Ma (2013) and Felder and Kifmann (2024).

[24] This result is similar in spirit to the one obtained by Felder and Kifmann (2024).

behave as if she were perfectly altruistic (since the term between square bracket measures the effort's marginal social benefit).

Moreover, in order to ensure that z_{All}^{eq} and z_1^{eq} correspond to their optimal levels (once the effort levels have been optimally chosen), we need to set payment functions $T_k(.)$ satisfying

$$T_1(\varepsilon_1^*) - T_{All}(\varepsilon_{All}^*) = (1-\alpha)\psi(\varepsilon_1^*), \tag{15}$$

$$T_0(\varepsilon_0^*) - T_1(\varepsilon_1^*) = (1-\alpha)[\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*)]. \tag{16}$$

We have seen when comparing the equilibrium and optimal values of the thresholds that doctors over-weigh their cost of effort when choosing who to test, and that this distortion can be counteracted by offering them a larger transfer when the effort required is larger. Moreover, this distortion decreases with how altruistic the doctor is. So, we obtain in both cases (eq. (15) and (16)) that the difference in transfers between two adjacent cases is $(1-\alpha)$ times the difference in (optimal) effort cost in each case. These conditions determine the difference between transfers in adjacent cases, while the participation constraint (11) will set the (minimum) absolute level of transfers.

It is clear from above that a fixed transfer, such as a capitation payment, cannot by itself decentralize the first-best allocation, because it will fail to incentivize the doctor to exert the optimal level of effort. Whether we can decentralize using a mixed payment scheme that would combine capitation and P4P depends on what we can observe and condition the contract upon. We treat the various possibilities, starting with the least constrained one.

6.2 Effort observable and contractible

In this section, we assume that the payment can be conditioned directly on the effort level. While it may not be the most realistic scenario in practice, this will serve as an interesting benchmark. We can then easily recover the simplest forms of the payment functions $T_0(.)$ and $T_1(.)$. Denoting from now on the part of the transfers that does not depend on effort (*i.e.*, the capitation component) with an upper bar, we set $T_{All}(\varepsilon_{All}) = \bar{T}$ so that $\varepsilon_{All}^{eq} = \varepsilon_{All}^* = 1/2$, and

we get that:

$$T_1(\varepsilon_1) = \bar{T} + (1-\alpha)\psi(\varepsilon_1), \tag{17}$$

$$T_0(\varepsilon_0) = \bar{T} + (1-\alpha)\psi(\varepsilon_0). \tag{18}$$

This means that, if the social planner could directly observe and condition transfers on effort, it could use a payment scheme composed of a capitation payment $\bar{T}$ and of a variable one depending on effort so as to induce the doctor both to exert the first-best effort level, and to choose the first-best testing strategies (*i.e.* z^*_{All} and z^*_1). The variable part of the transfer equals the effort cost weighted by how far from perfectly altruistic the physician is. The difference in the variable transfers obtained at equilibrium between two adjacent cases is then $(1-\alpha)$ times the difference in effort cost between cases, so that (15) and (16) are both satisfied provided that the *same* capitation level $\bar{T}$ is served in Cases 1 and 0.

In other words, as long as the social planner corrects the doctor's incentives at the margin (for the effort choices), there is no need for an additional correction (through the fixed –*i.e*, capitation– part of the transfer) of the testing decisions. This is because our specific scenario allows for transfers to be directly conditioned on the (observable) level of effort exerted. In the next section, we will show that this result does not hold anymore when effort is not contractible.

The constant term, common to both (17) and (18) and denoted by $\bar{T}$, is then set to satisfy the doctor's participation constraint (11). With no cost of public fund, we could set $\bar{T}$ arbitrarily high. It is nevertheless interesting to look at the lowest fixed part compatible with the doctor's participation. It is equal to α times the cost of effort (since the variable part is $(1-\alpha)\psi(\varepsilon^*_k)$, compensating for the lack of altruism of the doctor). Note that

$$\max\left[0, \alpha\psi(\varepsilon^*_1), \alpha\psi(\varepsilon^*_0)\right] = \alpha\psi(\varepsilon^*_0),$$

which then corresponds to the minimum (uniform) capitation (*i.e.*, to be paid in all cases) $\bar{T}$ to satisfy the doctor's participation constraint.

Note that the participation constraint is then binding in Case 0, but that physicians earn

rents (equal to $\alpha\left(\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*)\right)$) in Case 1, and especially in Case *All* ($\alpha\psi(\varepsilon_0^*)$).[25] Interestingly, these rents levels increase with the doctor's altruism level.[26]

We summarize these results in the following proposition.

Proposition 3 *When effort is observable and contractible, the social planner can decentralize the first-best allocation with a transfer composed of a capitation element together with a pay-for-performance component, as given by (17) and (18). The capitation part has to be the same in all cases. Its minimum level compatible with the physician's participation leaves rent to the latter in Cases* 1 *and All, rents which increase in her degree of altruism.*

The intuition for why the capitation payment increases with altruism stems from the planner's need to ensure physician participation. When physicians exhibit greater altruism, they receive a lower variable portion of the transfer from the health authority, necessitating a higher capitation level to adequately cover the costs associated with optimal effort, which remains constant across all physicians in the first-best scenario.

6.3 Number of patients correctly treated observable and contractible

In this section, we assume that effort is **not** contractible. We show that this does not prevent the decentralization of the optimum as long as the number of well-treated patients can be observed (and contracted upon).[27] It is due to the fact that, in both Cases 0 and 1, the number of correctly-treated/matched patients is a monotone function of the effort levels.

We define n_k as the number of correctly-treated patients in Case $k \in \{All, 1, 0\}$. We also denote by n_1^A the number of correctly-treated type-A patients in Case 1. In Case *All*, $n_{All} = 1$ so that T_{All} is a constant denoted by $\bar{T}_{All}$. In Case 0, we obtain from Table 1 that $n_0 = \varepsilon_0$ (half

[25] We define as *rent*, the difference between the monetary transfer received by the doctor and her cost of effort. While this rent does not impose an explicit cost on the social planner (as it constitutes a transfer that does not factor into social welfare in the absence of public fund costs), we still strive to minimize it.

[26] See Bardey and Siciliani (2021) for a similar result in a two-sided environment applied to nurses' market.

[27] It is common practice to assume, as Abaluck *et al.* (2016) for example, that since people with a missed diagnostic for a serious medical condition will likely return to see their health providers, whether they have been correctly treated is observable after the treatment. In the US, the Hospital Readmissions Reduction Program (HRRP) penalized hospitals with Medicare readmission rates that were higher than a given threshold (see Gupta, 2021). Wilding *et al.* (2022) study a similar English policy which imposed financial penalties on GPs when the fraction of hypertensive patients with blood pressure under control fell below a target.

of them being B types, the other half A types). We then obtain that condition (13) is satisfied if we use the following contract which is linear in the number of well-treated patients:

$$T_0(n_0) = \bar{T}_0 + (1-\alpha)[\frac{\Delta U_B + \Delta U_A}{2} + \gamma \Delta e] n_0. \tag{19}$$

The slope of the transfer is proportional to the marginal utility gain of better treating patients (*i.e.* $(\Delta U_A + \Delta U_B)/2 + \gamma \Delta e$), corrected by how egoistic the physician is.

In Case 1, we have $1/2$ correctly-treated B patients (since all type-B patients are identified either through the diagnostic effort or the diagnostic test) and $n_1^A = \varepsilon_1/2$ correctly-treated A-type patients, for a total of $n_1 = (1+\varepsilon_1)/2$ correctly-treated patients. Thus, the contract that allows the social planner to decentralize the optimal effort level in that case is given by

$$T_1(n_1) = \bar{T}_1 + (1-\alpha)[\Delta U_A + \gamma \Delta e] n_1, \tag{20}$$

since differentiating it satisfies (14).

Comparing the square bracket terms in (19) and (20), we see that they only differ by the first term, which is larger for (19) since $\Delta U_B > \Delta U_A$. For any given α, the planner needs to generate stronger incentives to exert effort in Case 0 because the marginal gains of a correct treatment (when one more agent is well treated thanks to more effort) concerns both A and B, while in Case 1 it only concerns type A (with a lower marginal gain ΔU_A). The second part in the square bracket is identical because the gain in number of visits does not depend on type (A or B) but only on the treatment being correct.[28]

We then obtain the following proposition.

[28] In Case 1, the number of correctly-treated type-B patients is a constant, so alternatively the social planner need not reward the correct treatment of these individuals. The transfer can thus be made contingent only on n_1^A, with

$$T_1(n_1^A) = \bar{T}_1^A + (1-\alpha)[\Delta U_A + \gamma \Delta e] n_1^A. \tag{21}$$

Differentiating this function with respect to ε_1 gives (14). The only difference between (21) and (20) is the value of the capitation part (which should be determined by the participation constraint, see below in the text). The reason the linear part is the same is simply because

$$\frac{\partial n_1}{\partial \varepsilon_1} = \frac{\partial n_1^A}{\partial \varepsilon_1} (= 1/2).$$

In practice, it may be difficult to reward the correct treatment of some patients (A) but not of others (B).

Proposition 4 *When the number of correctly-treated patients is observable and contractible, the social planner can decentralize the first-best allocation with a transfer composed of a capitation element together with a P4P component, as given by (19) and (20). The capitation part has to be lower in Cases 0 and 1 than in Case All, with* T_{All} *set at the same level than when effort is contractible. The rents enjoyed by physicians in Cases* 1 *and All increase with their degree of altruism.*

Proof. See Appendix 9.8. ■

The above proposition implies that the capitation levels in Cases 0 and 1 are lower than when effort is contractible. The rationale for this is as follows. Since the number of correctly-treated patients increases *linearly* with the physician's effort, the P4P component's slope is proportional to the marginal cost of providing the *optimal* effort level. As we show in the appendix, given that the effort cost is convex, this results in a variable portion of the transfer at equilibrium that is higher than when effort is contractible. Therefore, to ensure the correct testing decision, the capitation component must be reduced. In other words, while the total transfer at equilibrium remains the same as in the previous section for all three cases, the distribution between capitation and P4P differs (with lower capitation and higher P4P) in Cases 0 and 1 compared to the scenario where effort is contractible.

Since the P4P component of the contract decreases with the physician's level of altruism (as the need to incentivize effort diminishes), a lower portion of the effort cost must be covered through the capitation payment for less altruistic doctors. At the extreme and as we show in the appendix, for doctors with very low altruism, the P4P component can be so substantial that the capitation payment may even turn negative to prevent them from under-testing at equilibrium.

7 Asymmetric information on doctors' altruism: A second-best analysis

In this section, we assume that the physician's degree of altruism is her private information, and we look at how to decentralize her optimal diagnostic effort and testing decisions.[29] We concentrate on two-part tariffs, namely payment schemes consisting of both a fixed (capitation) level and a variable (pay-for-performance) part.

We then proceed as follows. In Section 7.1, we show that asymmetric information on the degrees of altruism is best tackled by proposing a pooling contract. In Section 7.2, we find the optimal levels of the P4P part of the physician's remunerations. Finally, in Section 7.3, we find the optimal levels of the capitation parts, which will depend on both the testing decision and the cost of the diagnostic test.

7.1 The second-best contract is pooling

From now on, we assume for simplicity that there exist two types of physicians, type-H physicians with a high degree of altruism, α_H, and type-L physicians with a low degree of altruism, α_L such that $\alpha_L < \alpha_H$.[30] There is a proportion ν of low-altruism doctors. The physician's altruism degree is her private information, and the contract can only be conditioned on the number of correctly-treated patients (*i.e.*, effort is not contractible).

We index all the variables by the (non observable) physician (altruism) type $i \in \{L, H\}$ and by the (observable and contractible) Case $k \in \{All, 0, 1\}$. Note that, even under asymmetric information, the first-best optimum can still be implemented in Case All by setting $T_{All} = \bar{T}_{All}$, inducing $\varepsilon_{i,All} = 1/2$ and $\psi(\varepsilon_{i,All}) = 0$. We then focus from now on Cases 0 and 1.

[29] Recall that physicians accept the payment scheme proposed by the planner before making their effort and test decisions. This differentiates our model from the commitment scenario in Liu and Ma (2013), where all doctors, except the least altruistic one, are constrained in their decisions by the binding participation constraint they accepted when committing to their future treatment decisions.

[30] Our finding that the second-best contract is pooling does not hinge on the restriction to two types. In fact, separating contracts are easier to sustain in a two-type setting, so our argument holds *a fortiori* with more types, including a continuum of types.

