WORKING PAPERS

N° 1595

January 2026

"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 Testing with Moral Hazard and Adverse Selection*

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

February 10, 2026

Abstract

We analyze a setting in which physicians, who differ in their degree of altruism, first exert diagnostic effort before deciding whether to administer a test to determine the most appropriate treatment. Diagnostic effort yields an imperfect private signal of the patient's type, whereas the test provides a more accurate assessment. Absent corrective transfers, physicians exert too little diagnostic effort and may rely excessively on testing. When altruism is either homogeneous or observable, the first-best allocation can be decentralized through a payment scheme consisting of (i) a pay-for-performance (P4P) component, based on the proportion of correctly treated patients, to induce the optimal diagnostic effort, and (ii) a fixed component to ensure both the optimal testing decision and physician participation. When altruism is heterogeneous and privately known to physicians, the two-part tariff that decentralizes the first-best is no longer incentive compatible. The optimal contract is pooling rather than separating, an instance of non-responsiveness. Its uniform P4P component induces more altruistic physicians to exert higher diagnostic effort, while the fixed component must be conditioned on diagnostic test costs in order to promote optimal testing decisions.

Keywords: diagnostic risk, personalized medicine, non-responsiveness, fixed-part 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), 1^{st} CEPR Health Economics Conference (TSE, Toulouse, 2024), 45^{th} Journés des Economistes Français de la Santé (Bordeaux), Digital Health Workshop (Toulouse, 2024), 63^{rd} Congrès de la Société Canadienne de Sciences Economiques (Montréal), 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 Toulouse Capitole, Toulouse, France. Email: philippe.dedonder@tse-fr.eu

§Département des Sciences Economiques, ESG-UQAM, Montréal, Canada; CIRANO, Montréal; CREEI, Montréal; CESifo, Munich, Germany. E-mail: leroux.marie-louise@uqam.ca

1 Introduction

The healthcare sector is a major part of developed economies, with average spending in OECD countries at 9.6% of GDP in 2022. The U.S. spends the most (over 16%), while countries like Canada, Germany, France, and the U.K. spend about 10-12%. These figures include public and private health services, pharmaceuticals, and long-term care (OECD, 2023), but do not take into account the indirect economic impacts of health spending (such as lost earnings, reduced leisure time, and diminished home production).

Physicians are key economic actors whose medical decisions strongly affect healthcare outcomes. Diagnostic errors -missed, incorrect, delayed, or miscommunicated diagnoses- cause 16% of preventable patient harm and are a leading source of malpractice claims.[1] Addressing their economic causes is vital to improving care and reducing costs.

The rational use of diagnostic tests is central to the efficient allocation of treatments. Classic examples include molecular profiling of microbes to distinguish between bacteria, fungi, or viruses, and the use of different antibiotic families to target specific bacterial strains or to address resistance. Precision medicine provides another illustration, as it relies on diagnostic tests to tailor therapies to patient subgroups, often defined by genetic characteristics, thereby reducing errors and improving outcomes. Notable applications include cancer biomarkers and pharmacogenomic testing to personalize drug dosages.[2] Despite accounting for less than 5% of healthcare spending (Akhmetov and Bubnov, 2015), diagnostic tests are becoming increasingly significant with advances in AI (Mullainathan and Obermeyer, 2017, 2019).

The efficient use of diagnostic tests depends on physicians' incentives to prescribe these tests efficiently. While some healthcare systems see an overuse of diagnostic tests which do not necessarily improve patients' outcomes,[3] underuse of diagnostic tests, particularly for detecting

[1] See WHO's report at https://www.who.int/campaigns/world-patient-safety-day/world-patient-safety-day-2024. Accessed October 27, 2025.

[2] In North America, spending on precision medicine reached 18.5 billion USD in 2022 and is projected to rise to nearly 67 billion by 2027 (https://www.statista.com/). In 2024, the FDA approved 18 new personalized medicines, accounting for 38% of all newly approved molecular entities (https:trial.medpath.com).

[3] There is a large literature on the overuse of imaging technology– see Felder and Kifmann (2024) for MRIs for instance. Kowalski (2023) 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

pathogens like bacteria, fungi, or viruses, can negatively impact patients' health. 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 physician payment schemes has examined how fee-for-service, salary, fixed payments, and pay-for-performance (P4P) affect medical care, using both positive and normative approaches (see Section 2). Yet, few studies analyze these schemes in the context of diagnostic tests. This manuscript fills that gap by exploring payment models when such tools are available.

We develop a model with two patient types (A, B) and two treatments (D, P). Think of urinary tract infections: type-B patients have E. coli, typically susceptible to ciprofloxacin (D), while type-A patients have Klebsiella pneumoniae, often resistant to ciprofloxacin but sensitive to cefuroxime (P). Initially, the patient's type is unknown to both the patient and the physician, with the physician only aware of the population distribution of types. In the absence of further information, the optimal approach is to treat everyone with the cheaper D, though type-A patients would benefit from P. A diagnostic test, such as urine culture with antibiogram, reveals the bacterium and the appropriate antibiotic. Physicians can identify patient type through traditional (lengthy) diagnostic effort -such as symptom checks or medical history- which is costly and yields imperfect signals, and / or through a diagnostic test, assumed perfectly accurate.

In our model, physicians first choose their level of diagnostic effort, and then, based on the initial diagnostic signal, determine whether to proceed with a diagnostic test. We assume that they are imperfectly altruistic. As Currie *et al.* (2024) put it, doctors "care about patient welfare, but also about their own welfare which makes them imperfect agents. (p.36)" In the first-best scenario, altruism is either homogeneous across physicians or observable by the planner. In the second-best, we adopt the more realistic assumption that altruism is both heterogeneous and privately known to physicians.[4]

technology, revealing that these patterns are often driven by supply- rather than demand-side considerations.

[4]Galizzi *et al.* (2023) document considerable heterogeneity in physicians' altruistic preferences. Gertler and Kwan (2024) also study how profit motives versus altruism influence medical decisions. In an audit experiment

In the first-best, the social planner chooses effort and test use based on diagnostic effort signals. Three cases arise: with low test costs, all patients are tested and effort is unnecessary; with intermediate costs, only signal-A patients are tested, consistent with practice;[5] with high costs, no tests are prescribed. Effort and testing act as substitutes: as testing expands, optimal costly effort declines because the test provides more reliable information.

Assuming effort is unobservable or non-contractible, we show that the social planner can implement the first-best allocation through a transfer system combining a P4P component tied to the proportion of well-treated patients[6] and a fixed payment that varies with the testing scenario (all patients tested, only signal-A, or none).

When altruism is heterogeneous and privately known to physicians, the second-best optimal contract takes the form of pooling: all physicians are offered the same P4P and fixed payment formulas. This outcome reflects the property of non-responsiveness. With only fixed transfers calibrated to ensure participation, low-altruism physicians exert less effort than their more altruistic counterparts, with both falling short of the first-best level. It is therefore optimal to design incentives that primarily target an increase in the effort of low-altruism physicians. However, because their weaker concern for patient welfare requires higher compensation, a misalignment arises between the planner's objectives and physicians' preferences, making a separating contract undesirable.

The pooling contract's P4P component is based on average altruism, causing less (resp. more) altruistic physicians to exert below- (resp. above-) first-best optimal diagnostic effort. Payments set just high enough to satisfy participation constraints leave rents to less altruistic physicians, which rise with effort. Since effort and testing are substitutes, this leads to under-use

with standardized patients in Kenya, they show that profit-motivated providers over-report positive malaria tests by about 30 percentage points compared to altruistic providers, despite identical clinical cases.

[5]For example, mild bacterial infections may be treated with standard antibiotics like ciprofloxacin, while severe cases require a test to target bacteria such as Klebsiella, treated with cefuroxime.

[6] Currie *et al.* (2024) more broadly argue that, while health economics addresses issues relevant to other high-skilled labor markets, it is distinct in that downstream impacts of expert decisions are often more readily observable in healthcare than in other sectors. For instance, missed diagnoses often become evident when patients return for care (Abaluck *et al.*, 2016). Policies often capitalize on this observability. In the United States, for instance, the *Hospital Readmissions Reduction Program* (HRRP) penalizes hospitals with excess Medicare readmissions (Gupta, 2021), while in England, general practitioners face penalties when hypertension control rates fall below target thresholds (Wilding *et al.*, 2022).

of tests. The social planner can correct this bias by increasing the fixed component of payments when more patients are tested, but this requires the second-best transfers to be conditioned both on the fraction of patients tested and on the test cost.

The manuscript is organized as follows. The next section reviews the related literature. Section 3 presents the set-up. Section 4 derives the first-best allocation, while Section 5 analyzes the physicians' problem. Section 6 shows how to decentralize the first-best allocation with a two-part tariff. Section 7 relaxes the assumption that physicians' altruism is observable and studies the optimal second-best contract. Section 8 concludes.

2 Related Literature

Our manuscript 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 studies examine how physicians with varying altruism choose treatments, assuming away diagnostic tests. Early work assumes altruism is public information,[7] while recent papers treat it as private. Jack (2005) studies non-contractible quality, and Choné and Ma (2011) assume contractible care quantities when physicians have private information on patients' illness; both show that decentralizing the first-best allocation is impossible. Liu and Ma (2013) find it is feasible only if physicians commit to a treatment plan before accepting a contract. In this literature, the article closest to ours is Choné and Ma (2011) who also obtain a pooling contract, but as a consequence of the multidimensional adverse selection structure. In our set-up, the pooling contract result comes from the non-responsiveness situation generated by the combination of adverse selection and moral hazard. Without moral hazard but with adverse selection on doctors' altruism, a menu of separating contracts would be optimal.

In the second branch of the literature, which studies incentives for both diagnostic testing and treatment choices, most papers account for moral hazard -hidden action (diagnostic effort) and hidden information (the signal generated by that effort)- but do not incorporate adverse

[7] See Allard *et al.* (2011), Chalkley and Malcomson (1998), Ellis and McGuire (1986, 1990), Rochaix (1989).

selection stemming from physicians' heterogeneity. Garcia Mariñoso and Jelovac (2003) model an income-maximizing physician who exerts costly diagnostic effort, receives an imperfect signal, and decides whether to treat or refer to a specialist. Optimal contracts combine a capitation payment, a bonus for successful treatment, and a fee for referral. First- and second-best allocations are achieved depending on whether participation constraints are satisfied ex ante or state-contingent. Beenk and Kifmann (2024) extend this to two tests: the first costly to the physician and imperfect, the second costly to the payer but perfect. The optimal contracts encourage selective use of the second test through capitation, performance-based payments, and fees for the second test.[8] Our setting also draws on Adida and Dai (2024), where an imperfectly altruistic physician first chooses diagnostic effort, generating a symmetrical imperfect signal, then decides whether to perform a perfect test. While they study how fee-for-service affects effort and testing incentives, we focus on the normative question of the optimal payment schemes decentralizing first- and second-best allocations.

Felder and Kifmann (2024), like our paper, bring together these two strands of the literature. They study physicians with varying degrees of altruism who treat patients with differing prior probabilities p of severe versus mild conditions. Physicians observe p at no cost. They do not exert effort but may use a costly, imperfect diagnostic test. The first-best allocation prescribes treating for the mild (resp. severe) condition without testing if p lies below (resp. above) a low (resp. high) threshold, and testing patients whose p falls in the intermediate range. With unobservable altruism, the planner offers a contract menu with reduced cost-sharing, leading all but the most altruistic physicians to overtreat some patients to limit informational rents.[9]

Compared to the existing literature, and to borrow from 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 (on the

[8] Pignataro (2024) studies adverse selection where physicians differ in unobservable diagnostic ability. Like Beenk and Kifmann (2024), doctors first exert costly, imperfect diagnostic effort, then decide on a perfect test. As in our paper, diagnostic effort and testing are substitutes. Brandt and Cassou (2024) apply this to prospective hospital payments, showing optimal cross-subsidization within care pathways.

[9] Ghamat *et al.* (2018) and Dai and Singh (2020) also introduce adverse selection in frameworks involving diagnostic tests. In the former, private information pertains to patients' characteristics, whereas in the latter it relates to physicians' ability.

degree of altruism of physicians), and to analyze both the first-best and second-best incentive schemes for physicians who exert a diagnostic effort and may subsequently decide 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. Neither the patient nor the physician knows the patient's type at the beginning of the period. We set $\lambda = 1/2$ for simplicity. We relax this assumption in Appendix 9.6, and solve the model for a generic value of $\lambda < 1/2$. Although the analysis becomes more intricate (for reasons summarized in footnote 22), our results do not qualitatively change.

There are two treatments available to patients, indexed by $j \in \{P, D\}$, where D stands for the "default" treatment while P stands 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.[10] 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.

[10] We do not explicitly model health insurance in this model, so that we can focus on decisions by physicians, and how they can be influenced by the social planner.

Assumption 1 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 average 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).[11]

3.2 Physicians' effort and diagnostic tests

Physicians can obtain information about a patient's true condition through two non-exclusive methods: diagnostic effort (or clinical assessment) and diagnostic tests. We explain the characteristics of these two approaches in turn.