As before, the contract includes both a P4P component as well as a capitation part:

$$T_{i,k} = \bar{T}_{i,k} + \beta_{i,k} n_{i,k}.$$

Recall that the number of correctly-treated patients in Cases 1 and 0 are respectively: $n_{i,1} = (1 + \varepsilon_{i,1})/2$ and $n_{i,0} = \varepsilon_{i,0}$. In such a context, the physician's utility becomes:

$$\alpha_i B_k(\varepsilon_{i,k}) + \bar{T}_{i,k} + \beta_{i,k} n_{i,k} - \psi(\varepsilon_{i,k}). \tag{22}$$

The first term above is the altruism term where $B_k(\varepsilon_{i,k})$ is the expected utility of patients depending on the case considered and the level of effort provided by physicians. In Case 1, it takes the following form

$$B_1(\varepsilon_{i,1}) = \frac{1}{2} U_B^D + \left(\frac{\varepsilon_{i,1}}{2}\right) U_A^P + \left(\frac{1-\varepsilon_{i,1}}{2}\right) U_A^D - \frac{z}{2} - \frac{\gamma}{2}[(1+\varepsilon_{i,1})e^M + (1-\varepsilon_{i,1})e^{NM}],$$

while in Case 0, it is equal to

$$B_0(\varepsilon_{i,0}) = \left(\frac{\varepsilon_{i,0}}{2}\right)(U_B^D + U_A^P) + \left(\frac{1-\varepsilon_{i,0}}{2}\right)(U_A^D + U_B^P) - \gamma[\varepsilon_{i,0} e^M + (1-\varepsilon_{i,0})e^{NM}].$$

The equilibrium levels of efforts are obtained as a solution to the maximization of the physicians' utility (22) with respect to $\varepsilon_{i,k}$:

$$\psi'(\varepsilon_{i,k}) = \alpha_i \bar{b}_k + \mathbb{1}_k \beta_{i,k}, \tag{23}$$

with $\mathbb{1}_1 = 1/2$, $\mathbb{1}_0 = 1$, and where $B'_k(\varepsilon_{i,k}) = \bar{b}_k$ are constants, differing across Cases k but independent of $\varepsilon_{i,k}$:

$$\bar{b}_1 = \frac{\Delta U_A + \gamma \Delta e}{2}, \tag{24}$$

$$\bar{b}_0 = \frac{\Delta U_A + \Delta U_B}{2} + \gamma \Delta e. \tag{25}$$

Proposition 5 *The Second-Best contracts in Cases 0 and 1 are pooling.*

Proof. See Appendix 9.1, where we first demonstrate that any set of separating contracts designed to increase the effort level of type L more than that of type H (by setting $\beta_{L,k} > \beta_{H,k}$) cannot simultaneously satisfy all incentive constraints (*i.e.*, one of the two types of doctors will

always want to mimic the other type). We then show that a contract where $\beta_{L,k} < \beta_{H,k}$ results in lower welfare compared to offering the same contract (with $\beta_{L,k} = \beta_{H,k}$) to both types. ■

The intuition for this result is as follows. At the *laissez-faire*, both physicians' type underprovide effort, with those with low altruism providing less effort than those with high altruism. The social planner's welfare function is increasing and concave in effort because the marginal social benefit of effort is constant while the effort cost is convex. Consequently, the planner aims to particularly incentivize the low-altruism type to increase her effort level. However, the low-altruism physician enjoys a lower net benefit from increasing her effort, as she places less importance on the patient's utility. Technically, this means that while the classical single-crossing property condition is met, the slope of the indifference curve (in the effort, transfer plane) is steeper for the low-altruism doctor so that she has to be compensated more for increasing her effort. This corresponds to what Laffont and Martimort (2002) call *non-responsiveness*, where "the sharp conflict between the principal's preferences and the incentive constraints (which reflect the agent's preferences) makes it impossible to use any information transmitted by the agent about his type" (p.55).

In our context, this results in the same, pooling contract being proposed to all physicians:[31]

$$T_{i,k} = \bar{T}_k + \beta_k n_{i,k}. \tag{26}$$

Note that the formal proof developed in Appendix 9.1 makes use of the following assumption:

Assumption 2 *The utility cost of effort takes the following quadratic form:*

$$\psi(\varepsilon) = \frac{(\varepsilon - 1/2)^2}{2}.$$

This assumption is made for simplicity (*i.e.* to obtain closed form solutions for the effort

[31] While non-responsiveness has been observed in various contexts, the specific mechanism we highlight, based on both moral hazard and adverse selection, is novel, to the best of our knowledge. Although Choné and Ma (2011) also achieve a similar result, their mechanism is different, relying on the limited liability constraint in a model without hidden action. Also, an alternative to the pooling contract described here, where both types of physicians participate, would be a "shutdown policy" (see Laffont and Martimort [2002]), where a single contract is designed to satisfy the participation constraint of only one type of physician. However, this option is not practical, as stressed by Currie *et al.* (2024): "chronic doctor shortages in many countries suggest that there will be continuing demand for the services of even the least skilled physicians (p. 36)."

levels), while Proposition 5 holds more generally for any convex effort cost function. We maintain Assumption 2 in the rest of the paper.

In the following sections, we first derive the optimal levels of β_k assuming that physicians choose the optimal testing decision–*i.e.*, that they correctly decide whether to test all, only type-A patients, or nobody. Next, we derive the fixed components $\bar{T}_k$ that induce the optimal testing decisions for the second-best optimal effort levels.

7.2 The optimal second-best pay-for-performance component

We now determine the optimal levels of β_k taking as given the physicians' testing decisions. Since we have shown that the contract is pooling, the remuneration components are identical for all physicians but vary across the different cases.

For a given Case $k = \{0, 1\}$, the social planner's problem is:

$$\max_{\beta_k, \bar{T}_k} SW = \nu[B_k(\varepsilon_{L,k}) - \psi(\varepsilon_{L,k})] + (1-\nu))[B_k(\varepsilon_{H,k}) - \psi(\varepsilon_{H,k})], \tag{27}$$

where $\varepsilon_{i,k}$ is chosen by type i-physicians and satisfies equation (23).

We solve this program in Appendix 9.2 and obtain the following proposition.

Proposition 6 *Under Assumption 2, the second-best slope of the P4P component for the pooling contract in Case 1 and Case 0 is:*

$$\beta_1^{SB} = (\Delta U_A + \gamma \Delta e)[1 - \bar{\alpha}], \tag{28}$$

$$\beta_0^{SB} = (\frac{\Delta U_A + \Delta U_B}{2} + \gamma \Delta e)[1 - \bar{\alpha}], \tag{29}$$

where $\bar{\alpha} = \nu\alpha_L + (1-\nu)\alpha_H$ is the average physician altruism. This second-best optimal contract generates the following ranking of efforts

$$\varepsilon_{L,k}^{SB} < \varepsilon_k^* < \varepsilon_{H,k}^{SB} \ \forall k = \{0, 1\}.$$

As in the first-best, and for similar reasons, the slope of the optimal payment scheme is higher in Case 0 than in Case 1 (*i.e.* $\beta_0^{SB} > \beta_1^{SB}$). But the second-best pooling contract induces

different effort levels, with more altruistic physicians exerting higher effort compared to their less altruistic counterparts. The first-best effort, which remains unaffected by the physician's degree of altruism, takes an intermediate value. This pattern holds in both scenarios: whether no patients are tested or only those with signal-A are tested.

In the next section, we determine the values of the capitation payments.

7.3 The optimal capitation payments

The capitation payments must be set to encourage optimal testing decisions from both types of physicians while ensuring their participation. We start by computing the capitation levels that satisfy the participation constraints while minimizing rents (Section 7.3.1). We then study how to modify those capitation levels to induce the correct testing decisions (Sections 7.3.2 and 7.3.3).

7.3.1 Feasible contracts

When the contract is pooling, there are no incentive constraints, so the set of feasible contracts is reduced to those satisfying the participation constraints.

Definition 2 *The set of transfers of the form (26) satisfying the participation constraints of both type-H and type-L physicians, at the second-best equilibrium while minimizing rents, are denoted by a star and given by:*

$$\begin{aligned}
T^*_{All} &= \bar{T}^*_{All} = 0, \\
T^*_{H,1} &= \psi(\varepsilon^{SB}_{H,1}),\ T^*_{H,0} = \psi(\varepsilon^{SB}_{H,0}), \\
\bar{T}^*_1 &= \psi(\varepsilon^{SB}_{H,1}) - \beta^{SB}_1 \frac{1+\varepsilon^{SB}_{H,1}}{2},\ \bar{T}^*_0 = \psi(\varepsilon^{SB}_{H,0}) - \beta^{SB}_0 \varepsilon^{SB}_{H,0}, \\
T^*_{L,1} &= \psi(\varepsilon^{SB}_{H,1}) - \beta^{SB}_1 \frac{\varepsilon^{SB}_{H,1} - \varepsilon^{SB}_{L,1}}{2},\ T^*_{L,0} = \psi(\varepsilon^{SB}_{H,0}) - \beta^{SB}_0 (\varepsilon^{SB}_{H,0} - \varepsilon^{SB}_{L,0}).
\end{aligned}$$

In Case *All*, no effort is done, and thus the minimum transfer satisfying participation is a capitation of zero for both types of physicians. As shown in the figure in Appendix 9.3, in both Case 1 and Case 0, a type-H physician exerts more effort than a type-L physician (see

Proposition 6). Since the effort cost is convex while the variable part of the reimbursement schedule is linear in effort, the participation constraint is tighter for type-H than for type-L physicians. Minimizing rents then means a zero rent for type-H physicians (with a total transfer equal to her cost of effort, as shown in the second line in Definition 2) while a type-L physician obtains a positive rent. From the zero rent for type-H physicians, one deduces the capitation level (by definition, the same for both physicians' types) in the third line of Definition 2, and the total payment (including rent) of type-L physicians in the last line.

In the next sections, we study how we should modify the capitation levels from the levels computed in Definition 2 in order to induce doctors to take the correct testing decisions. We start with the choice between testing all patients or only those with an A signal. Because the reasoning is very similar for the choice between testing A-signal patients and testing nobody, most of the latter developments are relegated to the appendix.

7.3.2 Implementation of the second-best switching cost, z_{All}

We first compute the equilibrium value of the switching threshold z_{All} when transfers are set as in Definition 2, and compare this equilibrium threshold with its second-best optimal value. We then study how to adjust the capitation parts of the transfers to make sure that doctors take the second-best testing decision, while still satisfying the participation constraints. As we show below, the capitation transfers will be a function of the value of the test cost z. So, unlike for the first-best decentralization, the second-best decentralization requires the test cost (and not only the test decision) to be observable and contractible.

For generic capitation transfers T_{All} and $T_{i,1}$, the equilibrium level of the threshold z_{All} at the second-best level of effort, for a physician with a degree of altruism of α_i, is given by:

$$z^{eq}_{All}(\alpha_i, T_{All}, T_{i,1}) = (1 - \varepsilon^{SB}_{i,1})\,(\Delta U_A + \gamma \Delta e) + \frac{2}{\alpha_i}(T_{All} - T_{i,1}) + \frac{2}{\alpha_i}\psi(\varepsilon^{SB}_{i,1}). \tag{30}$$

Note that the degree of altruism of a doctor influences z^{eq}_{All} both directly and indirectly through her effort choice $\varepsilon^{SB}_{i,1}$. The following lemma allows us to rank these thresholds z^{eq}_{All}, for the transfer levels introduced in Definition 2, according to the physicians' altruism level.

Lemma 2 $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1})$.

Proof. See Appendix 9.4.1. ■

The intuition behind this lemma is that less altruistic physicians are influenced by the opportunity to earn a rent when they test only signal-A patients (rather than receiving no rent when testing everyone). Consequently, they stop testing all patients at a lower threshold of test cost z.

Let us now turn to the second-best optimal level of z_{All}. This level is computed given the doctors' second-best optimal choice of effort, and thus depends on α_i only indirectly through the individual choice of $\varepsilon^{SB}_{i,1}$ (see equation (4)):

$$z^{SB}_{All}(\alpha_i) \equiv (1 - \varepsilon^{SB}_{i,1})(\Delta U_A + \gamma \Delta e) + 2\psi(\varepsilon^{SB}_{i,1}). \tag{31}$$

Recall that this second-best value does not depend on the value of the transfers.

We obtain the following lemma:

Lemma 3 $z^{SB}_{All}(\alpha_i) > z^{eq}_{All}(\alpha_i, T^*_{All}, T^*_{i,1}), i \in \{L, H\}$.

Proof. Comparing expressions (30) and (31), we obtain that the statement holds if and only if

$$\psi(\varepsilon^{SB}_{i,1}) > \frac{1}{\alpha_i}(T_{All} - (T_{i,1} - \psi(\varepsilon^{SB}_{i,1}))),$$

where the right-hand side is proportional to the difference between rents in Case *All* and in Case 1. The right-hand side is negative for L and zero for H when measured at $(T_{All} = T^*_{All}, T_{i,1} = T^*_{i,1})$ so that the above inequality holds for both types of physicians. ■

This reveals that, compared to the second-best optimum, doctors switch too early from testing everyone (Case *All*) to testing only signal-A patients (Case 1) when transfers are set to minimize rents while satisfying participation constraints (*i.e.*, for $T_{All} = T^*_{All}, T_{i,1} = T^*_{i,1}$). This applies even to type-H physicians, who do not earn any rent in either case. The higher second-best effort in Case 1 contributes to the social cost of transitioning from Case *All* to Case 1, but type-H physicians do not internalize this cost since they are compensated for it through

$T^*_{H,1}$. Additionally, type-L physicians face a second incentive to stop treating all patients at a lower-than-second-best-optimal threshold, as the rent they gain in Case 1 is strictly positive while it is nil in Case *All*.

In Appendix 9.4.2, we show that unless we make further assumptions, we cannot order $z^{SB}_{All}(\alpha_i)$ as a function of α_i. In order to avoid treating too many cases and because qualitatively it would not modify our conclusions, we make the following assumption.[32]

Assumption 3 *There is an equal proportion of type-H and type-L physicians:* $\nu = 1/2$.

Lemma 4 *Under Assumption 3,* $z^{SB}_{All}(\alpha_H) = z^{SB}_{All}(\alpha_L)$.

Proof. See Appendix 9.4.2 ■

From now on, to simplify the notation, we will omit α_i as an argument of z^{SB}_{All} when using Assumption 3.

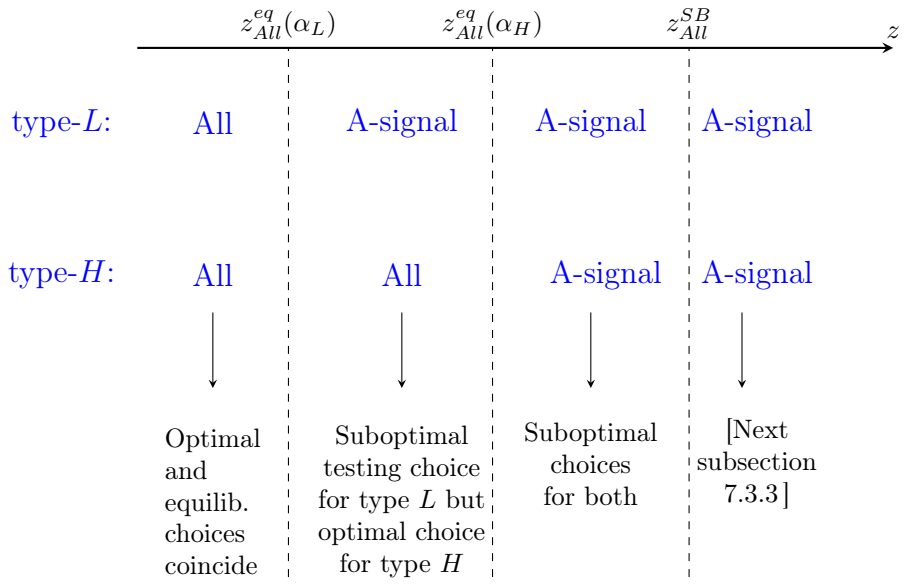
Put together, Lemmas 2, 3 and 4 lead to the following proposition:

Proposition 7 *Under Assumptions 2 and 3, we have:*

$$z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1}) < z^{SB}_{All}.$$

Figure 1 represents the different possible testing decisions as a function of the cost of the test, z. The first two rows show, respectively, the testing decisions of type-L and type-H physicians, while the third row compares equilibrium and optimal decisions.

[32] See Section 7.4 for a discussion of the consequences of lifting this assumption.



Note: For clarity reasons, we have removed the $(T^*_{All}, T^*_{i,1})$ in $z^{eq}_{All}(\alpha_L)$ and $z^{eq}_{All}(\alpha_H)$

Figure 1: Second-best equilibrium and optimal testing choices in Case *All* at $(T^*_{All}, T^*_{i,1})$

Figure 1 shows that there exist values of z for which at least one type of physician takes a sub-optimal testing decision. Unfortunately, it is impossible to find a pooling contract that would equalize the equilibrium and second-best optimum values of z_{All} for both types of physicians simultaneously. However, achieving this alignment is not necessary to induce the correct second-best testing decision for both types of physicians, as long as the (capitation) transfer can be made conditional on the value of z, as we now demonstrate.

For low values of z, below $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1})$, all physicians, whatever their degree of altruism, make the optimal testing decision by choosing to test all patients. In that case, $(T^*_{All}, T^*_{i,1})$ induces the optimal testing decisions.

For intermediate levels of z, either low-altruism physicians (if $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z \leqslant z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1})$) or both types of physicians (if $z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1}) < z \leqslant z^{SB}_{All}$) do not make the optimal testing decision when transfers are set as in Definition 2. More precisely, when $z^{eq}_{All}(\alpha_i, T^*_{All}, T^*_{i,1}) < z \leqslant z^{SB}_{All}$, type-$i$ physicians only test signal-A patients while they should

test all patients. Therefore, the authority must raise the (capitation) level of the transfer received for treating all patients in order to incentivize physicians to test everyone.[33] In doing so, the equilibrium threshold $z^{eq}_{All}(\alpha_i, T_{All}, T^*_{i,1})$ will be adjusted upward until it matches the observed level z.

We denote by $T_{i,All}(z)$ the lowest transfer level that ensures that type-i physicians treat all patients when $T_{i,1}$ is set as in Definition 2, namely such that $z^{eq}_{All}(\alpha_i, T_{i,All}(z), T^*_{i,1}) = z$. We then obtain

$$T_{i,All}(z) = T^*_{i,1} - \psi(\varepsilon^{SB}_{i,1}) + \frac{\alpha_i}{2}[z - (1 - \varepsilon^{SB}_{i,1})(\Delta U_A + \gamma \Delta e)]. \tag{32}$$

Our results are summarized in the following proposition.

Proposition 8 *Assume that the social planner can condition the transfers on the observed level of the cost of the diagnostic test, z. Under Assumptions 2 and 3 and asymmetric information on the α_i's, the rent-minimizing payment received by the physicians in the case where it is optimal to test all agents (*i.e.*, $z < z^{SB}_{All}$) should be set at $T_{i,k}(z) = T^*_{i,k}$, for $k \in \{1, 0\}$ and $i \in \{L, H\}$ together with*

$$T_{All}(z) = \begin{cases} T^*_{All} = 0, & \text{when } z \leqslant z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) \\ T_{L,All}(z), & \text{when } z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z \leqslant z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1}) \\ \max\{T_{L,All}(z), T_{H,All}(z)\}, & \text{when } z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1}) < z \leqslant z^{SB}_{All} \end{cases}$$

where $T_{i,All}(z)$ is defined by (32).

Since the contract is pooling and the transfers cannot be conditioned on α_i, when $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z < z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1})$, the capitation transfer (equal to the total transfer in Case *All*) will be set for every physician to $T_{L,All}(z)$, even if type-H's testing decision is already optimal with the lower T^*_{All}.[34] Finally, to ensure that both types of physicians take the correct testing decisions while minimizing rents, we set the capitation transfer to $T_{All}(z) = \max\{T_{L,All}(z), T_{H,All}(z)\}$ when $z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1}) < z < z^{SB}_{All}$.

We now move to the next cost threshold, z_1.

[33] Alternatively, decreasing the transfer received in Case 1 would violate the participation constraints.

[34] Increasing the capitation transfer above the minimum required level (as defined in Definition 2) in case *All* will only reinforce the decision of a type-H physician to continue testing all patients.

7.3.3 Implementation of the second-best switching cost, z_1

As the reasoning for determining z_1 and T_1 is very similar to that in the previous section, most of the mathematical developments and explanations are provided in Appendix 9.5.

First, we obtain the same ranking of equilibrium and second-best threshold values of test costs as in the previous section.

Proposition 9 *Under Assumptions 2 and 3, we have*

$$z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*) < z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*) < z_1^{SB}.$$

Proof. *See Appendix 9.5.1* ■

Similarly to Proposition 7, the above proposition indicates that, compared to the second-best optimum, doctors switch too readily (*i.e.*, for lower values of z) from testing only signal-A patients (Case 1) to testing no one (Case 0) when transfers are set to minimize rents while satisfying participation constraints. The intuition is the same as for Lemma 3: type-H physicians do not internalize the higher effort cost in Case 0, while type-L physicians are also biased by the higher rents received in the latter case (see Appendix 9.5.1 where we show that type-L physicians effectively earn a higher rent in Case 0 than in Case 1).

We also obtain that less altruistic physicians switch earlier than more altruistic ones to the no-test Case 0. This result is analogous to Lemma 2 from the previous section, and holds for a similar reason: as type-L physicians enjoy a higher rent in Case 0 than in Case 1 when transfers are set at the rent-minimizing levels from Definition 2, they stop testing all patients at a lower test cost z.

Figure 5 in Appendix 9.5 compares equilibrium and optimal testing decisions, and is very similar to Figure 1 in the previous section.

Likewise, the following proposition, which summarizes the optimal capitation payments levels when physicians should test only A-signal patients at the second-best, is similar to Proposition 8.

Proposition 10 *Assume that the government can condition the transfers on the observed level of the cost of the diagnostic test, z. Under Assumptions 2 and 3 and asymmetric information on the α_i's, the rent-minimizing capitation payments received by the physicians in the case where it is optimal to test only signal-A patients (*i.e., *when $z_{All}^{SB} < z < z_1^{SB}$), should be set at $T_{All}(z) = T_{All}^* = 0$, $T_{i,0}(z) = T_{i,0}^*$, $\forall i \in \{L, H\}$ together with*

$$\bar{T}_1(z) = \begin{cases} \bar{T}_1^* = \psi(\varepsilon_{H,1}^{SB}) - \beta_1 \left(\frac{1+\varepsilon_{H,1}^{SB}}{2} \right), & \text{when } z \leqslant z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*) \\ \bar{T}_{L,1}(z), & \text{when } z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*) < z \leqslant z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*) \\ max\{\bar{T}_{L,1}(z), \bar{T}_{H,1}(z)\}, & \text{when } z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*) < z \leqslant z_1^{SB} \end{cases}$$

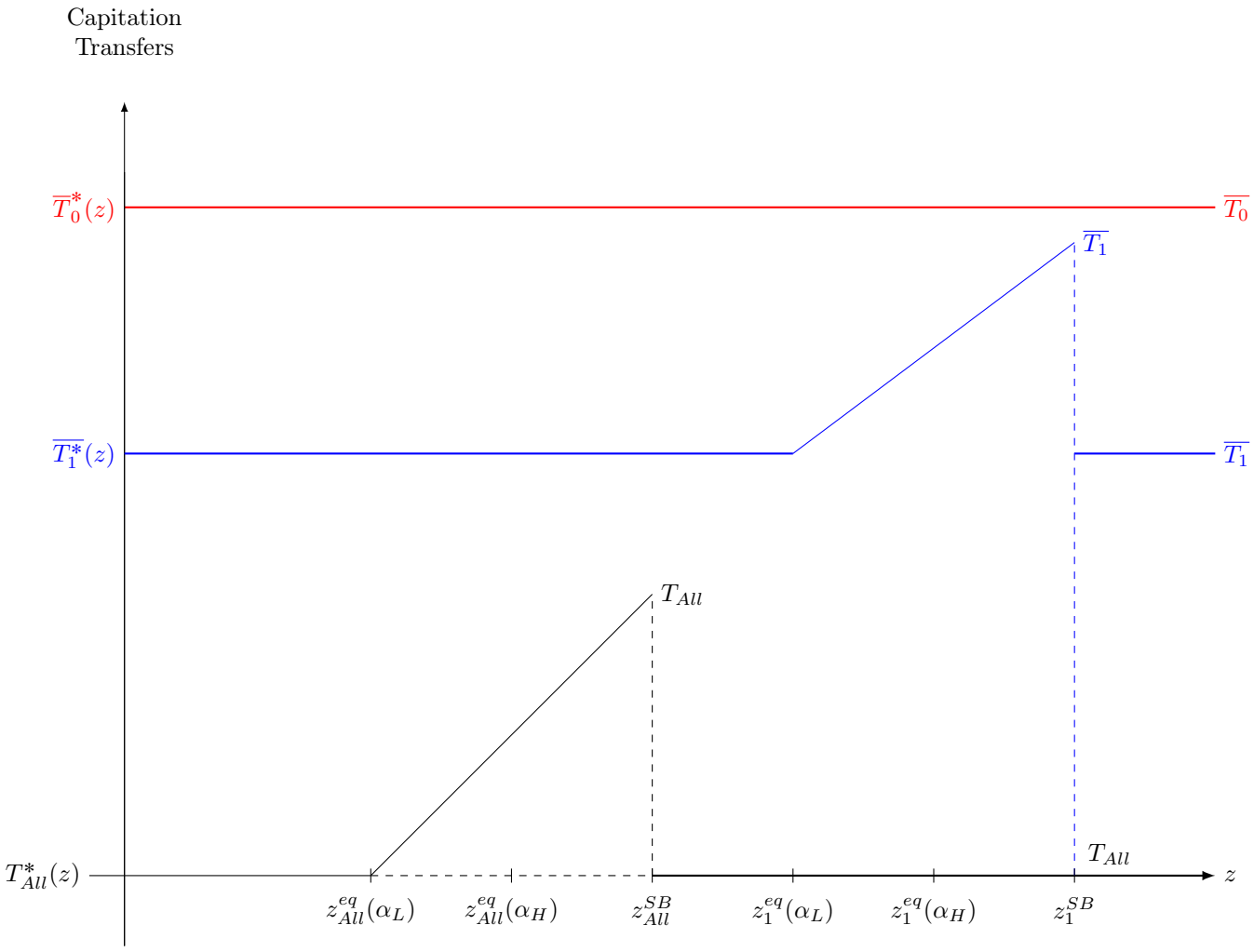
where $\bar{T}_{i,1}(z)$ is defined by (A.11).

Proof. *See Appendix 9.5.2* ■

Finally, when $z > z_1^{SB}$, both physicians make the optimal decision not to test anyone. The optimal set of total transfers is then given by $(T_{All}^*, T_{i,1}^*, T_{i,0}^*)$, set in Definition 2.