Diagnostic effort can be measured by factors such as the time spent with the patient, the thoroughness of symptom examination, and of the investigation into the patient's medical and family history. Formally, the physician exerts an effort, ε, which generates a signal about the patient's type. The signal $\sigma \in \{A, B\}$ has the following precision

$$\varepsilon \;=\; \Pr(\sigma = B \mid i = B) \;=\; \Pr(\sigma = A \mid i = A) \;\in\; [1/2, 1].$$

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 frequencies of the four cases (2 signals and 2 types) in the population. As patients are randomly selected from the population, the frequencies of each case also indicate the likelihood that any given patient falls into one of the four categories.

[11] In the generic case where $\lambda < 1/2$, Assumption 1 can be weakened to $(1-\lambda)\Delta U_B > \lambda U_A > 0$, a condition that may hold even when $\Delta U_B < \Delta U_A$. In this case, the default treatment is still the one best suited to type-B patients, not due to higher individual health benefits, but because basic-type patients are more prevalent and thus generate greater population-level benefits. See Appendix 9.6.

	True B	True A	Total
Signal B	$\varepsilon/2$	$(1-\varepsilon)/2$ (false neg.)	$1/2$
Signal A	$(1-\varepsilon)/2$ (false pos.)	$\varepsilon/2$	$1/2$
Total	$1/2$	$1/2$	1

Table 1: Population frequencies

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 frequency in the population of false positives and false negatives. In diagnostic test terminology, this corresponds to a technology with identical sensitivity and specificity. This effort has a cost to the physician, which is increasing and convex and denoted by $\psi(\varepsilon)$. We assume an Inada's condition such that $\psi(1/2) = 0$, $\psi'(1/2) = 0$ and $\lim_{\varepsilon \to 1} \psi'(\varepsilon) = +\infty$. Finally, we assume throughout that the signal on the patient's type the physician has received remains her private information.[12]

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.[13] Throughout the paper, we assume that both the fraction of patients tested and the test cost z are observable and potentially contractible by the health authority. However, the decision to test a *specific* individual and the test result remain the doctor's private information.[14]

[12]This assumption is the most realistic. It prevents the social planner from compelling physicians to test patients based on the signal they receive.

[13]We address both types of costs (treatment and diagnostic test) symmetrically by assuming that the patient bears them both (recall that patient's utility U_i^j is net of the treatment cost). 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. Also, our results do not hinge on the assumption of a perfect test. The Appendix 9.1 indeed shows that our main results follow through when introducing a test with precision less than 100%. These results rely on two features: signals are obtained sequentially and both entail costs. As a result, the test is used only if it can influence the physician's treatment decision, and its binary signal is always followed once performed. This extended model can thus be solved with the same structure as in the main text, yielding formulas that naturally extend the perfect-test case.

[14]If test results were observable, collusion-proof contracts would be necessary to prevent collusion between patient and physician (see, for instance, Wu *et al.* [2021]). This approach would be impractical in our setup, where the patient has a passive role.

The timing of the game runs as follows. First, the health authority proposes a payment scheme to physicians, who then decide to accept or reject it based on the participation constraints outlined in Section 5. If they reject it, the game ends. 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 socially optimal outcomes and guiding physicians' choices. The utility of a patient of type $i = \{A, B\}$ prescribed with treatment $j = \{D, P\}$ is $\tilde{U}_i^j = U_i^j - lz$, where $l = \{0, 1\}$ according to whether a diagnostic test is prescribed or not.

Doctors care both about their own income (*i.e.* any transfer T received from the social planner) net of their cost of effort, $\psi(\varepsilon)$, and about their patient's welfare. More precisely, we assume that a doctor puts a weight of $\alpha \in [0, 1]$ on the patient's utility, so that her utility is given by[16]

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

We denote by α the physician's degree of altruism, and we call her imperfectly altruistic if $\alpha < 1$. Note that we do not endogenize the number of patients seen by a doctor, which is taken as exogenous and independent of her effort and testing choices.[17]

We now proceed as follows. The 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.

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

[16] Altruism may reflect an intrinsic trait of the doctor, but it can also be seen as a proxy for reputational concerns, malpractice risk, patient bargaining power (as in Choné and Ma, 2011), or repeated interactions (as in Rochaix, 1989).

[17] In a previous version of this manuscript (Bardey *et al.*, 2024), we allow for the number of physician visits to decrease when patients are appropriately matched to a treatment, and for patients to dislike visiting their physician. These assumptions complicate the notation without providing additional insights, so we have omitted them from this manuscript.

4 The social optimum

We consider a utilitarian social planner maximizing the sum of patients' and physicians' utility while excluding the altruistic component of the physician's utility.[18] 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 social planner (*i.e.* there is no cost of public funds), and thus plays no role in this section.[19]

4.1 Optimal effort levels

We have to deal with three possible cases, where we denote the first-best 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

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

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

Case 0: No test is prescribed to anyone.

In such a case, welfare is a function of the effort level ε_0. After having exerted this effort, the physician can either treat everyone with D anyway (in which case the welfare level is $(U_A^D + U_B^D)/2 - \psi(\varepsilon_0)$, or treat the patient according to the signal received, with a welfare level of

$$W_0(\varepsilon_0) = -\psi(\varepsilon_0) + \frac{\varepsilon_0}{2}(U_B^D + U_A^P) + \left(\frac{1-\varepsilon_0}{2}\right)(U_A^D + U_B^P). \tag{2}$$

In the latter case, the first-order condition for ε_0 is:

$$\psi'(\varepsilon_0^*) = \frac{\Delta U_B + \Delta U_A}{2}. \tag{3}$$

[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).

[19] Incorporating the cost of public funds would have a quantitative, but not qualitative, impact on our findings. For example, it would lower the optimal effort levels calculated below, as the transfers required to incentivize effort become more costly to society. Similarly, it would bias optimal diagnostic decisions toward the least expensive options. However, the increase in mathematical complexity is not justified by the new insights.

The right-hand side shows that a marginal increase in effort decreases by one half the proportion 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 optimal welfare level is given by introducing ε_0^* in equation (2). We exclude from the analysis the uninteresting case where the welfare attained by following the signal at the optimal effort level ε_0^* is lower than the welfare obtained by treating all patients with the default treatment D under the minimum effort level. We then impose

$$W_0(\varepsilon_0^*) > \frac{1}{2}(U_B^D + U_A^D),$$

so that the diagnostic technology is sufficiently effective (through low cost and/or substantial gains in patient utility from correct diagnosis) that it is not dominated by the strategy of treating all patients with the default treatment without any diagnostic effort.

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 is given by

$$W_1(\varepsilon_1) = -\psi(\varepsilon_1) + \frac{1}{2}U_B^D + \frac{\varepsilon_1}{2}U_A^P + \left(\frac{1-\varepsilon_1}{2}\right)U_A^D - \frac{z}{2}, \tag{4}$$

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

$$\psi'(\varepsilon_1^*) = \frac{\Delta U_A}{2}, \tag{5}$$

with, compared to (3), a marginal gain only on false negatives (ΔU_A).The optimal welfare level is further obtained by introducing ε_1^* in equation (4).

We then obtain the following proposition:

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

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

The intuition runs as follows: as explained after equation (5), 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. Note that conditional on the test decision, the optimal effort levels are independent of the test cost z.[20]

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 signal B is dominated by testing only patients with signal A, whatever the effort level.*

Proof. The welfare level reached when only patients with signal B are tested is

$$\frac{1}{2}U_A^P + \frac{1-\varepsilon}{2}U_B^P + \frac{\varepsilon}{2}U_B^D - \psi(\varepsilon) - \frac{z}{2} < W_1(\varepsilon),$$

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., positives), with a per patient marginal gain of ΔU_A (resp., ΔU_B). Assumption 1 then ensures that testing only signal-A patients is socially preferred to testing only signal-B patients.[21]

4.2 Optimal diagnostic testing decision

The social planner must decide whether to test no one, only those with a signal A, or everyone, represented by Cases 0, 1 or *All* respectively. This decision is influenced by the test cost, z. We thus need to compare $W_{All}(\varepsilon^*_{All})$, $W_1(\varepsilon^*_1)$ and $W_0(\varepsilon^*_0)$ as a function of the value of z. More precisely, we focus on the richer scenario where, as the cost of the test increases, the optimal

[20] Empirically, our model's prediction of substitution between diagnostic effort and test use can be investigated using shocks to test costs or availability -such as reimbursement changes or technology rollouts- to observe corresponding shifts in physicians' time and attention to the patient. For instance, Ergun-Sahin *et al.* (2022) and Chu *et al.* (2024) use within-hospital variation in workload or fatigue to identify the extent of the substitution between testing and diagnostic effort.

[21] We show in Appendix 9.6 that this result still holds when the proportion λ of type A is inferior to 1/2.

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. 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 test cost level below which it is optimal to exert no effort and test the entire population. It is determined by the condition $W_{All}(1/2) = W_1(\varepsilon^*_1)$, so that

$$z^*_{All} \equiv (1 - \varepsilon^*_1)\Delta U_A + 2\psi(\varepsilon^*_1). \tag{6}$$

The intuition behind this formulation of the threshold is the following. The only gain (from a welfare perspective) from testing only patients with an A signal, rather than all patients, is that we only test half the population, saving half the tests costs. This gain is balanced by two losses, corresponding to the two (positive) terms defining z^*_{All}: (i) a reduction of correct treatments among type-A agents (those who receive a signal B), and (ii) a cost of effort for the physician. The larger these two 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 testing only patients signaling A and not testing anyone, $W_1(\varepsilon^*_1) = W_0(\varepsilon^*_0)$, so that:

$$z^*_1 \equiv 2\left(\psi(\varepsilon^*_0) - \psi(\varepsilon^*_1)\right) + (1 - \varepsilon^*_0)\,\Delta U_B - (\varepsilon^*_0 - \varepsilon^*_1)\,\Delta U_A. \tag{7}$$

The intuition behind this formulation is similar to that of z^*_{All}. Transitioning from testing only patients with signal A to testing no one allows saving on the test cost for half the population. This has three consequences corresponding to the three 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 when testing patients with an A signal, which 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.[22]

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.

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) and taking into account her participation constraint

$$T_k \geqslant \psi(\varepsilon_k^{eq}),\ \forall k \in \{All, 1, 0\}, \tag{8}$$

which requires that the transfer T_k received from the authority in any Case $k \in \{All, 0, 1\}$ has to compensate for the effort disutility. Recall that we assume throughout the manuscript that the social planner observes the fraction of patients tested, and thus knows the Case k in which physicians operate. For reasons of political feasibility and ethics, we exclude the term $\alpha\tilde{U}_i^j$ from the participation constraint, thereby preventing the planner from exploiting physicians' altruism to reduce the transfers needed to ensure participation. In other words, the social planner cannot free ride on doctors' intrinsic motivation to treat patients by offering compensation levels below the cost of effort. Moreover, for simplicity, we assume that physicians' altruism does not affect their outside utility either.[23]

[22] In Appendix 9.6, we show that with a generic proportion λ, the share of A signals in Table 1 differs from λ and depends on effort ε, with minority types over-represented. For $\lambda < 1/2$, the optimal effort ε^*_1 in (5) also depends on test cost z, creating a fixed-point issue in computing z^*_{All} and z^*_1. While adding complexity, this generalization yields limited additional insight.

[23] This simplification does not qualitatively affect our results, as explained in the discussion following Proposition 4. These participation constraints can be interpreted as limited-liability constraints, as in, among others, Liu and Ma (2013) and Felder and Kifmann (2024).

The physician's utility depends on whether she tests all patients (V_{All}), only those with an A signal (V_1) or nobody (V_0):

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

This yields the following (equilibrium) levels of efforts:

$$\psi'(\varepsilon_{All}^{eq}) = T'_{All}(\varepsilon_{All}^{eq}), \tag{9}$$

$$\psi'(\varepsilon_0^{eq}) = \alpha\left(\frac{\Delta U_B + \Delta U_A}{2}\right) + T'_0(\varepsilon_0^{eq}), \tag{10}$$

$$\psi'(\varepsilon_1^{eq}) = \alpha\left(\frac{\Delta U_A}{2}\right) + T'_1(\varepsilon_1^{eq}). \tag{11}$$

Comparing (3) to (10), and (5) to (11), we obtain that, when physician's payments are fixed (*i.e.* $T'(\varepsilon) = 0$), less-than-perfectly altruistic physicians under-provide effort in Cases 0 and 1. Moreover, total differentiation of (10) and (11) 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 threshold z_{All}^{eq} is such that $V_{All}(z_{All}^{eq}) = V_1(z_{All}^{eq})$, and this condition yields:

$$z_{All}^{eq} \equiv (1-\varepsilon_1^{eq})\Delta U_A + \frac{2}{\alpha}(T_{All} - T_1) + \frac{2}{\alpha}\psi(\varepsilon_1^{eq}). \tag{12}$$

Comparing the equilibrium threshold z_{All}^{eq} with its optimal counterpart z_{All}^{*} requires knowing the value of the fixed component of transfers to physicians at the social optimum. In the absence of any cost of public funds, these transfers are indeterminate (see Section 4). We therefore assume that the social planner minimizes transfers (and thus, rents) while still satisfying physicians' participation constraints. This assumption is reasonable -for example, on political economy grounds or with small costs of public funds- and we state it formally to clarify which results specifically rely on it.