7.4 Summing-up the second-best analysis

Figure 2 summarizes what we have learned in Section 7.3, and the take-home message of this paper in terms of capitation levels in the second-best case. The social planner should offer to physicians the choice between three contracts, each one conditional on the fraction of patients tested (all, only signal-A, or none) and on the value of the test cost z. Each contract should satisfy the physicians' participation constraint given the corresponding equilibrium effort choices, while minimizing rents. These capitation levels correspond to the values set in Definition 2. These values need not be distorted provided that all physicians choose the correct second-best testing decision at equilibrium. This is the case, on Figure 2, either if $z < z_{All}^{eq}(\alpha_L, T_{All}^*, T_{L,1}^*)$ (all physicians test all patients, as should be at the second-best), if $z_{All}^{SB} < z < z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*)$ (all physicians test only signal-A patients, as recommended at the second-best), or if $z > z_1^{SB}$ (no physician tests anybody, as required at the second-best). There are two situations where one of the three capitation transfers has to be distorted, namely when at least one type of physician's testing decision departs from the second-best optimal one: either some physicians test only



Note: All $z_k^{eq}(\alpha_i)$ are evaluated at $(T^*_{All}, T^*_{i,1}, T^*_{i,0})$

Figure 2: Levels of the capitation payments

signal-A patients while they should test everyone (if $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z < z^{SB}_{All}$), or they test nobody while they should test signal-A patients (if $z^{eq}_1(\alpha_L, T^*_{L,1}, T^*_{L,0}) < z < z^{SB}_1$). In both cases, the second-best decision can be decentralized by increasing the corresponding capitation (*i.e.*, $T_{All}(z)$ in the former case, and $\bar{T}_1(z)$ in the latter), leaving the other ones unchanged. The crucial insight is that it is not necessary (and, indeed, impossible) to equalize the equilibrium and second-best optimal thresholds of z for both physicians' types simultaneously, but sufficient to find the minimum amount of the capitation transfers that induces both types of physicians to take the second-best optimal testing decision.

Figure 2 is based on two important assumptions. First, we have assumed that $z^{SB}_{All} < z^{eq}_1(\alpha_L, T^*_{L,1}, T^*_{L,0})$. This assumption has allowed us to combine the results of sections 7.3.2 and 7.3.3 into a single figure while covering the largest number of possible situations. If we rather had $z^{SB}_{All} > z^{eq}_1(\alpha_i, T^*_{i,1}, T^*_{i,0})$, we would *not* have the configuration where a type-i physician optimally chooses to treat only signal-A patients. Rather, for $z^{eq}_1(\alpha_i, T^*_{i,1}, T^*_{i,0}) < z < z^{SB}_{All}$, a type-$i$ physician would choose to not test anyone, while it is optimal to test all patients. We would then have to increase T_{All} above its value set in Definition 2, in order to induce these physicians to treat all patients. While the optimal value of T_{All} would differ from the one reported in Proposition 8, the gist of our argument to decentralize the second-best testing decision would then remain unchanged.

The other important premise underlying Figure 2 is that both types of physicians should always make the same testing decision at the second-best (*i.e.*, that the thresholds z^{SB}_{All} and z^{SB}_1 do not vary with α_i). A sufficient condition for this result is that there are as many type-L as type-H physicians (see Lemma 4 and Proposition 9). We show in Appendix 9.6 that the core of our argument remains valid even if this assumption is relaxed. Specifically, the only scenario where it might be impossible to decentralize the second-best optimal testing decision as described in Propositions 8 and 10 is when three conditions are met simultaneously: (i) the optimal testing decision differs between the two types of physicians (*i.e.*, $z^{SB}_k(\alpha_i) < z < z^{SB}_k(\alpha_j)$ for $i \neq j$ and $k \in \{All, 1\}$); (2) at the capitation levels set in Definition 2, one physician type

makes her second-best optimal testing decision while the other type does not; (3) the increase in capitation necessary to change the testing decision of the latter type is large enough to also change the testing decision of the former type. While we cannot rule out such a scenario, the likelihood of all three conditions being met simultaneously is rather limited. Therefore, our conclusion that second-best testing decisions can be decentralized with test cost-dependent capitation levels generally applies in most situations, even when Assumption 3 does not hold.

8 Conclusion

This paper has studied the diagnostic effort and testing decisions of imperfectly altruistic physicians as they determine which of two treatments is more appropriate for their patients.

We first derive the first-best allocation, where the regulation authority can observe the physicians' degree of altruism, and compare it with the *laissez-faire*. In this latter case, both physicians exert insufficient effort due to their imperfect altruism: they accurately anticipate the marginal cost of their effort but underestimate its marginal benefit for their patients. We also show that they rely excessively on diagnostic tests. This occurs because efforts and tests are strategic substitutes, but effort is costly to the physician while the test is costly to the patient.

We then consider the second-best allocation, where physicians' levels of altruism are unobservable. First, we demonstrate that the second-best optimal contract is a *pooling* contract, offering the *same* P4P and capitation components to all physicians. This is a case of non-responsiveness, where the regulation authority's objectives are not aligned with the incentives required to motivate the physicians. In this scenario, the slope of the P4P component of the contract should be based on average altruism, resulting in high-altruism physicians exerting more effort than their low-altruism counterparts. Additionally, we show that if capitation transfers are set at the minimum level required for physicians participation, they will under-utilize diagnostic tests. This occurs because the regulation authority must compensate physicians for their increased diagnostic effort when they reduce testing, leaving rents to all except the most altruistic. Because of the substitutability between diagnostic effort and tests, rents are higher

when effort is higher and fewer patients are tested, leading to under-utilization of tests. This is then corrected by providing larger capitation transfers when more patients are tested.

Interestingly, decentralizing the second-best outcome requires making (capitation) transfers dependent on the test cost. This implies that as new technologies emerge and diagnostic test costs vary over time, the regulation authority could adjust physicians' remunerations accordingly.

Finally, this model relies on several assumptions. First, we have assumed that the test results are perfect, which may not reflect reality. Second, we have assumed that the consequences of mismatches between patient type and treatment are symmetrical, with the health authority observing both types of mismatch. In reality, treating B-patients with P, while not cost-efficient, might not lead to a higher number of physician visits, making it undetectable by the health authority. In that case, the only detectable mismatch would occur when a type-A patient is treated with the default treatment. Exploring these extensions is part of our research agenda.

Our model could also be extended to algorithmic decision-making tools, particularly AI models, which are increasingly employed in clinical settings. While clinical AI tools currently vary in performance (see Obermeyer *et al.*, 2019), ongoing advancements in data quality and machine learning techniques are likely to yield substantial improvements over time. It is essential to understand the impacts of these tools as they evolve, including the need to establish optimal financial incentives for their adoption and application. Ensuring that the economic environment supports effective and beneficial AI usage should be a key focus for future research in healthcare policy and economics.

9 Appendix

9.1 The second-best contract is pooling

The proof of Proposition 5 follows through whatever the Case $k = \{1, 0\}$. We proceed in two stages. We first prove that no menu of contracts with $\beta_{L,k} > \beta_{H,k}$ can satisfy simultaneously the incentive compatibility constraints of both types of physicians. We then show that welfare cannot be maximized with a menu of contracts such that $\beta_{L,k} < \beta_{H,k}$. We then obtain that the same contract with $\beta_{L,k} = \beta_{H,k}$ has to be proposed to both physicians' types at the second-best equilibrium.

9.1.1 No separating contract with $\beta_{L,k} > \beta_{H,k}$

Assume that the two following contracts $(\bar{T}_{L,k}, \beta_{L,k})$ and $(\bar{T}_{H,k}, \beta_{H,k})$ are designed for type L and type H, respectively. Each physician chooses her preferred contract among the two proposed. To specifically incentivize the effort of the low-altruism physician, we need to set $\beta_{L,k} > \beta_{H,k}$. This in turn implies that $\bar{T}_{L,k} < \bar{T}_{H,k}$ since, otherwise, both types would choose the contract devised for L.

Figure 3 illustrates the non-responsiveness argument. Assume that type L is indifferent between the two contracts (*i.e.* points X and Y are on the same indifference curve, denoted by I_L). In this situation, type H would be better off choosing the contract designed for type L (*i.e.* being at point Z on indifference curve I'_H), thereby violating the incentive constraint for type H. To satisfy this incentive constraint, one would need to either increase $\bar{T}_{H,k}$ or decrease $\bar{T}_{L,k}$, which would then violate the incentive constraint for type L and make her strictly prefer the contract designed for type H (since she was initially indifferent between the two contracts). Consequently, at least one incentive constraint is always violated, indicating that a separating equilibrium with $\beta_{L,k} > \beta_{H,k}$ cannot exist.

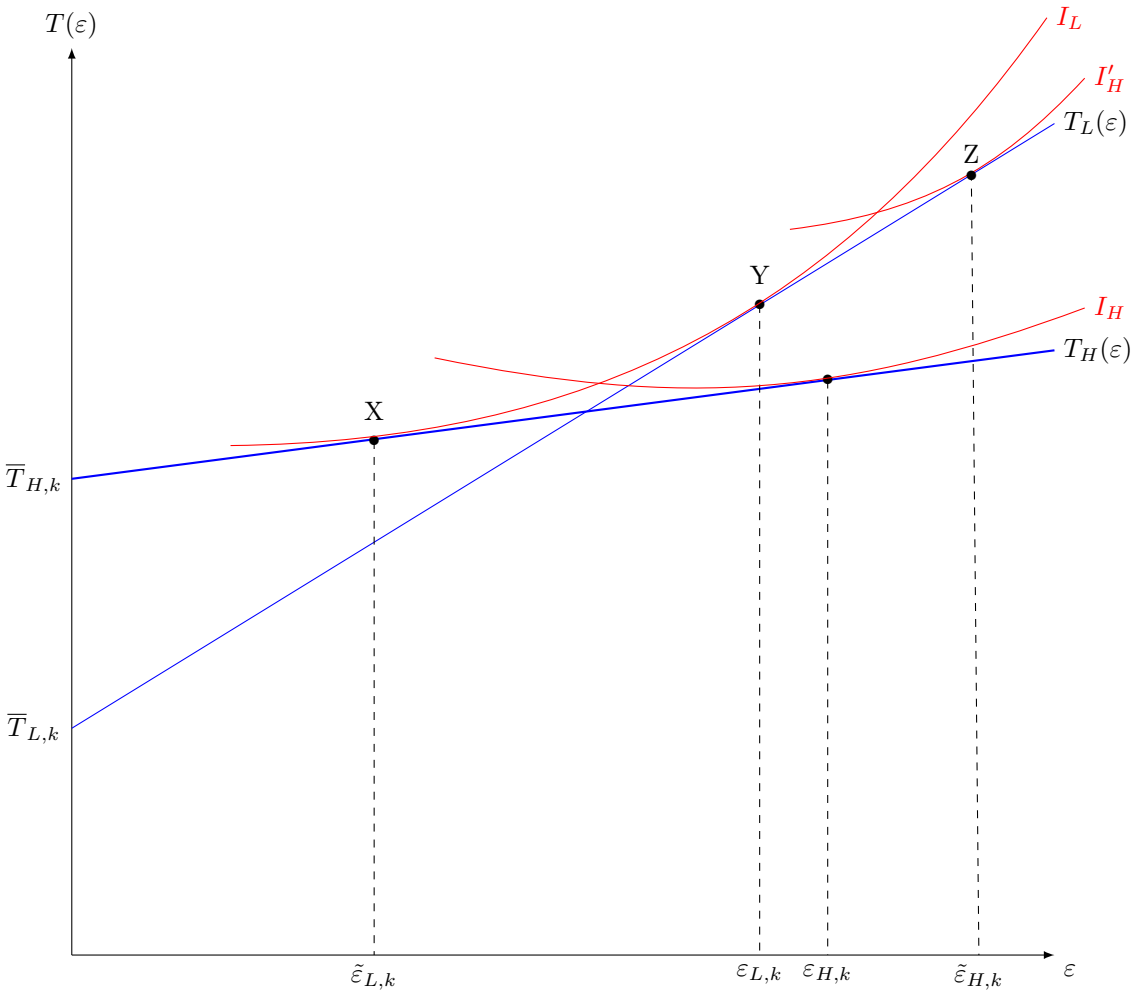


Figure 3: Second-best pooling equilibrium (contradiction argument)

We now turn to a formal proof [FOR ONLINE APPENDIX ONLY].

Assume that with the pair of contracts proposed, a type-L physician is indifferent between the contract $(\bar{T}_{L,k}, \beta_{L,k})$ devised for her and the one devised for type H-physicians, $(\bar{T}_{H,k}, \beta_{H,k})$:

$$\alpha_L B_k(\varepsilon_{L,k}) + \bar{T}_{L,k} + \beta_{L,k} n_{L,k} - \psi(\varepsilon_{L,k}) = \alpha_L B_k(\tilde{\varepsilon}_{L,k}) + \bar{T}_{H,k} + \beta_{H,k} \tilde{n}_{L,k} - \psi(\tilde{\varepsilon}_{L,k}), \quad \text{(A.1)}$$

where we use a tilde to denote the allocation where a physician mimics the other type by buying

the contract designed for the latter. More precisely, $\varepsilon_{L,k}$ is defined by (23) while $\tilde{\varepsilon}_{L,k}$ refers to the level of effort of a physician of type L claiming to be a type H and taking the contract $(\bar{T}_{H,k}, \beta_{H,k})$, such that[35]

$$\psi'(\tilde{\varepsilon}_{i,k}) \quad = \quad \alpha_i \bar{b}_k + \mathbb{1}_k \beta_{j,k}, \tag{A.2}$$

with $i \neq j$, $\bar{b}_k$ defined by (24) and (25), and $\mathbb{1}_1 = 1/2$ and $\mathbb{1}_0 = 1$.

In turn, $\tilde{n}_{L,k}$ refers to the number of correctly-treated patients when a type-L physician claims to be of type H. In Cases 0 and 1, we obtain:

$$\tilde{n}_{i,1} = \frac{1 + \tilde{\varepsilon}_{i,1}}{2}; \ \tilde{n}_{i,0} = \tilde{\varepsilon}_{i,0}.$$

Rearranging equation (A.1), we have that

$$\bar{T}_{H,k} - \bar{T}_{L,k} = \alpha_L [B_k(\varepsilon_{L,k}) - B_k(\tilde{\varepsilon}_{L,k})] - [\psi(\varepsilon_{L,k}) - \psi(\tilde{\varepsilon}_{L,k})] + \beta_{L,k} n_{L,k} - \beta_{H,k} \tilde{n}_{L,k}.$$

Let us then show that with such contracts $(\bar{T}_{L,k}, \beta_{L,k})$ and $(\bar{T}_{H,k}, \beta_{H,k})$, a type-$H$ physician would always want to mimic a type-L. This would be the case if and only if

$$\alpha_H B_k(\tilde{\varepsilon}_{H,k}) + \bar{T}_{L,k} + \beta_{L,k} \tilde{n}_{H,k} - \psi(\tilde{\varepsilon}_{H,k}) > \alpha_H B_k(\varepsilon_{H,k}) + \bar{T}_{H,k} + \beta_{H,k} n_{H,k} - \psi(\varepsilon_{H,k}), \tag{A.3}$$

with $\varepsilon_{H,k}$ and $\tilde{\varepsilon}_{H,k}$ defined by equations (23) and (A.2). This condition can be rewritten as:

$$\alpha_H [B_k(\tilde{\varepsilon}_{H,k}) - B_k(\varepsilon_{H,k})] - [\psi(\tilde{\varepsilon}_{H,k}) - \psi(\varepsilon_{H,k})] + \beta_{L,k} \tilde{n}_{H,k} - \beta_{H,k} n_{H,k} > \bar{T}_{H,k} - \bar{T}_{L,k},$$

and replacing for the expression of $(\bar{T}_{H,k} - \bar{T}_{L,k})$, we get:

$$\alpha_H [B_k(\tilde{\varepsilon}_{H,k}) - B_k(\varepsilon_{H,k})] - [\psi(\tilde{\varepsilon}_{H,k}) - \psi(\varepsilon_{H,k})] - \{\alpha_L [B_k(\varepsilon_{L,k}) - B_k(\tilde{\varepsilon}_{L,k})] - [\psi(\varepsilon_{L,k}) - \psi(\tilde{\varepsilon}_{L,k})]\}$$
$$> \beta_{L,k}(n_{L,k} - \tilde{n}_{H,k}) - \beta_{H,k}(\tilde{n}_{L,k} - n_{H,k}).$$

This condition can further be rewritten as:

$$\alpha_H [\tilde{\varepsilon}_{H,k} \bar{b}_k - \varepsilon_{H,k} \bar{b}_k] - [\psi(\tilde{\varepsilon}_{H,k}) - \psi(\varepsilon_{H,k})] - \{\alpha_L [\varepsilon_{L,k} \bar{b}_k - \tilde{\varepsilon}_{L,k} \bar{b}_k] - [\psi(\varepsilon_{L,k}) - \psi(\tilde{\varepsilon}_{L,k})]\}$$
$$> \beta_{L,k}(n_{L,k} - \tilde{n}_{H,k}) - \beta_{H,k}(\tilde{n}_{L,k} - n_{H,k}). \tag{A.4}$$

[35] Recall that $B'_k(\varepsilon_{i,k})$ is independent of $\varepsilon_{i,k}$.

Isolating $\bar{b}_k$ in equations (23) and (A.2) and replacing for their expression in (A.4), we obtain after some simplifications that it is equivalent to

$$[\tilde{\varepsilon}_{H,k}\psi'(\tilde{\varepsilon}_{H,k}) - \psi(\tilde{\varepsilon}_{H,k})] - [\varepsilon_{H,k}\psi'(\varepsilon_{H,k}) - \psi(\varepsilon_{H,k})]$$
$$-\{[\varepsilon_{L,k}\psi'(\varepsilon_{L,k}) - \psi(\varepsilon_{L,k})] - [\tilde{\varepsilon}_{L,k}\psi'(\tilde{\varepsilon}_{L,k}) - \psi(\tilde{\varepsilon}_{L,k})]\} > 0.$$

We now use the quadratic form of $\psi(.)$ (Assumption 2), which simplifies the above expression as follows:

$$[\tilde{\varepsilon}_{H,k} - \varepsilon_{H,k}][\tilde{\varepsilon}_{H,k} + \varepsilon_{H,k}] > [\varepsilon_{L,k} - \tilde{\varepsilon}_{L,k}][\varepsilon_{L,k} + \tilde{\varepsilon}_{L,k}].$$

Replacing further for the functional form of $\psi(.)$ in the first-order conditions (23) and (A.2), the above condition simplifies to

$$(\beta_{L,k} - \beta_{H,k})(1 + 2\alpha_H\bar{b}_k + \mathbb{1}_k(\beta_{L,k} + \beta_{H,k})) > (\beta_{L,k} - \beta_{H,k})(1 + 2\alpha_L\bar{b}_k + \mathbb{1}_k(\beta_{L,k} + \beta_{H,k})).$$

For $\alpha_H > \alpha_L$ and $\beta_{L,k} > \beta_{H,k}$, the above condition is then always satisfied.

We have then proved that condition (A.3) always holds: a type-H physician would always want to mimic a type-L physician if separating contracts were proposed. This is true for any Case k. So, as soon as a pair of contracts with $\beta_{L,k} > \beta_{H,k}$ makes one individual indifferent between her contract and mimicking the other type, the latter would strictly prefer the contract of the former.

9.1.2 No separating contract with $\beta_{L,k} < \beta_{H,k}$ can maximize welfare

[ONLINE APPENDIX ONLY]

The proof results from two properties of the welfare function: (i) welfare is increasing and concave in effort with a unique maximum (whatever α_i) at $\varepsilon = \varepsilon_k^*$ as given by (2) and (3); (ii) welfare is not affected by payments to doctors which are considered as pure transfers. This means that we can focus on the impact of $\beta_{L,k}$ and $\beta_{H,k}$ on welfare while abstracting from the specific values of $\bar{T}_k$ that satisfy doctors' incentive constraints. Moreover, we know (i) from equation (23) that $\varepsilon_{i,k}$ is monotonically increasing in $\beta_{i,k}$ (for $i \in \{L, H\}$), (ii) from equation

(A.2) that $\tilde{\varepsilon}_{i,k}$ is monotonically increasing in $\beta_{j,k}$ (for $i \neq j$) and (iii) that $\varepsilon_{L,k} < \varepsilon_{H,k}$ when $\beta_{L,k} = \beta_{H,k}$.

Take any $\beta_{L,k} < \beta_{H,k}$. Three situations may then occur:

(i) We have $\varepsilon_{L,k} < \varepsilon_{H,k} \leqslant \varepsilon_k^*$. This also implies that $\varepsilon_{L,k} < \tilde{\varepsilon}_{L,k} < \varepsilon_{H,k} \leqslant \varepsilon_k^*$. In that case, increasing $\beta_{L,k}$ up to $\beta_{H,k}$ increases $\varepsilon_{L,k}$ up to $\tilde{\varepsilon}_{L,k}$ as well as welfare since we have moved the low-altruism physician's effort choice closer to its first-best value.

(ii) We have $\varepsilon_k^* \leqslant \varepsilon_{L,k} < \varepsilon_{H,k}$ which implies that $\varepsilon_k^* \leqslant \varepsilon_{L,k} < \tilde{\varepsilon}_{H,k} < \varepsilon_{H,k}$. In that case, decreasing $\beta_{H,k}$ down to $\beta_{L,k}$ decreases $\varepsilon_{H,k}$ down to $\tilde{\varepsilon}_{H,k}$ and increases welfare since we have moved the high-altruism physician's effort choice closer to its first-best value.

(iii) We have $\varepsilon_{L,k} < \varepsilon_k^* < \varepsilon_{H,k}$. In that case, increasing $\beta_{L,k}$ and decreasing $\beta_{H,k}$ both increase welfare. We should stop increasing $\beta_{L,k}$ once $\varepsilon_{L,k} = \varepsilon_k^*$ or stop decreasing $\beta_{H,k}$ once $\varepsilon_{H,k} = \varepsilon_k^*$. Note that since $\varepsilon_{L,k} < \varepsilon_{H,k}$ for all $\beta_{L,k} = \beta_{H,k}$, at most one of these two situations (where either $\varepsilon_{L,k}$ or $\varepsilon_{H,k}$ equals ε_k^*) can occur. We then have $\beta_{L,k} = \beta_{H,k}$, where both $\varepsilon_{L,k}$ and $\varepsilon_{H,k}$ have been moved closer to their first-best value without overshooting it, so that welfare cannot be maximized when $\beta_{L,k} < \beta_{H,k}$.

9.2 The second-best optimal pay-for-performance transfer

Differentiating (27) with respect to β_k yields the following first-order condition, for each Case k

$$\frac{\partial SW}{\partial \beta_k} = \nu \frac{d\varepsilon_{L,k}}{d\beta_k}[\bar{b}_k - \psi'(\varepsilon_{L,k})] + (1-\nu)\frac{d\varepsilon_{H,k}}{d\beta_k}[\bar{b}_k - \psi'(\varepsilon_{H,k})] = 0,$$

where $d\varepsilon_{i,k}/d\beta_k = \mathbb{1}_k$ is a constant (see equation (23)). Using the quadratic functional form specified for $\psi(.)$ as well as the expression of $\psi'(\varepsilon_{i,k})$ in (23), we obtain that the above condition simplifies to

$$\frac{\partial SW}{\partial \beta_k} = \nu \mathbb{1}_k[\bar{b}_k - (\alpha_L \bar{b}_k + \mathbb{1}_k \beta_k))] + (1-\nu)\mathbb{1}_k[\bar{b}_k - (\alpha_H \bar{b}_k + \mathbb{1}_k \beta_k))] = 0$$

which, after some simplifications, leads to

$$\begin{aligned} \beta_1^{SB} &= 2\bar{b}_1[1-\bar{\alpha}], \\ \beta_0^{SB} &= \bar{b}_0[1-\bar{\alpha}], \end{aligned}$$

and expressions (28) and (29).

We now find the second-best optimal levels of effort by replacing for the values of β_k^{SB} in the first-order condition (23):

$$\psi'(\varepsilon_{i,k}) = (\alpha_i + 1 - \bar{\alpha})\bar{b}_k.$$

Under Assumption 2, we obtain that:

$$\varepsilon_{H,k} = (\alpha_H + 1 - \bar{\alpha})\bar{b}_k, \tag{A.5}$$

$$\varepsilon_{L,k} = (\alpha_L + 1 - \bar{\alpha})\bar{b}_k. \tag{A.6}$$

From equation (23) and comparing it to (2) and (3), we also obtain that ε_k^* is defined by $\psi'(\varepsilon_k^*) = \varepsilon_k^* = \bar{b}_k$. Straightforward algebra then shows that $\varepsilon_{L,k}^{SB} < \varepsilon_k^* < \varepsilon_{H,k}^{SB}$ $\forall k = \{0, 1\}$.

9.3 Second-best rents

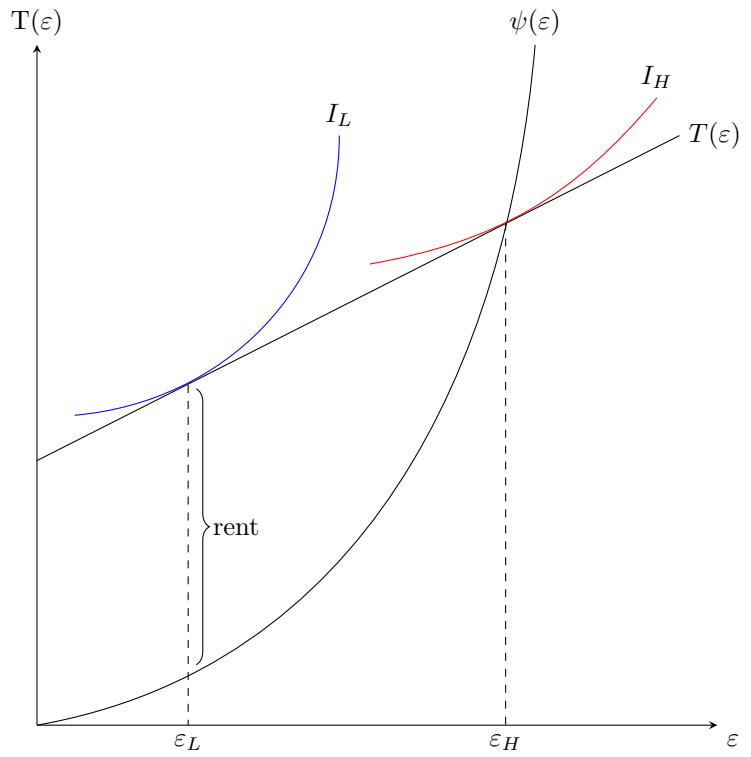


Figure 4: Measuring rents in the second-best allocation

9.4 Implementation of the second-best switching cost, z_{All}

In this appendix, we abuse notation and proceed as if there were a continuum of α levels. This shows that our approach can be generalized to a continuum of altruism types.