Assumption 2 *When indifferent between transfer levels, the social planner minimizes them, so that the participation constraint (8) binds for at least one Case k.*

Since transfers do not affect the determination of the optimal test threshold z^*_{All}, Assumption 2 implies here that they can be set so that (8) holds with equality in both Cases *All* and 1. Consequently, for a given level of effort, the second and third terms in (12) cancel out. Comparing (6) and (12), we then obtain that the physician stops testing all patients at a test cost threshold that is socially too low, since transfers fully offset their effort cost.

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

$$z_1^{eq} \equiv (1-\varepsilon_0^{eq})\Delta U_B - (\varepsilon_0^{eq} - \varepsilon_1^{eq})\Delta U_A + \frac{2}{\alpha}(T_1 - T_0) + \frac{2}{\alpha}(\psi(\varepsilon_0^{eq}) - \psi(\varepsilon_1^{eq})). \tag{13}$$

For a given level of effort, when transfers are set according to Assumption 2, the last two terms in (13) cancel out. This results in the physician stopping testing too early, as she does not account for her higher effort cost (unlike the social planner). We summarize those results in the following proposition.

Proposition 2 *When transfers are fixed (*i.e. *not dependent on effort and testing decisions), physicians exert insufficient effort ($\varepsilon_k^{eq} < \varepsilon_k^*, \forall k \in \{0,1\}$), and this under-provision of effort decreases as the physician's altruism degree, α, increases. Moreover, when transfers are set according to Assumption 2, the equilibrium features under-testing for any given level of effort.*

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)

In this section, we show how to decentralize the optimal effort levels as well as the testing decisions. We first state the general formulas for decentralization, before looking at how to implement them when the proportion of well-treated patients is observable and 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:

$$T'_{All}(\varepsilon^{eq}_{All}) = 0, \tag{14}$$

$$T'_{0}(\varepsilon^{eq}_{0}) = (1-\alpha)\left[\frac{\Delta U_B + \Delta U_A}{2}\right], \tag{15}$$

$$T'_{1}(\varepsilon^{eq}_{1}) = (1-\alpha)\left[\frac{\Delta U_A}{2}\right]. \tag{16}$$

The intuition behind these formulas is straightforward. 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 behave as if she were perfectly altruistic (since the term multiplying $1-\alpha$ measures the effort's marginal social benefit).

Moreover, in order to ensure that z^{eq}_{All} and z^{eq}_{1} correspond to their optimal levels (once the effort levels have been optimally chosen), we need to set payment functions $T_k(.)$ satisfying

$$T_1(\varepsilon^*_1) - T_{All}(\varepsilon^*_{All}) = (1-\alpha)\psi(\varepsilon^*_1), \tag{17}$$

$$T_0(\varepsilon^*_0) - T_1(\varepsilon^*_1) = (1-\alpha)[\psi(\varepsilon^*_0) - \psi(\varepsilon^*_1)]. \tag{18}$$

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 greater effort is required. Moreover, this distortion decreases with doctor's altruism. So, eq. (17) and (18) show that the difference in transfers between two adjacent cases must be $(1-\alpha)$ times the difference in (optimal) effort cost in each case. In addition, the participation constraint (8) will set the (minimum) level of transfers.

It is clear from above that a fixed transfer cannot by itself decentralize the first-best allocation. In the next section, we focus on the situation where the social planner can implement a payment scheme that depends on the proportion of well-treated patients.

6.2 Decentralization based on the proportion of correctly treated patients

We consider the realistic situation where effort cannot be contracted upon, while the proportion of correctly treated patients can. A payment scheme that depends on the proportion of well-treated patients is reasonable for a range of medical conditions where established medical protocols -such as blood tests or hospital readmissions- can help infer whether a patient received appropriate care. This assumption is widely used in the health economics literature and aligns with payment schemes used in practice (see footnote 6). Importantly, we do not require payments to be contingent on individual treatment outcomes, which would be a more stringent condition. Instead, we base payments on the *proportion* of correctly treated patients, which we denote by n_k in Case $k \in \{All, 1, 0\}$.[24]

More precisely, we allow for case-specific two-part tariffs, consisting of a fixed payment and a variable P4P component linked to the proportion of well-treated patients, n_k.[25] 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 of them being B types, the other half A types). We then obtain that condition (15) is satisfied if we use the following contract which is linear in the proportion of well-treated patients:

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

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

In Case 1, we have 1/2 correctly treated B-type patients (since all B-type patients are

[24]This assumption best fits antibiotic stewardship programs where it is easy to reward appropriate prescription in cases of bacterial infection, and where guidelines are clear about when antibiotics are or are not appropriate. For cardiovascular care such as acute myocardial infarction, clear protocols also exist (*e.g.*, administration of beta-blockers, aspirin, timely PCI), making it easier to assess appropriate treatment. By contrast, for some medical conditions it is more difficult to retrospectively and objectively determine the "correct" treatment, which makes assessing the proportion of appropriately treated patients less straightforward and may introduce additional challenges, such as cherry-picking patients or avoiding high-risk cases (see Miller and Babiarz, 2013). Our analysis focuses on settings where these difficulties are limited.

[25]In our setting, since each patient is treated only once, the fixed part of the payment can be seen either as a capitation or a fee-for-service. If, however, patients could return for additional visits, this component of the payment would be more appropriately described as fee-for-service. Bardey et *al.* (2021) distinguish fee-for-service from capitation by introducing a time constraint into physicians' programs.

identified either through the diagnostic effort or the diagnostic test) and $\varepsilon_1/2$ correctly treated A-type patients, for a total proportion 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

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

since differentiating it satisfies (16).

Comparing the terms multiplying n_k in (19) and (20), we see that they are larger in (19) since $\Delta U_B > \Delta U_A$. For any given α, the social planner needs to generate stronger incentives to exert effort in Case 0 because the marginal gains of a correct treatment apply both to types A and B, while in Case 1, it applies only to type A (with a lower marginal gain ΔU_A). We then obtain the following proposition.

Proposition 3 *When altruism is observable and the proportion of correctly treated patients is observable and contractible, the social planner can decentralize the first-best allocation with a transfer composed of a fixed component together with a P4P component, as given by (19) and (20). The fixed parts of the transfers are lower in Cases 0 and 1 than in Case All. When transfers are set according to Assumption 2, the rents enjoyed by physicians in Cases* 1 *and All increase with their degree of altruism.*

Proof. See Appendix 9.2. ■

The P4P component of the contract decreases with the physician's degree of altruism (as the need to incentivize effort diminishes). When transfers are set according to Assumption 2 (rent minimization), a lower portion of the effort cost must then be covered through the fixed part of the payment for less altruistic doctors. At the extreme and as we show in Appendix 9.2, for doctors with very low altruism, the P4P component can be so substantial that the fixed part of the payment may even turn negative to prevent them from under-testing at equilibrium.

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

Up to this point, we have assumed that physicians' altruism is either homogeneous or observable. In this section, we assume that differences in altruism are privately known by physicians. We look at how to decentralize their optimal diagnostic effort and testing decisions by concentrating on two-part tariffs, namely payment schemes consisting of both a fixed level and a variable (P4P) part.

We proceed as follows. We first focus on the effort level within any (observable) Case k. In Section 7.1, we show that asymmetric information on the altruism degree is best addressed by offering a pooling contract for each case. We also compute the levels of the P4P part of the physician's remunerations decentralizing the second-best effort level in each case. We then decentralize the physicians' testing decisions by looking at the optimal menu of three (*i.e.* one for each Case k) pooling contracts in Section 7.2. This allows us to find the optimal levels of the fixed parts of the pooling contracts, which will depend on both the testing decision and the cost of the diagnostic test.

7.1 Within cases: A pooling contract decentralizes the second-best effort levels

Following the standard two-types approach in the principal-agent literature, we assume from now on 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 $\alpha_L < \alpha_H < 1$. There is a proportion ν of low-altruism physicians. The physician's altruism degree is her private information, and the contract can only be conditioned on the proportion 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.

The social planner is allowed to offer a menu of contracts (one for each type of physicians) consisting as before, of a P4P component and a fixed part:

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

We now demonstrate that, for each Case k, the social planner should offer a single contract to both types of physicians, rather than separating ones.

Proposition 4 *The Second-Best contracts in Cases 0 and 1 are pooling.*

Proof. See Appendix 9.3. ■

The intuition for this result is as follows. With fixed transfers only, both physicians' type under-provide 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 social 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). While non-responsiveness has been documented in various contexts, the specific mechanism we highlight -arising from the conjunction of moral hazard and adverse selection within a two-type framework satisfying the single-crossing condition- is, to the best of our knowledge, novel.

Before turning to the implications of Proposition 4 for our setting, it is worth commenting on the robustness of the non-responsiveness result. First, this result does not rely on the restriction

to two types. In fact, separating contracts are generally easier to sustain in a two-type environment (see Laffont and Martimort, 2002). Hence, our argument holds *a fortiori* in settings with more types, including a continuum. Second, the result is also robust to the introduction of a noisy proxy for altruism. For example, suppose society is composed of observable groups of physicians, W (women) and M (men), with women being on average more altruistic than men, so that a physician's gender serves as a noisy signal of altruism. In this case, the optimal policy consists in offering a separate pooling contract to each group, with a steeper slope for men than for women. Third, the result also remains valid under alternative forms of participation constraints that account for type-dependent reservation utilities. This finding is consistent with Jullien (2000) and Laffont and Martimort (2002), who show that type-dependent participation constraints tend to make pooling contracts more likely to be optimal at the second best. Fourth, we examine the robustness of this result to the introduction of piecewise linear schedules in the online Appendix 9.7. Although Proposition A.2 shows that expanding the policy space in this manner allows for the construction of incentive-compatible separating contracts that can improve upon the pooling contract analyzed here, we stress there that such separating schemes are difficult to generalize to settings with multiple physician types -both in theory and in practice. Fifth, 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). We further discuss the robustness of the non-responsiveness result in the conclusion.

In our context, this results in the same contract $(\bar{T}_k, \beta_k)$ being proposed to all physicians in Case k, with the resulting payments:

$$T_{i,k} = \bar{T}_k + \beta_k n_{i,k}, \tag{21}$$

varying across types because effort levels (and thus proportion of correctly treated patients)

differ between type-L and type-H physicians even when they sign the same contract.

Note that the formal proof developed in Appendix 9.3 makes use of the following assumption:

Assumption 3 *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 levels), while Proposition 4 holds more generally for any convex effort cost function. We maintain Assumption 3 in the rest of the manuscript. We now determine the optimal levels of β_k.

Proposition 5 *Under Assumption 3, the second-best slope of the P4P component for the pooling contract in Case 1 and Case 0 is:*

$$\beta_1^{SB} = (1 - \bar{\alpha})\Delta U_A, \tag{22}$$

$$\beta_0^{SB} = (1 - \bar{\alpha})\left[\frac{\Delta U_A + \Delta U_B}{2}\right], \tag{23}$$

where $\bar{\alpha} = \nu\alpha_L + (1-\nu)\alpha_H$ is the average physician altruism. This second-best contract generates the following ranking of efforts: $\varepsilon_{L,k}^{SB} < \varepsilon_k^* < \varepsilon_{H,k}^{SB}$ $\forall k = \{0, 1\}$.

Proof. See Appendix 9.4 ■

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, for any case $k = \{0, 1\}$, 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 is unaffected by the physician's degree of altruism, takes an intermediate value.

In the next section, we determine the values of the payments' fixed part.

7.2 Across cases: The fixed part of the optimal payments decentralizing the second-best testing decisions

Before turning to the decentralization of the second-best testing decisions, we briefly compare them with their first-best counterparts. The second-best testing thresholds differ solely because effort levels diverge between the first- and second-best.[26] Regarding the choice between testing

[26] Compare $z_{All}^{SB}(\alpha_i)$ in (A.19) with z_{All}^* in (6), and $z_1^{SB}(\alpha_i)$ in (A.25) with z_1^* in (7).

everyone or only signal-A patients, the second-best distortion of effort reduces welfare in the latter case, whereas welfare under universal testing remains unchanged (since the minimum effort level is exerted anyway). As a result, physicians optimally choose universal testing at higher test costs in the second-best than in the first-best. At the same time, the welfare level attained when no patient is tested is also lower at the second-best, which generates two effects working in opposite directions on the second-best threshold $z_1^{SB}(\alpha_i)$, leaving the net impact indeterminate. This leads to the following Proposition.

Proposition 6 *(a)* $z_{All}^{SB}(\alpha_i) > z_{All}^{*}$*, (b)* $z_1^{SB}(\alpha_i) \gtrless z_1^{*}$.

Proof. Straightforward differentiation of (A.19) and (A.25) with respect to ε_1 and ε_0, using the rankings of first-best and second-best effort levels from Proposition 5. ■

We now turn to the decentralization of these second-best testing thresholds. The fixed component of the payments must be designed to induce optimal testing decisions from both physician types while ensuring participation. Our approach is as follows. First, we set transfers according to Assumption 2, *i.e.*, at their lowest levels consistent with physicians' participation constraints given the second-best effort levels defined above. We then compare equilibrium and second-best optimal testing thresholds to identify the distortions induced by such transfers. Finally, we determine how to adjust these transfers to restore second-best optimality. In this section, we highlight the main message, focusing on the choice between testing all patients or only those with an A signal, and we relegate full derivations to the appendices.