9.4.1 Proof of Lemma 2

We differentiate $z^{eq}_{All}(\alpha_i, T_{All}, T_{i,1})$ with respect to α_i:

$$\frac{dz^{eq}_{All}(\alpha_i, T_{All}, T_{i,1})}{d\alpha_i} = \frac{d\varepsilon^{SB}_{i,1}}{d\alpha_i}\left[\frac{2}{\alpha_i}\psi'(\varepsilon^{SB}_{i,1}) - (\Delta U_A + \gamma\Delta e)\right] - \frac{2}{\alpha_i^2}(T_{All} - T_{i,1} + \psi(\varepsilon^{SB}_{i,1})) - \frac{2}{\alpha_i}\frac{dT_{i,1}}{d\alpha_i}, \tag{A.7}$$

with $T_{i,1}$ given by (26), so that

$$\frac{dT_{i,1}}{d\alpha_i} = \frac{\beta_1}{2}\frac{d\varepsilon_{i,1}^{SB}}{d\alpha_i},$$

so that together with the first-order condition (23) on effort, equation (A.7) yields

$$\frac{dz_{All}^{eq}(\alpha_i, T_{All}, T_{i,1})}{d\alpha_i} = \frac{2}{\alpha_i^2}(T_{i,1} - \psi(\varepsilon_{i,1}^{SB}) - T_{All}),$$

which is positive when measured at $(T_{All} = T_{All}^*, T_{L,1} = T_{L,1}^*)$.

9.4.2 Proof of Lemma 4

We differentiate the expression of $z_{All}^{SB}(\alpha_i)$ with respect with α_i, and obtain

$$\begin{aligned}\frac{dz_{All}^{SB}(\alpha_i)}{d\alpha_i} &= [-(\Delta U_A + \gamma\Delta e) + 2\psi'(\varepsilon_{i,1}^{SB})]\frac{d\varepsilon_{i,1}^{SB}}{d\alpha_i} \\ &= 2[\alpha_i - \bar{\alpha}]\frac{\Delta U_A + \gamma\Delta e}{2}\frac{d\varepsilon_{i,1}^{SB}}{d\alpha_i}.\end{aligned}$$

Hence, $z_{All}^{SB}(\alpha_i)$ is a U-shaped function of α_i, with a minimum in $\bar{\alpha}$ so that, unless we make further assumptions, we cannot conclude whether $z_{All}^{SB}(\alpha_H) \gtrless z_{All}^{SB}(\alpha_L)$.

Under Assumption 2, we have that the $\varepsilon_{i,k}$'s are defined by (A.5) and (A.6). Replacing for these values in equation (31) and for the functional form of $\psi(.)$, we obtain, after some simplifications, that

$$z_{All}^{SB}(\alpha_i) = \Delta U_A + \gamma\Delta e + \bar{b}_1^2(1 + \alpha_i - \bar{\alpha})(\alpha_i - \bar{\alpha} - 1),$$

where it is possible to show that under Assumption 3, $(1 + \alpha_i - \bar{\alpha})(\alpha_i - \bar{\alpha} - 1)$ is the same for α_H and α_L. Hence, $z_{All}^{SB}(\alpha_H) = z_{All}^{SB}(\alpha_L)$ as stated in Lemma 4.

9.5 Implementation of the second-best switching cost, z_1

9.5.1 Proof of Proposition 9

We have:

$$\begin{aligned}z_1^{eq}(\alpha_i, T_{i,1}, T_{i,0}) &= (1 - \varepsilon_{i,0}^{SB})\Delta U_B - (\varepsilon_{i,0}^{SB} - \varepsilon_{i,1}^{SB})\Delta U_A + \gamma(1 + \varepsilon_{i,1}^{SB} - 2\varepsilon_{i,0}^{SB})\Delta e \\ &+ \frac{2}{\alpha_i}(T_{i,1} - T_{i,0}) + \frac{2}{\alpha_i}(\psi(\varepsilon_{i,0}^{SB}) - \psi(\varepsilon_{i,1}^{SB})). \qquad \text{(A.8)}\end{aligned}$$

Differentiating equation (A.8) with respect to α_i and using the envelope theorem for $\varepsilon_{i,k}^{SB}$ (*i.e.* equation (23)), as well as $T_{i,1}$ given by (26) with $\beta_{i,k} = \beta_1$, so that

$$\frac{dT_{i,1}}{d\alpha_i} = \frac{\beta_1}{2}\frac{d\varepsilon_{i,1}^{SB}}{d\alpha_i},$$
$$\frac{dT_{i,0}}{d\alpha_i} = \beta_0\frac{d\varepsilon_{i,0}^{SB}}{d\alpha_i},$$

we obtain after some rearrangements,

$$\frac{dz_1^{eq}(\alpha_i, T_{i,1}, T_{i,0})}{d\alpha_i} = -\frac{2}{\alpha_i^2}[(T_{i,1} - \psi(\varepsilon_{i,1}^{SB})) - (T_{i,0} - \psi(\varepsilon_{i,0}^{SB}))]. \tag{A.9}$$

In order to sign this expression, we define the square bracket in the right-hand side of equation (A.9) as the difference in rents between Cases 1 and 0,

$$\begin{aligned}\Delta R &\equiv (T_{i,1} - \psi(\varepsilon_{i,1}^{SB})) - (T_{i,0} - \psi(\varepsilon_{i,0}^{SB})) \\ &\equiv \bar{T}_1 - \bar{T}_0 + \beta_1\frac{1+\varepsilon_{i,1}^{SB}}{2} - \psi(\varepsilon_{i,1}^{SB}) - (\beta_0\varepsilon_{i,0}^{SB} - \psi(\varepsilon_{i,0}^{SB}))\end{aligned}$$

and we differentiate it with respect to α_i:

$$\begin{aligned}\frac{d\Delta R}{d\alpha_i} &= \alpha_i\bar{b}_0\frac{d\varepsilon_{i,0}^{SB}}{d\alpha_i} - \alpha_i\bar{b}_1\frac{d\varepsilon_{i,0}^{SB}}{d\alpha_i} \\ &= -\alpha_i(\bar{b}_1^2 - \bar{b}_0^2) > 0,\end{aligned}$$

where the second line is obtained using equation (23) together with Assumption 2, and $d\varepsilon_{i,k}^{SB}/d\alpha_i = \bar{b}_k$.

When measured at $(\alpha_H, T_{H,1}^*, T_{H,0}^*)$ (so that the participation constraints of a type-H physician are binding in both cases), we have $\Delta R = 0$, leading to $\Delta R < 0$ when measured at $(\alpha_L, T_{L,1}^*, T_{L,0}^*)$. This implies that for type-L physicians, the rents (as set in Definition 2) are increasing when transitioning from Case 1 to Case 0.

This in turn means that $z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*) < z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*)$, so that the analogous of Lemma 2 holds for z_1^{eq} as well.

We then turn to the optimal second-best switching cost $z_1^{SB}(\alpha_i)$, given the second-best

optimal level of effort, defined by:

$$z_1^{SB}(\alpha_i) \equiv 2\left(\psi(\varepsilon_{i,0}^{SB}) - \psi(\varepsilon_{i,1}^{SB})\right) + \left(1 - \varepsilon_{i,0}^{SB}\right)\Delta U_B - \left(\varepsilon_{i,0}^{SB} - \varepsilon_{i,1}^{SB}\right)\Delta U_A + \gamma(1 + \varepsilon_{i,1}^{SB} - 2\varepsilon_{i,0}^{SB})\Delta e. \tag{A.10}$$

Comparing equations (A.8) with (A.10), we obtain, after some rearrangements, that $z_1^{eq}(\alpha_i, T_{i,1}^*, T_{i,0}^*) < z_1^{SB}(\alpha_i)\,\forall i$ if $\psi(\varepsilon_{i,0}^{SB})) - \psi(\varepsilon_{i,1}^{SB})) > \Delta R/\alpha_i\,\forall i$. Since $\Delta R \leqslant 0$ and $\varepsilon_{i,0}^{SB} > \varepsilon_{i,1}^{SB}\,\forall i$, this inequality is satisfied at $(\alpha_i, T_{i,1}^*, T_{i,0}^*)$, so that the analogous of Lemma 3 holds for the comparison between $z_1^{SB}(\alpha_i)$ and $z_1^{eq}(\alpha_i, T_{i,1}^*, T_{i,0}^*)$ as well.

Using the expressions (A.5) and (A.6) of $\varepsilon_{i,k}$ under Assumptions 2 and 3, it is possible to show that

$$z_1^{SB}(\alpha_i) = (\Delta U_A + \gamma\Delta e) + (\bar{b}_0^2 - \bar{b}_1^2)(1 + \alpha_i - \bar{\alpha})(\alpha_i - \bar{\alpha} - 1).$$

Since the last term is invariant to α_i under Assumption 3, we obtain that $z_1^{SB}(\alpha_H) = z_1^{SB}(\alpha_L)$, so that the analogous of Lemma 4 holds for $z_1^{SB}(\alpha_i)$ as well. For simplicity, in the rest of the manuscript, we remove the argument in $z_1^{SB}(\alpha_i)$.

This completes the proof of Proposition 9.

9.5.2 Proof of Proposition 10

We now determine the levels of the capitation payments $\bar{T}_1$ in Case 1 which induce the optimal testing decision. As shown in Figure 5, four cases are possible depending on the value of z.

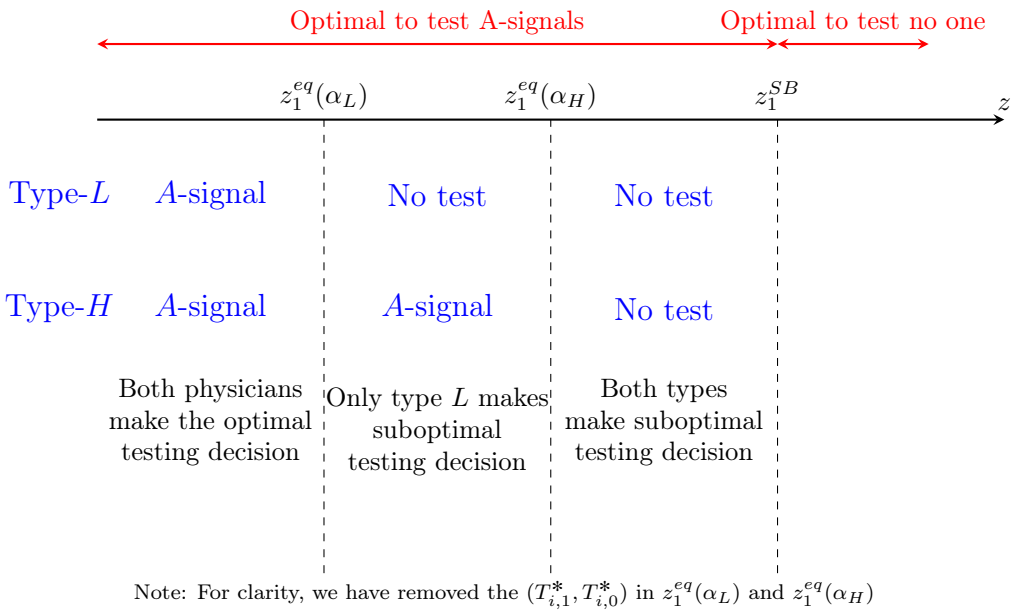


Figure 5: Second-best equilibrium and optimal testing choices in Case 1 at $(T_{i,1}^*, T_{i,0}^*)$.

When z is below $z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*)$, both physicians make the optimal testing decision, since z is also below z_1^{SB} for both. In that situation, there is no need to distort their testing choices and the payment scheme only needs to be set so as to satisfy the participation constraints of both type-H and type-L physicians. In that case, transfers $T_{i,1}(z)$ are given by Definition 2.

For intermediate levels of z, such that $z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*) < z < z_1^{SB}$, either only type-$L$ physicians (when $z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*) < z \leqslant z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*)$) or both types (when $z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*) < z < z_1^{SB}$) make sub-optimal testing decisions. In these intervals, the planner then has to distort the physicians' testing decision threshold $z_1^{eq}(\alpha_i, T_{i,1}^*, T_{i,0}^*)$ to (at least) the observed cost z by increasing the capitation transfer $\bar{T}_1$. We define $T_{i,1}(z)$ as the minimum (total) transfer level ensuring that, for type i, $z_1^{eq}(\alpha_i, T_{i,1}(z), T_{i,0}^*) = z$:

$$T_{i,1}(z) = T_{i,0}^* - \psi(\varepsilon_{i,0}^{SB}) + \psi(\varepsilon_{i,1}^{SB}) + \frac{\alpha_i}{2}\left[z - ((1-\varepsilon_{i,0}^{SB})\Delta U_B - (\varepsilon_{i,0}^{SB} - \varepsilon_{i,1}^{SB})\Delta U_A + \gamma(1 + \varepsilon_{i,1}^{SB} - 2\varepsilon_{i,0}^{SB})\Delta e)\right].$$

To provide physician i with this total transfer, the capitation payment then has to be set at the

following level:

$$\begin{aligned}\bar{T}_{i,1}(z) &\equiv T_{i,1}(z) - \beta_1^{SB}\frac{1+\varepsilon_{i,1}^{SB}}{2} \\ &= T_{i,0}^* - \psi(\varepsilon_{i,0}^{SB}) + \psi(\varepsilon_{i,1}^{SB}) - \beta_1^{SB}\frac{1+\varepsilon_{i,1}^{SB}}{2} \\ &+ \frac{\alpha_i}{2}\big[z - ((1-\varepsilon_{i,0}^{SB})\Delta U_B - (\varepsilon_{i,0}^{SB} - \varepsilon_{i,1}^{SB})\Delta U_A + \gamma(1+\varepsilon_{i,1}^{SB} - 2\varepsilon_{i,0}^{SB})\Delta e)\big], \quad \text{(A.11)}\end{aligned}$$

Since the contract is pooling, the capitation transfer has to be set at the same level for both physicians' types. When $z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*) < z < z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*)$, the transfer then has to be set at $\bar{T}_{L,1}(z)$ for all physicians (even though type-H physicians already make the optimal testing decision when offered the capitation transfers set in Definition 2). When $z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*) < z < z_1^{SB}$, the transfer then has to be set at the maximum between $\bar{T}_{L,1}(z)$ and $\bar{T}_{H,1}(z)$ to ensure the optimal test decision while minimizing rents. We then have proved Proposition 10.

9.6 Second-Best optimal testing decisions varying with altruism degree [ONLINE APPENDIX]

In this Appendix, we study the decentralization of the second-best optimal decisions when Assumption 3 does not hold. This implies that Lemma 4 as well as $z_1^{SB}(\alpha_H) = z_1^{SB}(\alpha_L)$ also do not hold anymore.

In this appendix, we focus on the choice between treating all patients or only those with an A signal (*i.e.*, on the threshold $z_{All}^{SB}(\alpha_i)$), but a similar analysis applies to the choice between treating only signal-A patients or nobody (*i.e.*, on the threshold $z_1^{SB}(\alpha_i)$).

Since Lemmas 2 and 3 still hold while Lemma 4 does not, there are three novel potential rankings of equilibrium and second-best optimal testing thresholds not covered in the text: (1) $z_{All}^{eq}(\alpha_L, T_{All}^*, T_{L,1}^*) < z_{All}^{eq}(\alpha_H, T_{All}^*, T_{L,1}^*) < z_{All}^{SB}(\alpha_L) < z_{All}^{SB}(\alpha_H)$, (2) $z_{All}^{eq}(\alpha_L, T_{All}^*, T_{L,1}^*) < z_{All}^{SB}(\alpha_L) < z_{All}^{eq}(\alpha_H, T_{All}^*, T_{L,1}^*) < z_{All}^{SB}(\alpha_H)$ and (3) $z_{All}^{eq}(\alpha_L, T_{All}^*, T_{L,1}^*) < z_{All}^{eq}(\alpha_H, T_{All}^*, T_{L,1}^*) < z_{All}^{SB}(\alpha_H) < z_{All}^{SB}(\alpha_L)$.

The implementation procedure for determining the optimal levels of the capitation transfers, as outlined in the text, can be applied to all configurations, except if z is located simultaneously

between the two second-best thresholds (so that the optimal testing decision varies between physicians types), and between the equilibrium and second-best threshold of the physician type who has the larger second-best threshold. This can occur in each of the three different novel rankings just defined, provided that: (1) $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{L,1}) < z^{SB}_{All}(\alpha_L) < z < z^{SB}_{All}(\alpha_H)$, (2) $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z^{SB}_{All}(\alpha_L) < z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{L,1}) < z < z^{SB}_{All}(\alpha_H)$ and (3) $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{L,1}) < z^{SB}_{All}(\alpha_H) < z < z^{SB}_{All}(\alpha_L)$.

In these three instances, the second-best optimal testing decision can only be decentralized by increasing the capitation level T_{All} for either type H (in the first and second instances) or type L (in the last instance). If this increase is substantial enough, we then run the risk of changing as well the testing decision of the other physician's type, moving her away from her second-best optimal decision (to test only A-signal). Except in these very specific circumstances, the planner can still decentralize the second-best testing decision as described in the text when $z^{SB}_{All}(\alpha_i)$ varies with physicians' types.

9.7 Unequal proportions of type A and B patients. [ONLINE APPENDIX]

In this section, we assume a generic proportion λ (resp. $(1-\lambda)$) of patients of type A (resp. type B). We keep the assumption that the signal precision is equal to ε for both patients' types (as in Garcia-Mariñoso and Jelovac [2003], for instance). Table 1 reflects the result of Bayesian updating by doctors and is now modified as follows:

Type → Signal ↓	B	A	Total
B	$(1-\lambda)\varepsilon$	$\lambda(1-\varepsilon)$ false neg.	$\lambda + \varepsilon(1-2\lambda)$
A	$(1-\lambda)(1-\varepsilon)$ false pos.	$\lambda\varepsilon$	$(1-\lambda) - \varepsilon(1-2\lambda)$
	$(1-\lambda)$	λ	1

Note that the proportion of patients with a signal reflecting their type remains ε (and thus those with an incorrect signal represent the complementary fraction $1-\varepsilon$) independently of the value of λ. The fraction of signal-i type is now a function of ε (last column). More precisely, when $\lambda \neq 1/2$, there is an over-representation (resp. under-representation) of the signal corresponding to

the minority (resp. majority) type, with the gap between signal- and type-frequency decreasing with ε and disappearing when $\varepsilon = 1$. In our context of precision medicine, it is assumed that $\lambda < 1/2$, as more people should be treated with the default treatment D than with the personalised one.

9.7.1 Optimal effort levels

We first compute the social optimum. We proceed exactly as in the paper.

Case All: Test all patients. In such a case, welfare is given by

$$W_{All}(\varepsilon_{All}) = -\psi(\varepsilon_{All}) + \lambda U_A^P + (1-\lambda)U_B^D - z - \gamma e^M,$$

because true types are revealed after the test and there is a proportion $(1-\lambda)$ of type B and a proportion λ of type A. As before, effort is then useless (*i.e.* $\varepsilon^*_{All} = 1/2$ and $\psi(\varepsilon^*_{All}) = 0$), because it is costly to exert, while the test anyway will reveal the patient's type with certainty.

Case 0: No test is prescribed to anyone.

In such a case, the welfare function is:

$$W_0(\varepsilon_0) = -\psi(\varepsilon_0)+(1-\lambda)\varepsilon_0 U_B^D+\lambda(1-\varepsilon_0)U_A^D+(1-\lambda)(1-\varepsilon_0)U_B^P+\lambda\varepsilon_0 U_A^P-\gamma(\varepsilon_0 e^M+(1-\varepsilon_0)e^{NM}),$$

where the third and fourth terms come from the classification errors: the false negatives (A types treated with D because mistaken for types B) and false positives (B types treated with P because mistaken for types A). The first-order condition for ε_0 (optimal effort in the absence of a diagnostic test) is:

$$\psi'(\varepsilon_0^*) = (1-\lambda)\Delta U_B + \lambda\Delta U_A + \gamma\Delta e. \tag{A.12}$$

The intuition for the first two terms is that a marginal increase in effort decreases by $1-\lambda$ the proportion of false positives (with a per person gain of ΔU_B) and by λ the false negatives (with a per person gain of ΔU_A). The intuition for the last term is that we forgo Δe visits each time the doctor makes more effort ($(1-\lambda)$ of type B and λ of type A).

Note that we exclude the possibility that a solution where no effort is made and everyone is treated with D is preferred to a solution with no test and exerting ε_0^*. This is equivalent to assuming that ε_0^* satisfies

$$W_0(\varepsilon_0^*) > (1-\lambda)U_B^D + \lambda U_A^D - \gamma((1-\lambda)e^M + \lambda e^{NM}).$$

Case 1: Test prescribed (after effort ε chosen) on signal A only

When the test is prescribed only after observing a signal A, the welfare function becomes

$$\begin{aligned} W_1(\varepsilon_1) &= -\psi(\varepsilon_1) + (1-\lambda)U_B^D + \lambda[\varepsilon_1 U_A^P + (1-\varepsilon_1)U_A^D] - z((1-\lambda)(1-\varepsilon_1) + \varepsilon_1\lambda) \\ &- \gamma[(1-\lambda+\lambda\varepsilon_1)e^M + (1-\varepsilon_1)\lambda e^{NM}], \end{aligned}$$

where ε_1 denotes the effort level in this case, and where the test allows to get rid of the false positives (B types who sent a A signal) at the test cost z for proportion $((1-\lambda)(1-\varepsilon_1) + \varepsilon_1\lambda)$ of the sample that has sent a A signal. In such a case, the initial false positives receive the D treatment.

The first-order condition for ε_1 is

$$\psi'(\varepsilon_1^*) = \lambda[\Delta U_A + \gamma\Delta e] + z(1-2\lambda). \tag{A.13}$$

Note that, unlike our initial formulation (3), ε_1^* now depends on z. As explained above, a greater effort decreases the fraction of type-A signal when $\lambda < 1/2$, and thus the fraction to be tested. This provides an additional reason to exert effort in Case 1, compared to the situation where $\lambda = 1/2$. We highlight this point in footnote 19 in the text.

Proposition A.1 *Effort and test are strategic substitutes:*
$\varepsilon_{All}^* < \varepsilon_1^* < \varepsilon_0^* \iff z(1-2\lambda) < (1-\lambda)(\Delta U_B + \gamma\Delta e)$.

Proof. This result is obtained by comparing first-order conditions (A.12) and (A.13). ■

Proposition A.1 generalizes Proposition 1. Intuitively, the ranking of optimal effort across cases remains the same, provided that the additional incentive to exert effort in Case 1 (to

decrease the fraction of patients to be tested) is small enough. This will be the case provided that λ is not too small or z not too large.

9.7.2 Optimal diagnostic test decisions

First, we show that $W_{All}(1/2)$ and $W_1(\varepsilon_1^*(z))$ intersect (at most) only once as z increases from zero. As in the text, the slope of the derivative of $W_{All}(1/2)$ with respect to z is -1. Using the envelope theorem, the slope of $W_1(\varepsilon_1^*(z))$ with respect to z is,

$$\frac{\partial W_1(\varepsilon_1^*(z))}{\partial z} = -[(1-\lambda)(1-\varepsilon_1^*) + \varepsilon_1^*\lambda],$$

which is negative, with an absolute value corresponding to the total fraction of signals A received, and belonging to the interval $[1-\varepsilon_1^*(z), \varepsilon_1^*(z)]$, and thus lower than 1. As the proportion of signals A increases with λ, the slope of $W_1(\varepsilon_1^*(z))$ with respect to z increases (in absolute value) with λ.

Assuming as in the text that it is optimal to test all if the test cost is nil (*i.e.*, that $W_{All}(1/2) > W_1(\varepsilon_1^*(z))$ when $z = 0$), there is a single value of z which equalizes the two, and which is such that

$$z_{All}^* \equiv \frac{\lambda(1-\varepsilon_1^*(z_{All}^*))(\Delta U_A + \gamma\Delta e) + \psi(\varepsilon_1^*(z_{All}^*))}{\lambda + (1-2\lambda)\varepsilon_1^*(z_{All}^*)},$$

where the denominator is positive when $\lambda < 1/2$.

We then move to the test cost threshold level which renders the planner indifferent between treating only those signalling A and not treating anyone, $W_1(\varepsilon_1^*(z)) = W_0(\varepsilon_0^*)$, and obtain that:

$$z_1^* \equiv \frac{(\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*(z_1^*)) + (1-\lambda)(1-\varepsilon_0^*)\Delta U_B - \lambda(\varepsilon_0^* - \varepsilon_1^*(z_1^*))\Delta U_A + \gamma[\lambda\varepsilon_1^*(z_1^*) - \varepsilon_0^* + (1-\lambda)]\Delta e}{1-\lambda+\varepsilon_1^*(z_1^*)(2\lambda-1)}.$$

The intuition behind this formulation is similar to the one provided in Section 4.2, with only the proportions in front of the different terms in both the numerator and the denominator changing.