We denote with a star the minimum transfers satisfying the participation constraints (equation 8) at the second-best effort levels described in Proposition 5 (see Definition 2 in Appendix 9.5.1 for their specific values). Figure A.2 in Appendix 9.5.1 shows that, under these transfer values, only type-L physicians receive positive rents, in Cases 0 and 1. This arises because effort costs are convex, whereas payments (eq. 21) are linear in effort, so the binding constraint is that of type-H physicians.

We then compare the equilibrium testing thresholds with these transfers, $z_{All}^{eq}(\alpha_L, T_{All}^{*}, T_{L,1}^{*})$, with their second-best levels, $z_{All}^{SB}(\alpha_i)$, and obtain the following ranking.

Lemma 2 *(a)* $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1})$. *(b)* $z^{SB}_{All}(\alpha_i) > z^{eq}_{All}(\alpha_i, T^*_{All}, T^*_{i,1}), \forall i = \{L, H\}$.

Proof. See Appendix 9.5.2. ■

Part (a) shows that less altruistic physicians cease testing all patients at a lower test-cost threshold. This reflects their incentive to earn a rent by testing only A-signal patients, whereas testing everyone yields no rent. Part (b) indicates that, relative to the second-best optimum, physicians switch too early from testing all patients (Case *All*) to testing only A-signal patients (Case 1). This occurs even for type-H physicians, who earn no rents in either case, because they do not internalize the higher second-best effort required in the transition from Case *All* to Case 1, being fully compensated for it by the monetary transfer received. Type-L physicians have an additional incentive to under-test, as they earn strictly positive rents in Case 1 but none in Case *All*.

To adjust transfers from Definition 2 in order to decentralize the second-best testing decision, we further need to rank $z^{SB}_{All}(\alpha_i)$ by α_i. As shown in Appendix 9.5.3, this ranking is in general ambiguous. For tractability, we consider the case where the threshold is independent of α_i by making the following assumption:

Assumption 4 *There is an equal proportion of type-H and type-L physicians:* $\nu = 1/2$.

More general cases are treated in Appendix 9.5.6, where we show that, although the exposition is more complex, the qualitative results are unchanged. Assumption 4 enables the following ranking of thresholds.

Lemma 3 *Under Assumptions 3 and 4, we have:*

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

Proof. Appendix 9.5.3 proves that Assumption 4 is a sufficient condition to insure that z^{SB}_{All} is independent of α_i. The rest of the proof results from Lemma 2. ■

We now study how to decentralize the second-best testing decisions, starting from the fixed part of the transfers studied in this subsection.

Lemma 3 (and its equivalent for z_1, Lemma 4 in Appendix 9.5.4) show that, at the transfer levels specified in Definition 2, the equilibrium test-cost thresholds are strictly below their second-best counterparts. Consequently, there exist values of z for which at least one physician type makes a suboptimal testing decision -for instance, testing only A-signal patients when the second-best calls for testing all patients. Importantly, no pooling contract can equalize the equilibrium and second-best thresholds z_{All} and z_1 for both physician types simultaneously. Such alignment, however, is not required to induce second-best testing decisions for both types, provided that the fixed part of the transfers can be conditioned on z, as we demonstrate in Figure 1.

Figure 1 reports the rent-minimizing fixed parts of transfers, $T_{All}(z)$, $\bar{T}_1(z)$ and $\bar{T}_0(z)$, which decentralize the second-best testing decisions while satisfying participation constraints. When $z \leqslant z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1})$, both physician types optimally test all patients, so the transfers in Definition 2 suffice. When $z^{eq}_{All}(\alpha_L, T^*_{All}, T^*_{L,1}) < z \leqslant z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1})$, type-$L$ physicians test only A-signal patients, which is suboptimal. To induce them to continue testing all patients, $T_{All}(z)$ must be raised, as shown in Figure 1. Participation constraints preclude reducing transfers in Case 1 instead. This higher $T_{All}(z)$ also shifts the equilibrium threshold z_{All} for type-H physicians upward, but their testing decision remains second-best. A similar logic applies when $z^{eq}_{All}(\alpha_H, T^*_{All}, T^*_{H,1}) < z \leqslant z^{SB}_{All}$, where $T_{All}(z)$ is increased above T^*_{All} to induce both types to keep on testing all patients.[27]

Once z exceeds z^{SB}_{All}, a similar logic applies. As long as $z^{SB}_{All} < z \leqslant z^{eq}_1(\alpha_L, T^*_{L,1}, T^*_{L,0})$, no adjustment to the transfers in Definition 2 is required, with the transfer in Case *All* reverting to T^*_{All} as z reaches z^{SB}_{All}. When $z^{eq}_1(\alpha_L, T^*_{L,1}, T^*_{L,0}) < z \leqslant z^{eq}_1(\alpha_H, T^*_{H,1}, T^*_{H,0})$ (resp. $z^{eq}_1(\alpha_H, T^*_{H,1}, T^*_{H,0}) < z \leqslant z^{SB}_1$), it becomes necessary to incentivize type-L (resp. both types of) physicians to test A-signal patients by raising the fixed part of the transfer in Case 1, $\bar{T}_1(z)$. Finally, when $z > z^{SB}_1$, no physician tests any patient, which is socially optimal. Accordingly, all three fixed parts of the transfers return to the levels defined in Definition 2.

[27] In this case, the minimum $T_{All}(z)$ required to induce both physicians to test all patients differs across types. Since the second-best contract is pooling, the social planner selects the larger of the two, as shown in Proposition 7 in Appendix 9.5.5.

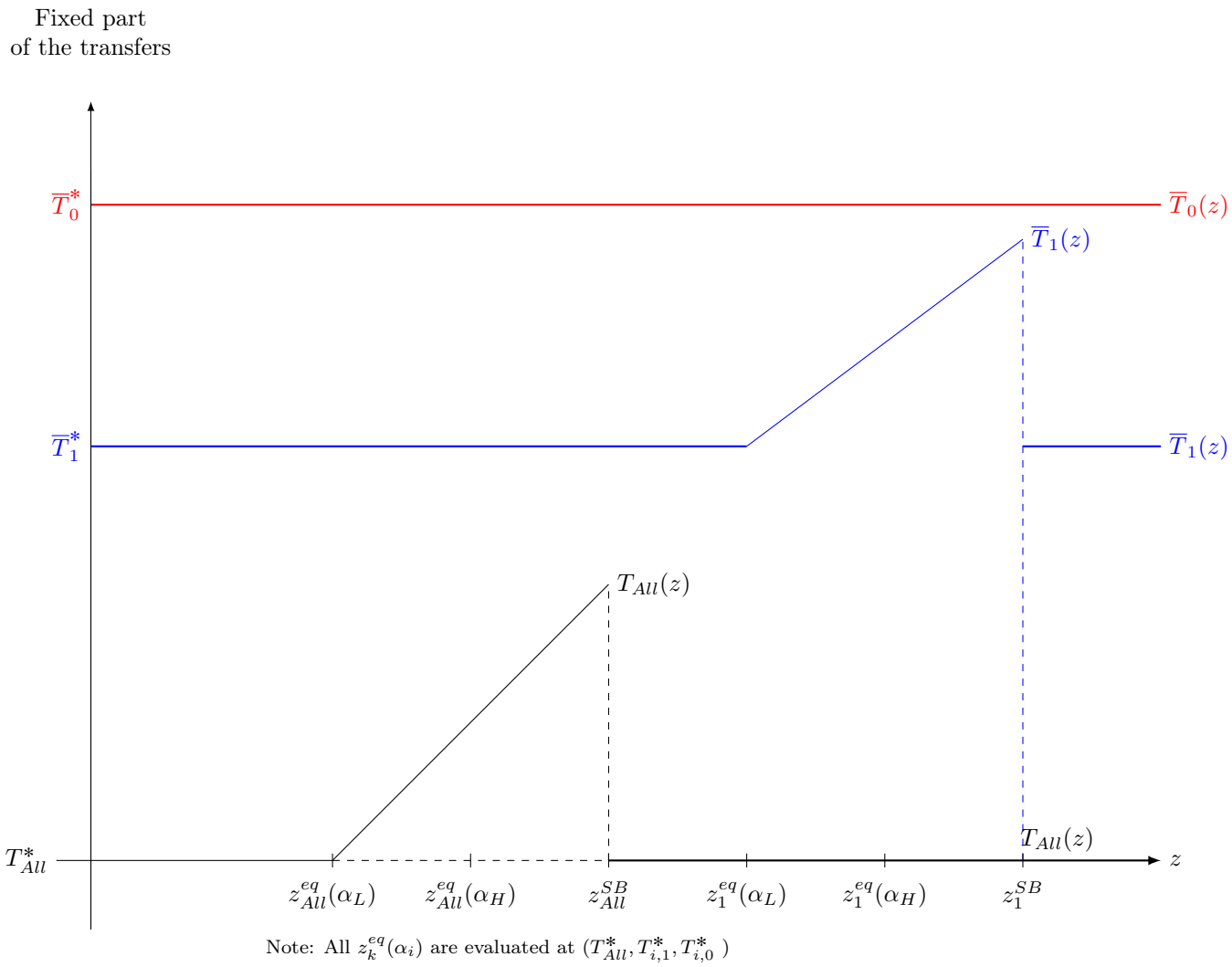

Figure 1: Levels of the fixed part of payments: $T_{All}(z)$, $\overline{T}_1(z)$ and $\overline{T}_0(z)$.

The main message of this section (presented analytically in Proposition 7 in Appendix 9.5.5) is then that second-best testing decisions, along with effort, can be decentralized by appropriately designing physicians' payments, provided these can be conditioned on the test cost z.[28]

8 Conclusion

This manuscript examines how imperfectly altruistic physicians choose diagnostic effort and testing when selecting between two treatments. We first derive the first-best allocation, where the regulation authority can observe the physicians' degree of altruism, and compare it with a situation where fixed (*i.e.* independent of physicians' decisions) transfers are set just high enough to satisfy their participation constraints. In this case, physicians exert insufficient effort and underutilize tests.

We then consider the second-best allocation, where physicians' levels of altruism are unobservable. We demonstrate that the second-best contract is a *pooling* contract, offering the *same* P4P and fixed part 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 physicians. In this scenario, the slope of the contract's P4P component should be based on average altruism, resulting in high-altruism physicians exerting more effort than low-altruism ones. Additionally, we show that if the fixed parts of the transfers are set in each case at the minimum level required for physicians participation, they will under-utilize diagnostic tests. Interestingly, decentralizing the second-best outcome requires making the fixed parts of the transfers dependent on the test cost. This implies that as new technologies emerge and diagnostic test costs vary over time, the regulation authority should adjust physicians' remunerations accordingly.

Our model highlights several avenues for future research. A first avenue concerns the (in)observability of treatment outcomes and its implications for outcome-based P4P schemes. We have assumed symmetric and observable consequences of treatment mismatches; in practice, treating regular patients with advanced therapies-though inefficient- may not increase physician

[28] In Appendix 9.5.5, we show that our results are robust to two key assumptions underlying Figure 1, (i) that $z_{All}^{SB} < z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*)$, and (ii) that at the second-best, both physician types always make the same testing decision (see also the online Appendix 9.5.6 for the latter).

visits and thus remain undetected, leaving only one type of mismatch observable. Real-world P4P programs, such as the UK's Quality and Outcomes Framework and the US Medicare value-based payment programs, provide a natural empirical context for our framework. While these contracts can mitigate moral hazard by rewarding diagnostic accuracy, imperfect observability and measurement noise may distort incentives, potentially encouraging risk-averse behavior, cherry-picking of patients, or excessive testing. When diagnostic tests serve as substitutes for costly effort, less altruistic physicians may respond to outcome-based incentives by over-testing to secure performance targets, underscoring the need for careful policy design that balances accuracy, efficiency, and fairness.

A second area for exploration concerns physician altruism. Although modeled as an unobservable trait, altruism likely correlates with observable characteristics and career choices. Future empirical work could proxy altruism through self-reported motivations, willingness to work in underserved areas, participation in public or pro bono care, or differential responsiveness to financial versus non-financial rewards. Relatedly, endogenous selection into the physician pool could affect participation: if more altruistic doctors have higher reservation utilities, they may be harder to retain, potentially leading to shortages or longer waiting times. We maintain here a no-shutdown assumption to ensure participation of all types, but relaxing it and allowing for self-selection across institutional settings (for instance between public and private facilities) would be a promising extension (see for instance Barigozzi *et al.* (2025) who model two firms, one for-profit and one mission-oriented).

Finally, for reasons of tractability, our analysis deliberately focuses on a simple framework. Nevertheless, we discuss in the paper (after Proposition 4) several aspects of the robustness of the non-responsiveness result to alternative assumptions. Future research could extend this analysis by incorporating richer environments with multi-dimensional heterogeneity among physicians, correlated signals, screening menus trading off risk and reward, or long-term and repeated interactions shaping reputation and type revelation.