9.7.3 The physician's problem

The doctor's utility now writes:

$$\begin{aligned}
V_{All} &= \alpha[\lambda U_A^P + (1-\lambda)U_B^D - z - \gamma e^M] + T_{All}(.) - \psi(\varepsilon_{All}), \\
V_1 &= \alpha\{(1-\lambda)U_B^D + \lambda[\varepsilon_1 U_A^P + (1-\varepsilon_1)U_A^D] - z((1-\lambda)(1-\varepsilon_1) + \lambda\varepsilon_1) \\
&\quad -\gamma[(1-\lambda+\lambda\varepsilon_1)e^M + (1-\varepsilon_1)\lambda e^{NM}]\} \\
&+ T_1(.) - \psi(\varepsilon_1), \\
V_0 &= \alpha\{(1-\lambda)\varepsilon_0 U_B^D + \lambda(1-\varepsilon_0)U_A^D + (1-\lambda)(1-\varepsilon_0)U_B^P + \lambda\varepsilon_0 U_A^P - \gamma(\varepsilon_0 e^M + (1-\varepsilon_0)e^{NM})\} \\
&+ T_0(.) - \psi(\varepsilon_0).
\end{aligned}$$

We obtain the following (equilibrium) levels of efforts:

$$\begin{aligned}
\psi'(\varepsilon_{All}^{eq}) &= T'_{All}(\varepsilon_{All}^{eq}), \\
\psi'(\varepsilon_0^{eq}) &= \alpha\{(1-\lambda)\Delta U_B + \lambda\Delta U_A + \gamma\Delta e\} + T'_0(\varepsilon_0^{eq}), \\
\psi'(\varepsilon_1^{eq}) &= \alpha\{\lambda[\Delta U_A + \gamma\Delta e] + z(1-2\lambda)\} + T'_1(\varepsilon_1^{eq}).
\end{aligned}$$

Like the optimal level of effort, ε_1^{eq} now also depends on z. Just as in our baseline case with equal proportions of type A and type B patients, we find that, under the *laissez-faire* scenario, imperfectly altruistic physicians under-provide effort in Cases 0 and 1, and that effort increases with altruism.

We finally compute the equilibrium partition of whether to test or not, namely the thresholds z_{All}^{eq} and z_1^{eq}. The threshold z_{All}^{eq} is such that $V_{All}(z_{All}^{eq}) = V_1(z_{All}^{eq})$, so that

$$\begin{aligned}
z_{All}^{eq} &= \frac{\lambda(1-\varepsilon_1^{eq}(z_{All}^{eq}))(\Delta U_A + \gamma\Delta e)}{\lambda + (1-2\lambda)\varepsilon_1^{eq}(z_{All}^{eq})} + \frac{T_{All} - T_1}{\alpha(\lambda + (1-2\lambda)\varepsilon_1^{eq}(z_{All}^{eq}))} \\
&+ \frac{\psi(\varepsilon_1^{eq}(z_{All}^{eq}))}{\alpha(\lambda + (1-2\lambda)\varepsilon_1^{eq}(z_{All}^{eq}))}.
\end{aligned}$$

This expression is similar to equation (9), except that the weights in front of the different terms in the numerator and the denominator are now different from 1/2 and involve ε_1^{eq}, which is itself measured at z_{All}^{eq}.

We proceed in the same way for z_1^{eq}, which is such that $V_0(z_1^{eq}) = V_1(z_1^{eq})$ and we obtain that:

$$\begin{aligned} z_1^{eq} &= \frac{(1-\lambda)\,(1-\varepsilon_0^{eq})\,\Delta U_B - \lambda\,(\varepsilon_0^{eq} - \varepsilon_1^{eq}(z_1^{eq}))\,\Delta U_A + \gamma[\lambda\varepsilon_1^{eq}(z_1^{eq}) - \varepsilon_0^{eq} + (1-\lambda)]\Delta e}{1-\lambda+\varepsilon_1^{eq}(z_1^{eq})(2\lambda-1)} \\ &+ \frac{T_1 - T_0}{\alpha[(1-\lambda)+\varepsilon_1^{eq}(z_1^{eq})(2\lambda-1)]} + \frac{\psi(\varepsilon_0^{eq}) - \psi(\varepsilon_1^{eq}(z_1^{eq}))}{\alpha[(1-\lambda)+\varepsilon_1^{eq}(z_1^{eq})(2\lambda-1)]}. \end{aligned}$$

This threshold level is very similar to equation (10). However, the right-hand-side of the above expression includes ε_1^{eq} now measured at the threshold level z_1^{eq}.

In contrast to our baseline scenario where the proportions of type A and type B are equal, it is now more challenging to directly compare the equilibrium levels of z with the optimal ones. This is because both at the equilibrium and at the optimum, we can only establish a system of two equations and two unknowns (specifically, ε_1 is determined by z while both thresholds z_{All} and z_1 are determined in turn by ε_1). We summarize this point in footnote 21.

9.8 Proof of Proposition 4

We first have to find the values of the fixed (*i.e.*, capitation) component of the payment scheme that align incentives for the optimal testing decisions (*i.e.*, that $z_{All}^{eq} = z_{All}^*$ and $z_1^{eq} = z_1^*$), and then check that the doctors' participation constraints are satisfied in all 3 cases.

Combining equations (15) and (20) allows us to obtain the optimal capitation value in Case 1, namely

$$\bar{T}_1 = \bar{T}_{All} + (1-\alpha)[\psi(\varepsilon_1^*) - (\Delta U_A + \gamma\Delta e)n_1^*], \tag{A.14}$$

where $n_1^* = (1+\varepsilon_1^*)/2$ is the number of correctly treated patient in Case 1 when the optimum effort level is exerted.

Using equations (16) and (19), we obtain

$$T_1(n_1^*) + (1-\alpha)[\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*)] = \bar{T}_0 + (1-\alpha)[(\frac{\Delta U_B + \Delta U_A}{2} + \gamma\Delta e)n_0^*].$$

Using equation (20) and (A.14) then allows us to find the value of the optimal capitation level in Case 0:

$$\bar{T}_0 = \bar{T}_{All} + (1-\alpha)[\psi(\varepsilon_0^*) - (\frac{\Delta U_B + \Delta U_A}{2} + \gamma\Delta e)n_0^*], \tag{A.15}$$

where $n_0^* = \varepsilon_0^*$ is the number of correctly treated patient in Case 0 when the optimal effort level is exerted.

Observe that using the first-order conditions for the optimum (2) and (3), the bracket term in both equations (A.14) and (A.15) can be rewritten as $\psi(\varepsilon_1^*) - \psi'(\varepsilon_1^*)(1+\varepsilon_1^*)$ and $\psi(\varepsilon_0^*) - \psi'(\varepsilon_0^*)\varepsilon_0^*$. Since the effort cost is increasing and convex (implying that $\psi(\varepsilon) < \psi'(\varepsilon)\varepsilon$), we obtain that both brackets are negative, so that both $\bar{T}_0$ and $\bar{T}_1$ are lower than $\bar{T}_{All}$.

We then have to set $\bar{T}_{All}$ such as the participation constraints are satisfied in all 3 cases. This corresponds to ensuring that

$$\begin{aligned} \bar{T}_{All} &\geqslant \psi(\varepsilon_{All}^*) = 0, \\ T_1 &\geqslant \psi(\varepsilon_1^*), \\ T_0 &\geqslant \psi(\varepsilon_0^*). \end{aligned}$$

Replacing for eq. (19), (20), (A.14) and (A.15) in the above expressions, we obtain after some rearrangements that $\bar{T}_{All} = \max\{0, \alpha\psi(\varepsilon_0^*), \alpha\psi(\varepsilon_1^*)\} = \alpha\psi(\varepsilon_0^*)$ since $\varepsilon_0^* > \varepsilon_1^*$.

From this, we are able to obtain the capitation part of the payment scheme doctors receive in Cases $\{0, 1\}$:

$$\begin{aligned} \bar{T}_0 &= \psi(\varepsilon_0^*) - (1-\alpha)[\frac{\Delta U_B + \Delta U_A}{2} + \gamma\Delta e)]n_0^*, \\ \bar{T}_1 &= \psi(\varepsilon_1^*) - (\Delta U_A + \gamma\Delta e)n_1^* + \alpha[\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*) + (\Delta U_A + \gamma\Delta e)n_1^*]. \end{aligned}$$

Note that the capitation levels are increasing in α and are negative when α is low enough.

We also see that the rent made by doctors increases with their degree of altruism both in Case *All* and Case 1 (with zero rent by construction in Case 0), with

$$\begin{aligned} \bar{T}_{All} &= \alpha\psi(\varepsilon_0^*), \\ T_1(n_1^*) &= \alpha\psi(\varepsilon_0^*) + (1-\alpha)\psi(\varepsilon_1^*) = \psi(\varepsilon_1^*) + \alpha(\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*)), \\ T_0(n_0^*) &= \psi(\varepsilon_0^*). \end{aligned}$$

References

[1] Adida, E. and Dai, T. 2024. Impact of Physician Payment Scheme on Diagnostic Effort and Testing, Management Science, 70 (8), pp.5408-5425.

[2] Allard, M., Jelovac, I. and Leger, P.T., 2011. Treatment and referral decisions under different physician payment mechanisms. Journal of Health economics, 30(5), pp.880-893.

[3] Bardey, D. and Siciliani, L. 2021. Nursing Homes' Competition and Distributional Implications when the Market is Two-Sided, Journal of Economics & Management Strategy, vol 30, Issue 2, p. 472-500.

[4] Bardey, D., Gromb, D., Martimort, D., Pouyet, J. 2020. Controlling Sellers Who Provide Advice: Regulation and Competition, Journal of Industrial Economics, vol 69, issue 3, p. 409-444.

[5] Beenk, K. and M. Kifmann. 2024. Optimal Financial Incentives for Physician's Sequential Diagnostic Testing and Treatment Choice, Research Paper, Hamburg Center for Health Economics.

[6] Brandt N. and Cassou, M., 2024. Care protocol adherence and health care contracting: managing information incompleteness. Available at SSRN: https://ssrn.com/abstract=4774356

[7] Chalkley, M. and Malcomson, J. 1998. Contracting for health services when patient demand does not reflect quality, Journal of Health Economics, vol 17 (1), p.1-19

[8] Choné, P. and Ma, C.T.A., 2011. Optimal health care contract under physician agency. Annals of Economics and Statistics/Annales d'Économie et de Statistique, pp.229-256.

[9] Chu, B., B. Handel, J. Kolstad, J. Knecht, U. Malmendier and F. Matejka. 2024. Cognitive Capacity, Fatigue and Decision Making: Evidence from the Practice of Medicine, UC Berkeley.

[10] Currie, J., W. B. MacLeod and K. Musen, 2024. First do not harm? Doctor decision making and patients outcomes, NBER Working Papers No. 32788.

[11] Dai, T. and Singh, S., 2020. Conspicuous by its absence: Diagnostic expert testing under uncertainty. Marketing Science, 39(3), pp.540-563.

[12] Ellis, R. and McGuire, T., 1986. Provider behavior under prospective reimbursement cost sharing and supply, Journal of Health Economics, 5, 129-151.

[13] Ellis, R. and McGuire, T., 1990. Optimal payment systems for health services, Journal of Health Economics, 9, 375-396.

[14] Felder S. and Kifmann, S., 2024. Reimbursing physicians with unknown altruism under diagnostic risk, mimeo.

[15] Garcia Mariñoso, B. and Jelovac, I., 2003. GPs payment contracts and their referral practice. Journal of Health Economics, 22(4), pp.617-635.

[16] Ghamat, S., Zaric, G., Pun, H. 2018. Contracts to Promote Optimal Use of Optional Diagnostic Tests in Cancer Treatment, Production and Operation Management, Vol. 27, No. 12, p. 2184-2200.

[17] Gupta, A. 2021. Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program. American Economic Review 111 (4), p. 1241-1283.

[18] Jack, W., 2005. Purchasing health care services from providers with unknown altruism, Journal of Health Economics, vol. 24(1), p. 73-93.

[19] Kowalski, A., 2023. Behaviour within a Clinical Trial and Implications for Mammography Guidelines, The Review of Economic Studies, 90, p. 432-462.

[20] Laffont, J.-J., and Martimort, D., 2002. The Theory of Incentives: The Principal-Agent Model. Princeton University Press.

[21] Liu, T. and Ma, C.T.A., 2013. Health insurance, treatment plan, and delegation to altruistic physician. Journal of Economic Behavior & Organization, 85, pp.79-96.

[22] McGuire T. G., 2000. Chapter 9 - Physician Agency, *in* Handbook of Health Economics. Elsevier, 1, 461-536.

[23] Mullainathan, S. and Obermeyer, Z., 2017. Does machine learning automate moral hazard and error?. American Economic Review, 107(5), pp.476-480.

[24] Mullainathan, S. and Obermeyer, Z., 2019. A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions. National Bureau of Economic Research.

[25] Obermeyer, Z., B. Powers, C. Vogeli and S. Mullainathan. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science 366 (6464), p. 447-453.

[26] OECD, 2023. Health at a Glance 2023: OECD Indicators, OECD Publishing, Paris, https://doi.org/10.1787/7a7afb35-en.

[27] Pignataro, Giuseppe, 2024. Genetic Testing, Diagnostic Effort and Physicians Incentives, mimeo.

[28] Rochaix, L., 1989. Information asymmetry and search in the market for physician services, Journal of Health Economics, 8, 53-84.

[29] WHO, Developing guideline recommendations for tests or diagnostic tools Handbook for Guideline Developments, 2014, 2nd edition (Chap 17).

[30] Wilding, A., L. Munford, B. Guthrie, E. Kontopantelis and M. Sutton. 2022. Family Doctor Responses to Changes in Target Stringency under Financial Incentives. Journal of Health Economics 85.

[31] Wu, Y., Bardey, D., Chen, Y. and Li, S., 2021. Health care insurance policies when the provider and patient may collude. Health Economics, 30(3), pp.525-543.