9 Appendix

9.1 Imperfect diagnostic test

In this Appendix, we generalize our model to the case where the diagnostic test is imperfect. We assume that the precision of the test is given by $1/2 < q \leqslant 1$, which represents the probability that the test delivers a result in line with the true type of the patient tested. We keep the symmetry assumption used in the paper, and assume that this precision is the same for both types. We further assume that the two signals (the one obtained thanks to the diagnostic effort ε, and the one obtained from the diagnostic test) are conditionally independent.

The sequence of the model is identical to our baseline model. Physicians decide how much effort ε to exert, observe the signal, and decide whether to run the diagnostic test or not. They are aware of the precision of both signals when taking their decision. We solve the model as in the text, starting with the social optimum. Table A.1 updates Table 1 and reports the frequencies of the triplet (true type, signal from diagnostic effort, signal from diagnostic test) in the population.

Type	**Signals**	**Frequency in population**
A	(A, A)	$\varepsilon q/2$
A	(A, B)	$\varepsilon(1-q)/2$
A	(B, A)	$(1-\varepsilon)q/2$
A	(B, B)	$(1-\varepsilon)(1-q)/2$
B	(A, A)	$(1-\varepsilon)(1-q)/2$
B	(A, B)	$(1-\varepsilon)q/2$
B	(B, A)	$\varepsilon(1-q)/2$
B	(B, B)	$\varepsilon q/2$

Table A.1: Population frequencies with imperfect test

We have to decide what the treatment will be for each of the possible signals received. Observe that the diagnostic test, being costly, can only be prescribed at equilibrium if the treatment decision can be affected by the test signal. For instance, if the test is prescribed after having observed the first signal A, then it must be the case that the prescription decision must

vary according to the second signal, otherwise, we would save the test cost by not prescribing it in that case. This, in turn, means that, when a test is run, the physician always follows the signal the test produces. In other words, the first signal (obtained from effort) is used to decide whether to run the test. If the test is not run, then the patient is treated according to the first signal received. If the test is run, the treatment decision depends only on the test signal.

Table A.2 then summarizes the treatment decisions made following the reception of the effort signal and, when applicable, the test result.

Testing decision after first signal	**Signals**	**Treatment**
Test All	(A, A)	P
Test All	(A, B)	D
Test All	(B, A)	P
Test All	(B, B)	D
Test only A	(A, A)	P
Test only A	(A, B)	D
Test only A	$(B, \varnothing)$	D
Test only B	(B, A)	P
Test only B	(B, B)	D
Test only B	$(A, \varnothing)$	P
No Test	$(A, \varnothing)$	P
Test only	$(B, \varnothing)$	D

Table A.2: Treatment decision after observing effort and test signals.

We now compute the welfare function corresponding to the four possible cases in the preceding table, using both the frequencies in the populations, and the decision rule used.

When every patient is tested, the aggregate welfare is:

$$W_{All}(\varepsilon_{All}) = -\psi(\varepsilon_{\text{All}}) + \frac{1}{2}\left[\, q(U_A^P + U_B^D) + (1-q)(U_A^D + U_B^P) \,\right] - z.$$

This formulation naturally reduces to the benchmark case presented in the main text when $q = 1$. More importantly, the result that diagnostic effort is useless remains valid for all values of q. This is because the physician's treatment decision is always based solely on the test result, regardless of the signal obtained from effort. The rationale for this behavior differs from that

in the main text: it is no longer due to the test being perfect, but rather stems from a strategic behavior. Specifically, a test is only ordered if it has the potential to alter the treatment decision regardless of the first signal received. This implies that the treatment decision is entirely dictated by the test result, which in turn renders diagnostic effort redundant.

Alternatively, when the test is prescribed only when the first signal is A, we obtain

$$\begin{aligned} W_1(\varepsilon_1) &= -\psi(\varepsilon_1) + \frac{1}{2}\left[U_A^P \varepsilon q + U_B^D(\varepsilon + (1-\varepsilon)q) \right. \\ &\quad \left. + U_A^D(1-\varepsilon+\varepsilon(1-q)) + U_B^P(1-\varepsilon)(1-q) \right] - \frac{z}{2}, \end{aligned}$$

where we made use of the proportions in Table A.2 . The first-order condition for ε_1 is

$$\psi'(\varepsilon_1^*) = \frac{1}{2}\left[q\Delta U_A + (1-q)\Delta U_B \right]. \tag{A.1}$$

Since we have that $\Delta U_A < \Delta U_B$, we have that $\partial \varepsilon_1^* / \partial q < 0$. As is intuitive, a lower effort ε is required when the second information source is more precise.

Finally, the case where the test is not run on anyone is the same as in the main text, so that ε_0^* is the same and is not affected by q, as is intuitive since the test plays no role in this case.

Comparing the first-order conditions (eq. 3) for ε_0^* and (eq. A.1) for ε_1^*, we see that $\varepsilon_1^*(q) < \varepsilon_0^* \, \forall q \in [0,1]$, so that Proposition 1 still holds- *i.e.*, effort and test are strategic substitutes.

The following results are also robust to the introduction of an imperfect test:

1) Lemma 1: testing only B-signal patients is dominated by testing only A-signal patients.

Proof.

When testing only B-signal patients, total welfare is equal to

$$\begin{aligned} W_{1B}(\varepsilon_{1B}) &= -\psi(\varepsilon_{1B}) + \frac{1}{2}\left[U_A^P(\varepsilon + (1-\varepsilon)q) + U_B^D \varepsilon q \right. \\ &\quad \left. + U_A^D(1-\varepsilon)(1-q) + U_B^P(1-\varepsilon+\varepsilon(1-q)) \right] - \frac{z}{2}. \end{aligned}$$

We now show that $W_{1B}(\varepsilon) < W_1(\varepsilon), \ \forall \varepsilon$. This is the case if and only if

$$\begin{aligned} & U_A^P\left(\varepsilon + (1-\varepsilon)q\right) + U_B^D\,\varepsilon q + U_A^D\,(1-\varepsilon)(1-q) + U_B^P\,(1-\varepsilon+\varepsilon(1-q)) \\ < \;& U_A^P\,\varepsilon q + U_B^D\,(\varepsilon + (1-\varepsilon)q) + U_A^D\,(1-\varepsilon+\varepsilon(1-q)) + U_B^P\,(1-\varepsilon)(1-q) \\ \Leftrightarrow \;& X(U_A^P - U_B^D) + Y(U_A^D - U_B^P) < 0, \end{aligned}$$

where $X = \varepsilon + q - 2\varepsilon q = -Y > 0$ so that

$$X(U_A^P - U_B^D) + Y(U_A^D - U_B^P) < 0 \Leftrightarrow \quad X(\Delta U_A - \Delta U_B) < 0,$$

which always holds under Assumption 1. ■

2) Computing the thresholds z_{All}^* and z_1^* yields

$$z_{All}^* = q(1-\varepsilon_1^*)\Delta U_A - \varepsilon_1^*(1-q)\Delta U_B + 2\psi(\varepsilon_1^*),$$

so that

$$\frac{\partial z_{All}^*}{\partial q} = (1-\varepsilon_1^*)\Delta U_A + \varepsilon_1^*\Delta U_B > 0,$$

where we made use of the envelope theorem. The intuition for this variation is straightforward, as a lower value of q decreases more the welfare level attained when more patients are tested (*i.e.*, in Case *All* compared to Case 1) and thus, makes testing all patients less attractive. We then switch to testing only patients with an A-signal for a lower threshold test cost value z_{All}. In other words, a less informative test should be used less often.

Similarly for z_1^*, we obtain:

$$z_1^* = 2(\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*)) - (\varepsilon_0^* - q\varepsilon_1^*)\Delta U_A + (q(1-\varepsilon_1^*) - (\varepsilon_0^* - \varepsilon_1^*))\Delta U_B,$$

so that

$$\frac{\partial z_1^*}{\partial q} = (1-\varepsilon_1^*)\Delta U_B + \varepsilon_1^*\Delta U_A > 0, \tag{A.2}$$

where we made use of the envelope theorem. The intuition is similar: we stop using the test for lower test cost values when the test is less informative.

The rest of the analysis follows the same structure as in the body of the paper, with more complex computations due to the introduction of q, while the analysis remains robust to the case $q < 1$.

Finally, note that the use of the test requires $q > \varepsilon^*$. Indeed, (i) for the test to be used, it must influence the treatment decision, and (ii) it cannot be optimal to rely on a test that is less precise than the effort technology. Then, our analysis remains robust even when the test is imperfect, provided that the test is sufficiently accurate to add value relative to diagnostic effort.

9.2 Proof of Proposition 3

We first have to find the values of the fixed 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 (17) and (20) allows us to obtain the optimal fixed part in Case 1, namely

$$\bar{T}_1 = \bar{T}_{All} + (1-\alpha)[\psi(\varepsilon_1^{*}) - \Delta U_A n_1^{*}], \tag{A.3}$$

where $n_1^{*} = (1+\varepsilon_1^{*})/2$ is the proportion of correctly treated patients in Case 1 when the optimum effort level is exerted. Using equations (18) and (19), we obtain

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

Equations (20) and (A.3) determine the value of the optimal fixed part level in Case 0:

$$\bar{T}_0 = \bar{T}_{All} + (1-\alpha)[\psi(\varepsilon_0^{*}) - \frac{\Delta U_B + \Delta U_A}{2} n_0^{*}], \tag{A.4}$$

where $n_0^{*} = \varepsilon_0^{*}$ is the proportion of correctly treated patients in Case 0 when the optimal effort level is exerted.

Observe that using the first-order conditions for the optimum (3) and (5), the bracket term in both equations (A.3) and (A.4) can be rewritten as $\psi(\varepsilon_1^{*}) - \psi'(\varepsilon_1^{*})(1+\varepsilon_1^{*})$ and $\psi(\varepsilon_0^{*}) - \psi'(\varepsilon_0^{*})\varepsilon_0^{*}$, respectively. 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.3) and (A.4) in the above expressions, and setting transfers at

the lowest levels in accordance with Assumption 2, 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 fixed part of the physicians' payment scheme in Cases 0 and 1:

$$\bar{T}_0 = \psi(\varepsilon_0^*) - (1-\alpha)\frac{\Delta U_B + \Delta U_A}{2} n_0^*, \tag{A.5}$$

$$\bar{T}_1 = \psi(\varepsilon_1^*) - (1-\alpha)\Delta U_A n_1^* + \alpha[\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*)]. \tag{A.6}$$

Note that the fixed part levels are increasing in α and are negative when α is low enough. We also see that the rent obtained by doctors increases with their degree of altruism both in Case *All* and Case 1 (with zero rent by construction in Case 0), with

$$\begin{aligned} \bar{T}_{All} &= \alpha\psi(\varepsilon_0^*), \\ T_1(n_1^*) &= \psi(\varepsilon_1^*) + \alpha(\psi(\varepsilon_0^*) - \psi(\varepsilon_1^*)), \\ T_0(n_0^*) &= \psi(\varepsilon_0^*). \end{aligned} \tag{A.7}$$

9.3 Proof of Proposition 4

Recall that the proportion of correctly treated patients in Cases 1 and 0 are respectively: $n_{i,1} = (1+\varepsilon_{i,1})/2$ and $n_{i,0} = \varepsilon_{i,0}$. Assuming that each physician selects the contract designed for her, her utility is given by:

$$V_{i,k} = \alpha_i B_k(\varepsilon_{i,k}) + \bar{T}_{i,k} + \beta_{i,k} n_{i,k} - \psi(\varepsilon_{i,k}). \tag{A.8}$$

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

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

while in Case 0, it is equal to

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

The equilibrium levels of efforts are obtained as a solution to the maximization of the physicians' utility $V_{i,k}$ (A.8) with respect to $\varepsilon_{i,k}$:

$$\psi'(\varepsilon_{i,k}) = \alpha_i \bar{b}_k + \mathbb{1}_k \beta_{i,k}, \tag{A.9}$$

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 $\varepsilon_{i,k}$:

$$\bar{b}_1 = \frac{\Delta U_A}{2}, \tag{A.10}$$

$$\bar{b}_0 = \frac{\Delta U_A + \Delta U_B}{2}. \tag{A.11}$$

The proof of Proposition 4 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 has to be proposed to both physicians' types at the second-best equilibrium.

9.3.1 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 A.1 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.

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, $(\bar{T}_{H,k}, \beta_{H,k})$:

$$\alpha_L B_k(\varepsilon_{L,k}) + \bar{T}_{L,k} + \beta_{L,k} n_{L,k} - \psi(\varepsilon_{L,k}) = \alpha_L B_k(\tilde{\varepsilon}_{L,k}) + \bar{T}_{H,k} + \beta_{H,k} \tilde{n}_{L,k} - \psi(\tilde{\varepsilon}_{L,k}), \quad \text{(A.12)}$$

where we use a tilde to denote the allocation where a physician mimics the other type by taking the contract designed for the latter. More precisely, $\varepsilon_{L,k}$ is defined by (A.9) 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[29]

$$\psi'(\tilde{\varepsilon}_{i,k}) \quad = \quad \alpha_i \bar{b}_k + \mathbb{1}_k \beta_{j,k}, \tag{A.13}$$

with $i \neq j$, $\bar{b}_k$ defined by (A.10) and (A.11), and $\mathbb{1}_1 = 1/2$ and $\mathbb{1}_0 = 1$. In turn, $\tilde{n}_{L,k}$ refers to the proportion 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.12), 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

$$\alpha_H B_k(\tilde{\varepsilon}_{H,k}) + \bar{T}_{L,k} + \beta_{L,k} \tilde{n}_{H,k} - \psi(\tilde{\varepsilon}_{H,k}) > \alpha_H B_k(\varepsilon_{H,k}) + \bar{T}_{H,k} + \beta_{H,k} n_{H,k} - \psi(\varepsilon_{H,k}), \tag{A.14}$$

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

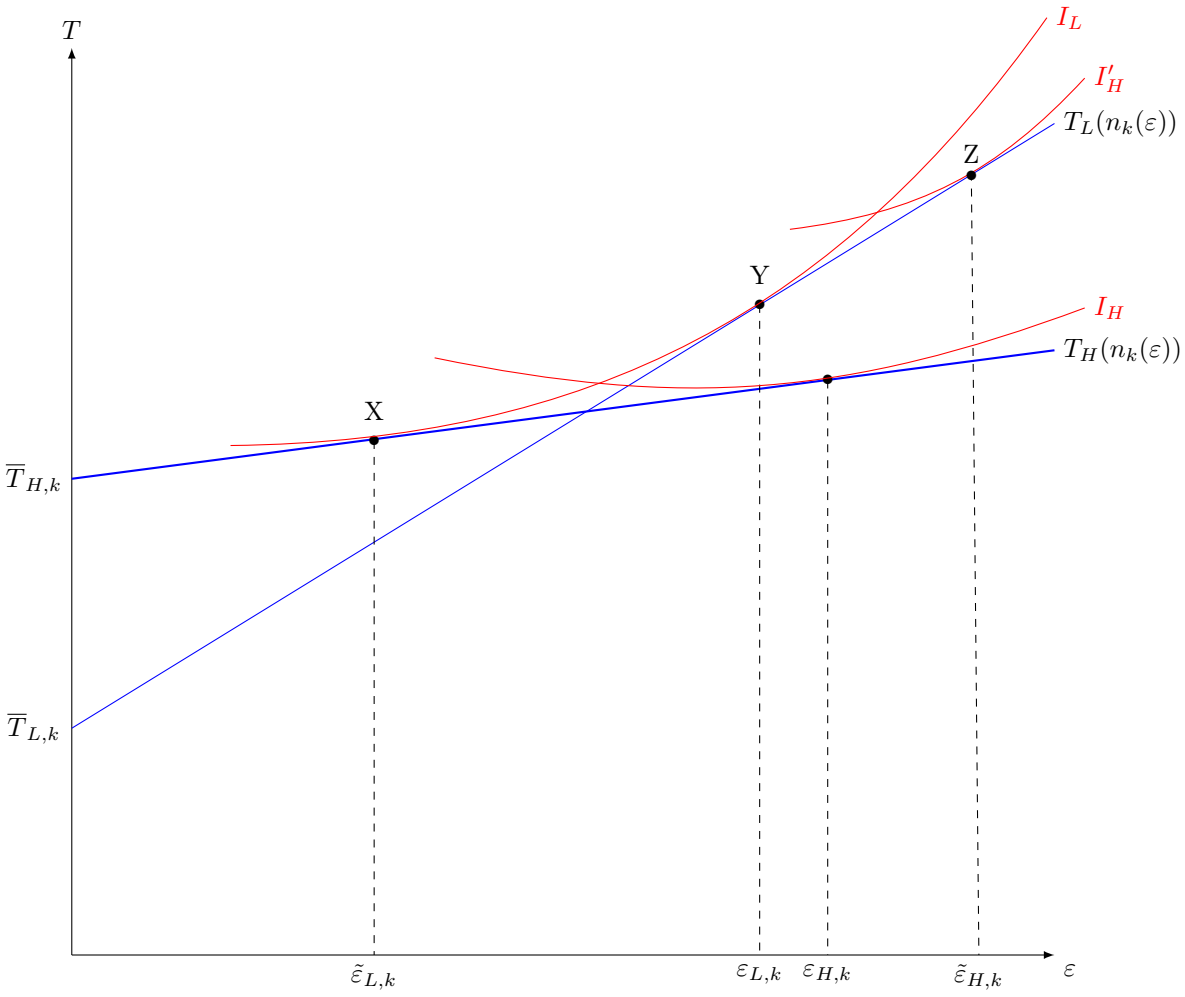

Figure A.1: Second-best pooling equilibrium (contradiction argument)

with $\varepsilon_{H,k}$ and $\tilde{\varepsilon}_{H,k}$ defined by equations (A.9) and (A.13). 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:

$$\alpha_H[\tilde{\varepsilon}_{H,k}\bar{b}_k - \varepsilon_{H,k}\bar{b}_k] - [\psi(\tilde{\varepsilon}_{H,k}) - \psi(\varepsilon_{H,k})] - \{\alpha_L[\varepsilon_{L,k}\bar{b}_k - \tilde{\varepsilon}_{L,k}\bar{b}_k] - [\psi(\varepsilon_{L,k}) - \psi(\tilde{\varepsilon}_{L,k})]\}$$
$$> \beta_{L,k}(n_{L,k} - \tilde{n}_{H,k}) - \beta_{H,k}(\tilde{n}_{L,k} - n_{H,k}) \quad \text{(A.15)}$$

Isolating $\bar{b}_k$ in equations (A.9) and (A.13) and replacing for their expression in (A.15), 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 3), 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 (A.9) and (A.13), 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.14) 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.

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

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 (3) and (5); (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 (A.9) that $\varepsilon_{i,k}$ is monotonically increasing in $\beta_{i,k}$ (for $i \in \{L, H\}$), (ii) from equation (A.13) 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 $\beta_{L,k} < \beta_{H,k}$.

9.4 Proof of Proposition 5

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

$$\max_{\beta_k} SW = \nu[B_k(\varepsilon_{L,k}) - \psi(\varepsilon_{L,k})] + (1-\nu))[B_k(\varepsilon_{H,k}) - \psi(\varepsilon_{H,k})], \tag{A.16}$$

where $\varepsilon_{i,k}$ is chosen by type i-physicians and satisfies equation (A.9).

Differentiating (A.16) 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 (A.9)). Using Assumption 3 as well as the expression of $\psi'(\varepsilon_{i,k})$ in (A.9), 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 (22) and (23).

We now find the second-best levels of effort by replacing for the values of β_k^{SB} in the first-order condition (A.9):

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

Under Assumption 3, we obtain that:

$$\varepsilon_{H,k}^{SB} = (\alpha_H + 1 - \bar{\alpha})\bar{b}_k + \frac{1}{2}, \tag{A.17}$$

$$\varepsilon_{L,k}^{SB} = (\alpha_L + 1 - \bar{\alpha})\bar{b}_k + \frac{1}{2}. \tag{A.18}$$

Comparing these equations to (3) and (5), we also obtain that ε_k^* is defined by $\psi'(\varepsilon_k^*) = \varepsilon_k^* - 1/2 = \bar{b}_k$. Straightforward algebra then shows that $\varepsilon_{L,k}^{SB} < \varepsilon_k^* < \varepsilon_{H,k}^{SB}\ \forall k = \{0,1\}$.

9.5 The optimal fixed parts of payments decentralizing the second-best testing decisions

9.5.1 Rent-minimizing transfers

Definition 2 *The set of transfers of the form (21) 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:

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

In Case *All*, no effort is exerted, so the minimum transfer consistent with participation is a zero fixed payment for both physician types. In Cases 0 and Case 1, type-H physicians exert more effort than type-L physicians (see Proposition 5). Because the effort cost is convex whereas the variable part of the reimbursement is linear in effort, the participation constraint is tighter for type-H than for type-L physicians (see Figure A.2). Minimizing financial rents then implies 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). From the zero rent for type-H physicians, one deduces the fixed part of the transfer (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.

Figure A.2 shows the rents accruing to physicians of both types in Case $k = \{0, 1\}$ with the transfers in Definition 2.

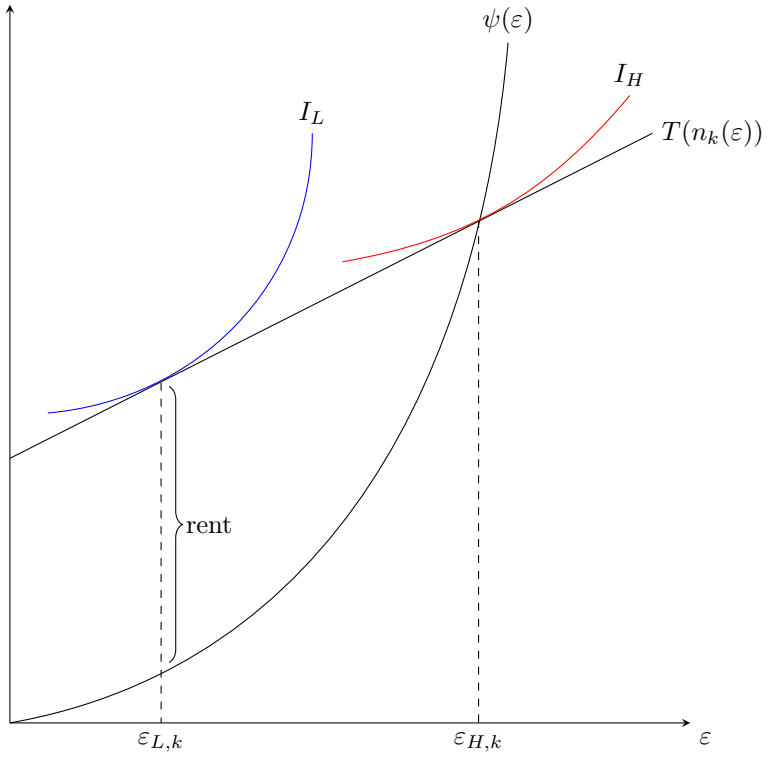

Figure A.2: Measuring rents in the second-best allocation

9.5.2 Proof of Lemma 2

The second-best level of z_{All} (which takes a form similar to equation (6)) is independent of the transfer values and is computed given the doctors' second-best effort choices. It thus depends on α_i only indirectly, through the individual choice of $\varepsilon_{i,1}^{SB}$:

$$z_{All}^{SB}(\alpha_i) \equiv (1 - \varepsilon_{i,1}^{SB})\Delta U_A + 2\psi(\varepsilon_{i,1}^{SB}). \tag{A.19}$$

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

$$z_{All}^{eq}(\alpha_i, T_{All}, T_{i,1}) = (1 - \varepsilon_{i,1}^{SB})\Delta U_A + \frac{2}{\alpha_i}(T_{All} - T_{i,1}) + \frac{2}{\alpha_i}\psi(\varepsilon_{i,1}^{SB}). \tag{A.20}$$

Note that a doctor's degree of altruism affects z_{All}^{eq} both directly and indirectly through their

effort choice $\varepsilon_{i,1}^{SB}$. We now prove Lemma 2.[30]

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

$$\frac{dz_{All}^{eq}(\alpha_i, T_{All}, T_{i,1})}{d\alpha_i} = \frac{d\varepsilon_{i,1}^{SB}}{d\alpha_i}\left[\frac{2}{\alpha_i}\psi'(\varepsilon_{i,1}^{SB}) - \Delta U_A\right] - \frac{2}{\alpha_i^2}(T_{All} - T_{i,1} + \psi(\varepsilon_{i,1}^{SB})) - \frac{2}{\alpha_i}\frac{dT_{i,1}}{d\alpha_i}, \tag{A.21}$$

with $T_{i,1}$ given by (21), so that

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

Together with the first-order condition (A.9) on effort, equation (A.21) 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 $(T_{All} = T_{All}^*, T_{L,1} = T_{L,1}^*)$.

(b) Comparing expressions (A.19) and (A.20), we obtain that the statement holds if and only if

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

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.

9.5.3 Proof of Lemma 3

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

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

Hence, $z_{All}^{SB}(\alpha_i)$ is a U-shaped function of α_i, with a minimum in $\bar{\alpha}$ so that the ranking between $z_{All}^{SB}(\alpha_H)$ and $z_{All}^{SB}(\alpha_L)$ is *a priori* ambiguous.

[30] We sometimes abuse notation and proceed as if there were a continuum of α_i levels. This shows that our approach can be generalized to a continuum of altruism types.

Replacing for the expressions of $\varepsilon_{i,k}^{SB}$ (defined by (A.17) and (A.18)), for the expression of $\bar{b}_1$ (defined by (A.10)) as well as for the functional form of $\psi(.)$ (Assumption 3) in equation (A.19), we obtain:

$$z_{All}^{SB}(\alpha_i) = \Delta U_A - 2\bar{b}_1[\bar{b}_1(1+\alpha_i-\bar{\alpha})+1/2] + [\bar{b}_1(1+\alpha_i-\bar{\alpha})]^2.$$

Rearranging terms, this expression simplifies to

$$z_{All}^{SB}(\alpha_i) = \frac{\Delta U_A}{2} + \bar{b}_1^2(1+\alpha_i-\bar{\alpha})(\alpha_i-\bar{\alpha}-1). \tag{A.22}$$

Under Assumption 4, $\bar{\alpha} = \frac{1}{2}\alpha_H + \frac{1}{2}\alpha_L$ so that

$$(1+\alpha_i-\bar{\alpha})(\alpha_i-\bar{\alpha}-1) = (1+\frac{1}{2}(\alpha_H-\alpha_L))(\frac{1}{2}(\alpha_H-\alpha_L)-1),$$

for both α_H and α_L. Hence, $z_{All}^{SB}(\alpha_H) = z_{All}^{SB}(\alpha_L)$. To simplify notation, we omit the α_i-argument in z_{All}^{SB} as stated in Lemma 3.

The influence of the physicians' altruism degree (α_L or α_H) on their second-best threshold z_{All}^{SB} is entirely driven by their second-best effort level, as shown in equation (A.19). A higher effort level has dual effects: a benefit (it reduces diagnostic errors, as shown in the first term in (A.19)), thereby lowering the optimal threshold z_{All}^{SB} and a cost, which increases the threshold z_{All}^{SB} (second term in (A.19)). The net effect on z_{All}^{SB} of a higher second-best effort by type H compared to type L is generally indeterminate. However, under Assumption 4, this net effect is zero and both types of physicians should optimally switch from testing all patients to testing only signal-A patients at the same test cost threshold.

9.5.4 Proof of Lemma 4

The following lemma is the equivalent of Lemma 3 for Case 1.

Lemma 4 *Under Assumptions 3 and 4, we have*

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

To prove the above lemma, we start by comparing equilibrium threshold cost levels. We have:

$$\begin{aligned} z_1^{eq}(\alpha_i, T_{i,1}, T_{i,0}) &= (1-\varepsilon_{i,0}^{SB})\Delta U_B - (\varepsilon_{i,0}^{SB} - \varepsilon_{i,1}^{SB})\Delta U_A \\ &+ \frac{2}{\alpha_i}(T_{i,1} - T_{i,0}) + \frac{2}{\alpha_i}(\psi(\varepsilon_{i,0}^{SB}) - \psi(\varepsilon_{i,1}^{SB})). \end{aligned} \tag{A.23}$$

Differentiating equation (A.23) with respect to α_i and using the envelope theorem for $\varepsilon_{i,k}^{SB}$ (*i.e.* equation (A.9)), as well as $T_{i,k}$ given by (21) with β_k^{SB} defined by (22) and (23), 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,

$$\frac{dz_1^{eq}(\alpha_i, T_{i,1}, T_{i,0})}{d\alpha_i} = -\frac{2}{\alpha_i^2}[(T_{i,1} - \psi(\varepsilon_{i,1}^{SB})) - (T_{i,0} - \psi(\varepsilon_{i,0}^{SB}))]. \tag{A.24}$$

In order to sign this expression, we define the square bracket in the right-hand side of equation (A.24) 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:

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

using equation (A.9) together with Assumption 3, 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 (obtained from the transfers 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(a) 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 level of effort, defined by:

$$z_1^{SB}(\alpha_i) \equiv 2\left(\psi(\varepsilon_{i,0}^{SB}) - \psi(\varepsilon_{i,1}^{SB})\right) + \left(1 - \varepsilon_{i,0}^{SB}\right)\Delta U_B - \left(\varepsilon_{i,0}^{SB} - \varepsilon_{i,1}^{SB}\right)\Delta U_A. \tag{A.25}$$

Comparing equations (A.23) with (A.25), 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 2(b) holds for the comparison between $z_1^{SB}(\alpha_i)$ and $z_1^{eq}(\alpha_i, T_{i,1}^*, T_{i,0}^*)$ as well.

Proceeding in the same way as for the computation of $z_{All}^{SB}(\alpha_i)$ (equation A.22), we now show that $z_1^{SB}(\alpha_i)$ is independent of α_i. Using the expressions (A.17) and (A.18) of $\varepsilon_{i,k}^{SB}$, the expressions of $\bar{b}_0$ and $\bar{b}_1$ (equations (A.10) and (A.11)) together with Assumption 3, we obtain that

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

As already shown the last term is invariant to α_i under Assumption 4 (see Appendix 9.5.3), we obtain that $z_1^{SB}(\alpha_H) = z_1^{SB}(\alpha_L)$, so that the independence of second-best test cost thresholds with respect to the degree of altruism 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 Lemma 4.

9.5.5 Summary and robustness of Section 7.2

Our results are summarized in the following proposition. We denote by $T_{i,All}(z)$ the minimum transfer level that ensures that type-i physicians effectively test all patients, given that $T_{i,1}^*$ is set according to Definition 2. This means $z_{All}^{eq}(\alpha_i, T_{i,All}(z), T_{i,1}^*) = z$. Similarly, $T_{i,1}(z)$ represents the minimum transfer level that ensures that, for type i, $z_1^{eq}(\alpha_i, T_{i,1}(z), T_{i,0}^*) = z$, with $T_{i,0}^*$ set as in Definition 2.

Proposition 7 *Assume that the social planner can condition the transfers on the observed level of the cost of the diagnostic test, z. Under Assumptions 3 and 4 and asymmetric information on the α_i's, the rent-minimizing payment received by the physicians should be set at $T_{All}(z) = T_{All}^* = 0$, $T_{i,k}(z) = T_{i,k}^*$, $\forall k \in \{1,0\}$ and $\forall i \in \{L,H\}$, except in the following cases:*

*A. If it is optimal to test all agents (*i.e. $z < z_{All}^{SB}$*), then*

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

where

$$T_{i,All}(z) = T_{i,1}^* - \psi(\varepsilon_{i,1}^{SB}) + \frac{\alpha_i}{2}[z - (1 - \varepsilon_{i,1}^{SB})\Delta U_A]. \tag{A.26}$$

*B. If it is optimal to test only patients with an A signal (*i.e. *when* $z_{All}^{SB} < z < z_1^{SB}$*), then*

$$T_1(z) = \begin{cases} T_{L,1}(z), & \text{when } z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*) < z \leqslant z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*) \\ max\{T_{L,1}(z), T_{H,1}(z)\}, & \text{when } z_1^{eq}(\alpha_H, T_{H,1}^*, T_{H,0}^*) < z \leqslant z_1^{SB}, \end{cases}$$

where

$$T_{i,1}(z) = T_{i,0}^* - \psi(\varepsilon_{i,0}^{SB}) + \psi(\varepsilon_{i,1}^{SB}) + \frac{\alpha_i}{2}\big[z - ((1 - \varepsilon_{i,0}^{SB})\Delta U_B - (\varepsilon_{i,0}^{SB} - \varepsilon_{i,1}^{SB})\Delta U_A\big]. \tag{A.27}$$

Figure 1 and Proposition 7 rest on two key premises. First, we have assumed that $z_{All}^{SB} < z_1^{eq}(\alpha_L, T_{L,1}^*, T_{L,0}^*)$. This allowed us to combine Lemmas 3 and 4 into a single figure while covering the broadest range of cases. If instead $z_{All}^{SB} > z_1^{eq}(\alpha_i, T_{i,1}^*, T_{i,0}^*)$, there would be no configuration in which a type-i physician optimally tests *only* A-signal patients. For $z_1^{eq}(\alpha_i, T_{i,1}^*, T_{i,0}^*) < z < z_{All}^{SB}$, type-$i$ physicians would choose to not test anyone, although the optimum is to test all patients. In this case, the fixed transfer $T_{All}(z)$ would need to be raised above its level in Definition 2 to induce physicians to test all patients. Although the precise value of $T_{All}(z)$ would differ from that in Proposition 7, the substance of our decentralization argument remains unchanged.

Second, Assumptions 3 and 4 imply that, at the second-best, both physician types always make the same testing decision (*i.e.* the thresholds z_{All}^{SB} and z_1^{SB} do not vary with α_i). A sufficient condition for this result is that the equal proportion of type-L and type-H physicians (Assumption 4). Appendix 9.5.6 shows that our core argument remains valid even when this assumption is relaxed. The only potential failure arises when three conditions hold simultaneously: (i) the optimal testing decision differs between the two types of physicians (*i.e.* $z_k^{SB}(\alpha_i) < z < z_k^{SB}(\alpha_j)$ for $i \neq j$ and $k \in \{All, 1\}$); (ii) at the fixed parts of the transfer levels of Definition 2, one type makes its second-best testing decision while the other does not; (iii)

the increase in the fixed part required to align the latter type's testing decision is large enough to alter the former's. Although such a configuration cannot be excluded, its likelihood is low. Hence, our conclusion -that second-best testing decisions can generally be decentralized with fixed parts of transfers depending on the test cost- holds in most cases, even without Assumption 4.

9.5.6 Second-best testing decisions varying with altruism degree [ONLINE APPENDIX]

In this Appendix, we study the decentralization of the second-best decisions when Assumption 4 does not hold. This implies that second-best test cost thresholds z_{All} and z_1 now vary with α. We focus on the choice between testing 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 testing only patients with an A signal or no one (*i.e.* to the threshold $z_1^{SB}(\alpha_i)$).

Since second-best test cost thresholds now depend on α_i, there are three novel potential rankings of equilibrium and second-best 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. $z_{All}^{eq}(\alpha_L, T_{All}^*, T_{L,1}^*) < z_{All}^{SB}(\alpha_L) < z_{All}^{eq}(\alpha_H, T_{All}^*, T_{L,1}^*) < z_{All}^{SB}(\alpha_H)$;

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 fixed parts of the 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 differs 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. $z_{All}^{eq}(\alpha_L, T_{All}^*, T_{L,1}^*) < z_{All}^{eq}(\alpha_H, T_{All}^*, T_{L,1}^*) < z_{All}^{SB}(\alpha_L) < z < z_{All}^{SB}(\alpha_H)$;

2. $z_{All}^{eq}(\alpha_L, T_{All}^*, T_{L,1}^*) < z_{All}^{SB}(\alpha_L) < z_{All}^{eq}(\alpha_H, T_{All}^*, T_{L,1}^*) < z < z_{All}^{SB}(\alpha_H)$;

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 testing decision can only be decentralized by increasing the fixed part 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 decision (to test only patients with an A signal). Except in these very specific circumstances, the social planner can still decentralize the second-best testing decision as described in Section 7.2 even when $z^{SB}_{All}(\alpha_i)$ varies with physicians' types.

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

First, we introduce the new set-up with $\lambda \neq 1/2$. Second, we show that Lemma 1 is robust to $\lambda < 1/2$. Third, we derive the optimal efforts in three cases. Fourth, we characterize the thresholds of z that determine the optimal testing decision. Finally, we characterize the thresholds that arise from the physician's optimization problem.

9.6.1 Modified set-up

In this section, we assume a generic proportion λ (resp. $(1-\lambda)$) of patients of type A (resp. type B). As mentioned in footnote 11, we weaken Assumption 1 to become

Assumption 5 $(1-\lambda)\Delta U_B > \lambda \Delta U_A > 0$.

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 A.3 reports the frequencies of the four cases (two signals and two types) in the population:

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

Table A.3: Frequencies in population as a function of λ

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 a 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 personalized one.

9.6.2 Optimal effort levels

We first compute the social optimum. We proceed exactly as in the main text.

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

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

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 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:

$$W_0(\varepsilon_0) = -\psi(\varepsilon_0) + (1-\lambda)\varepsilon_0 U_B^D + \lambda(1-\varepsilon_0)U_A^D + (1-\lambda)(1-\varepsilon_0)U_B^P + \lambda\varepsilon_0 U_A^P,$$

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:

$$\psi'(\varepsilon_0^*) = (1-\lambda)\Delta U_B + \lambda \Delta U_A. \tag{A.28}$$

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).

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

$$W_0(\varepsilon_0^*) > (1-\lambda)U_B^D + \lambda U_A^D.$$

Case 1: Test prescribed (after effort ε chosen) only to patients with an A signal

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

$$W_1(\varepsilon_1) \quad = \quad -\psi(\varepsilon_1) + (1-\lambda)U_B^D + \lambda[\varepsilon_1 U_A^P + (1-\varepsilon_1)U_A^D] - z((1-\lambda)(1-\varepsilon_1) + \varepsilon_1\lambda),$$

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 produce an A signal) at the test cost z for proportion $((1-\lambda)(1-\varepsilon_1)+\varepsilon_1\lambda)$ of the sample that produces an A signal. In such a case, the initial false positives receive treatment D. The first-order condition for ε_1 is given by

$$\psi'(\varepsilon_1^*) = \lambda \Delta U_A + z(1-2\lambda). \tag{A.29}$$

Note that, unlike our initial formulation (eq. 5), ε_1^* now depends on z. As explained above, a greater effort decreases the fraction of A signals 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 $\lambda = 1/2$.

Proposition A.1 *Effort and test are strategic substitutes:*

$\varepsilon^*_{All} < \varepsilon^*_1 < \varepsilon^*_0 \iff z(1-2\lambda) < (1-\lambda)\Delta U_B$.

Proof. This result is obtained by comparing first-order conditions (A.28) and (A.29). ■

Proposition A.1 generalizes Proposition 1. Intuitively, the ranking of optimal efforts 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 is not too large.

9.6.3 Robustness of Lemma 1

In order to show that Lemma 1 holds for $\lambda < 1/2$, consider the welfare level if we test after receiving a B signal rather than after an A signal. It is written as follows:

$$W_{1B}(\varepsilon) = \lambda U_A^P - (1-\varepsilon)(1-\lambda)U_B^P + \varepsilon(1-\lambda)U_B^D - \psi(\varepsilon) - z(\lambda + \varepsilon(1-2\lambda)), \qquad \text{(A.30)}$$

for a generic effort level ε. To continue testing only A signals under unequal proportions of types A and B requires $W_{1B}(\varepsilon) < W_1(\varepsilon)$, or equivalently that

$$\begin{aligned} -\psi(\varepsilon) + (1-\lambda)U_B^D + \lambda[\varepsilon U_A^P + (1-\varepsilon)U_A^D] - z((1-\lambda)(1-\varepsilon) + \varepsilon\lambda) > \\ \lambda U_A^P - (1-\varepsilon)(1-\lambda)U_B^P + \varepsilon(1-\lambda)U_B^D - \psi(\varepsilon) - z(\lambda + \varepsilon(1-2\lambda)). \end{aligned}$$

After some rearrangements, we find that this is effectively the case if

$$(1-\lambda)(1-\varepsilon)\Delta U_B - z((1-\lambda)(1-\varepsilon) + \varepsilon\lambda) > \lambda(1-\varepsilon)\Delta U_A - z(\lambda + \varepsilon(1-2\lambda)). \qquad \text{(A.31)}$$

Under Assumption 5, the first term on the left-hand side is greater than the first term on the right-hand side. The parenthesis terms factoring z on both sides represent the number of tests performed on A and B signals, respectively. One can also show that

$$(\lambda + \varepsilon(1-2\lambda) > (1-\lambda)(1-\varepsilon) + \varepsilon\lambda,$$

when $\varepsilon > 1/2$, that is, the number of A signals tested is smaller than the number of B signals tested. This ensures that expression (A.31) is always satisfied, and thus, that testing only

patients with a B signal is dominated by testing only patients with an A signal. This shows that Lemma 1 still applies when $\lambda < 1/2$.

9.6.4 Optimal diagnostic testing 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 A signals received, and belonging to the interval $[1-\varepsilon_1^*(z), \varepsilon_1^*(z)]$, and thus lower than 1. As the proportion of A signals 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

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

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

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

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

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.

9.6.5 The physician's problem

The physician's utility now writes:

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

We obtain the following (equilibrium) levels of efforts:

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

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, without correcting transfers, 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

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

This expression is similar to equation (12), 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:

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

This threshold level is very similar to equation (13). 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 types A and 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 these points in footnote 22.

9.7 Robustness of the non-responsiveness result to non-linear contracts [ONLINE APPENDIX]

In this appendix, we use figures to illustrate how relaxing the constraints on physician contracts may enable the planner to improve upon a pooling contract by offering separating contracts. Our focus is on various forms of non-linear contracts, and we base the illustrations on Figure A.1 in Appendix 9.3.1.

9.7.1 Graphical conditions for incentive compatibility of separating contracts

To examine the robustness of the non-responsiveness result to the introduction of non-linear schedules, we begin with the best pooling contract defined in Section 7.1 (for a generic case $k = \{0, 1\}$) and reproduce it in Figure A.3.[31] We denote by X the choice of physician L (with corresponding effort level ε_L) and by Y the choice of physician H (with corresponding effort level ε_H). As established in Proposition 5, we have $\varepsilon_L < \varepsilon^* < \varepsilon_H$.

[31] We omit the SB exponents, and indices k.

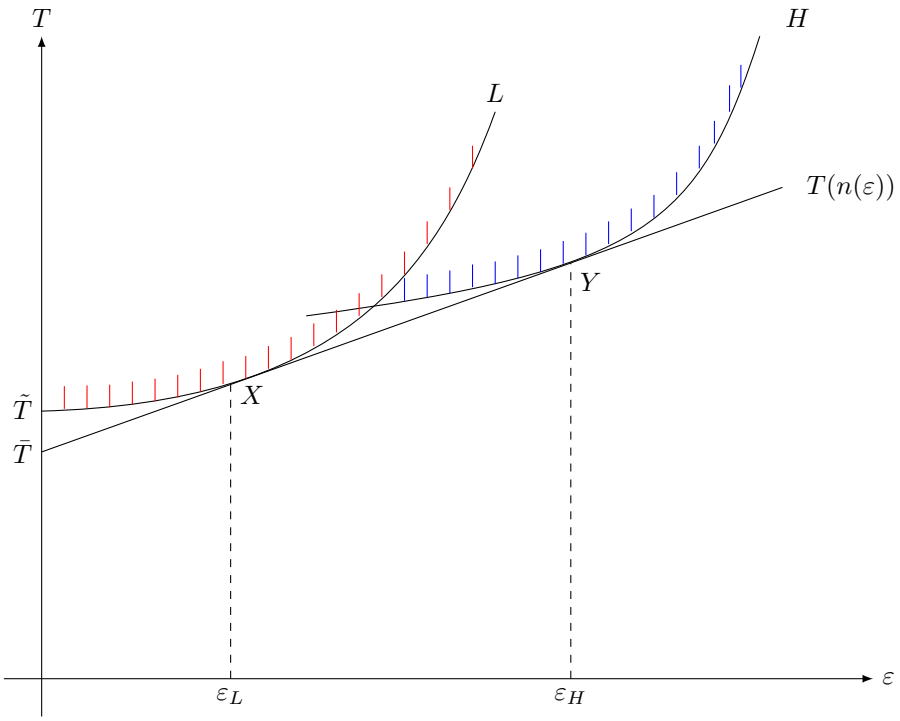

Figure A.3: Best pooling contract

We now construct a pair of separating contracts. For simplicity, we assume that the contract designed for type-L physicians coincides with the best pooling contract shown in Figure A.3, so that her chosen effort level ε_L remains unchanged. Our goal is then to design a non-linear contract for type-H physicians that lowers her effort, bringing it closer to the first-best level, while still satisfying all incentive and participation constraints. Figure A.3 illustrates the generic requirements for such a contract, after which we examine different families of non-linear contracts to determine whether they meet these requirements.

First, for type H to prefer the contract designed for her over point Y, the contract curve must pass through the blue-shaded area in Figure A.3 (*i.e.*, above type H's indifference curve through Y). Second, to prevent type L from mimicking H, the contract intended for H must remain outside the red-shaded area (*i.e.*, above type L's indifference curve through X). Third,

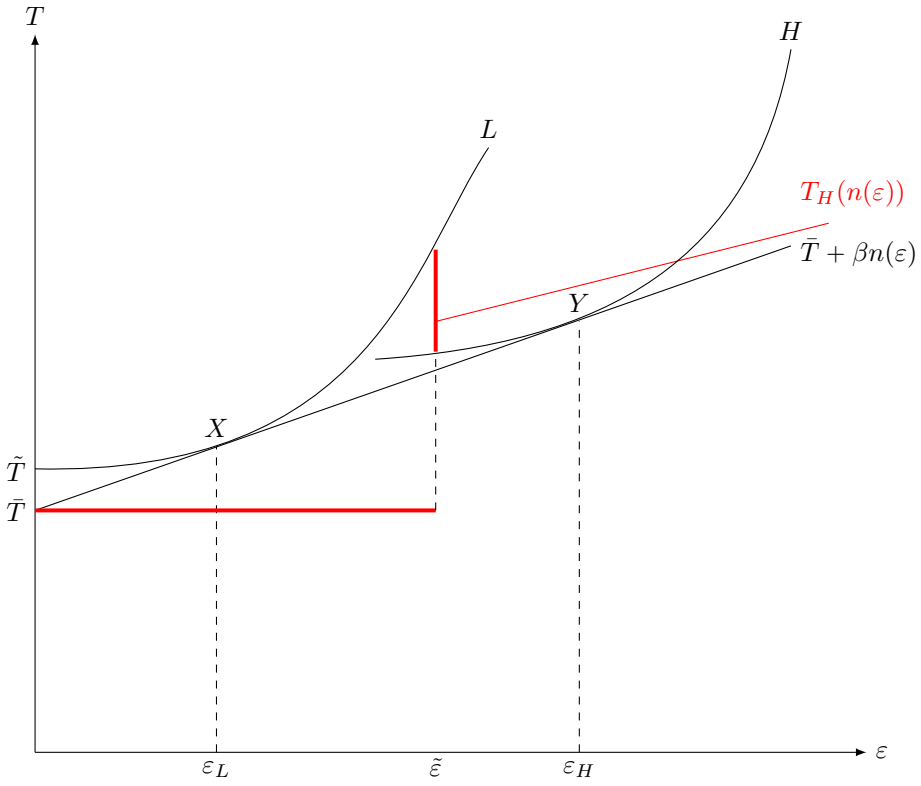

Figure A.4: Discontinuous non-linear contract

given the contract curve designed for her, type H's effort choice must be strictly below ε_H (*i.e.*, closer to ε^*).

We now use Figure A.4 to examine whether these three constraints can be satisfied simultaneously with different families of non-linear contracts.

9.7.2 Piecewise linear contracts with threshold-based bonuses

We first consider piecewise linear contracts with thresholds, above which physicians receive a bonus. They are specified as follows:

$$\begin{aligned} T_H(n(\varepsilon)) &= T_0 + \beta_0 n(\varepsilon), \quad \text{if } n(\varepsilon) < n(\tilde{\varepsilon}), \\ &= T_1 + \beta_H (n(\varepsilon) - n(\tilde{\varepsilon})), \quad \text{if } n(\varepsilon) \geqslant n(\tilde{\varepsilon}). \end{aligned} \tag{A.32}$$

To be incentive compatible (*i.e.*, chosen by type H but not by type L when the pooling contract $\bar{T} + \beta n(\varepsilon)$ is also available), the parameters of this contract $\{\beta_0, \beta_H, T_0, T_1, \tilde{\varepsilon}\}$ must satisfy the following four requirements:

1. $T_0 + \beta_0 n(\varepsilon) < \bar{T} + \beta n(\varepsilon)$ for $n(\varepsilon) < n(\tilde{\varepsilon})$ (so that type L does not mimic type H);
2. $\varepsilon_L < \tilde{\varepsilon} < \varepsilon_H$: the threshold granting access to the bonus lies strictly between the effort levels chosen by both types under the pooling contract;
3. $\beta_H < \beta$: to reduce the over-provision of effort by type H under the pooling contract;
4. T_1 must be sufficiently large for type H to prefer this contract to the pooling contract, but not so large as to attract type-L physicians. This admissible range is represented by the red vertical interval at $\tilde{\varepsilon}$ in Figure A.4.

Contracts satisfying these four requirements do exist and Figure A.4 illustrates one such contract (in red) designed for type-H physicians, where T_0 is set to $\bar{T}$ and $\beta_0 = 0$.[32]

We emphasize next that the bonus amount $T_1 - (T_0 + \beta_0 n(\tilde{\varepsilon}))$ paid when the effort threshold is reached must constitute an independent degree of freedom for the planner, since it has to be

[32] Since this portion of the contract is not chosen by type H at equilibrium, it can be replaced by this simpler specification in which a flat transfer is offered for any $\varepsilon < \tilde{\varepsilon}$.

chosen within a specific range. For instance, a contract of the type

$$\begin{aligned} T_H(\varepsilon) &= \breve{T}, \quad \text{if } n(\varepsilon) < n(\tilde{\varepsilon}), \\ &= \breve{T} + \beta_H(n(\varepsilon) - n(\tilde{\varepsilon})), \quad \text{if } n(\varepsilon) \geqslant n(\tilde{\varepsilon}) \end{aligned} \tag{A.33}$$

where $\beta_H < \beta$ cannot satisfy simultaneously the incentive compatibility constraints of both types. Such a contract is kinked but continuous at $\varepsilon = \tilde{\varepsilon}$ and cannot restore the responsiveness of the second-best contract.

Hence, having a discontinuity at $\varepsilon = \tilde{\varepsilon}$ is a necessary, although not sufficient, condition for a separating contract to be incentive compatible for both types. Indeed, the simple contract of the form

$$\begin{aligned} T_H(\varepsilon) &= \bar{T}_H, \quad \text{if } n(\varepsilon) < n(\tilde{\varepsilon}), \\ &= \bar{T}_H + \beta_H n(\varepsilon), \quad \text{if } n(\varepsilon) \geqslant n(\tilde{\varepsilon}) \end{aligned} \tag{A.34}$$

violates either type L's IC (if $\bar{T}_H$ is set high enough for the other conditions to be satisfied, which is the case depicted on Figure A.5)[33] or type H's IC constraint (if $\bar{T}_H \leqslant \tilde{T}$, with $\tilde{T}$ defined on Figure A.5).

[33]The simultaneous requirements that this contract has a lower slope than the one offered to type L, while also passing through the blue-shaded area, imply that it must encroach upon type L's incentive compatibility constraint.

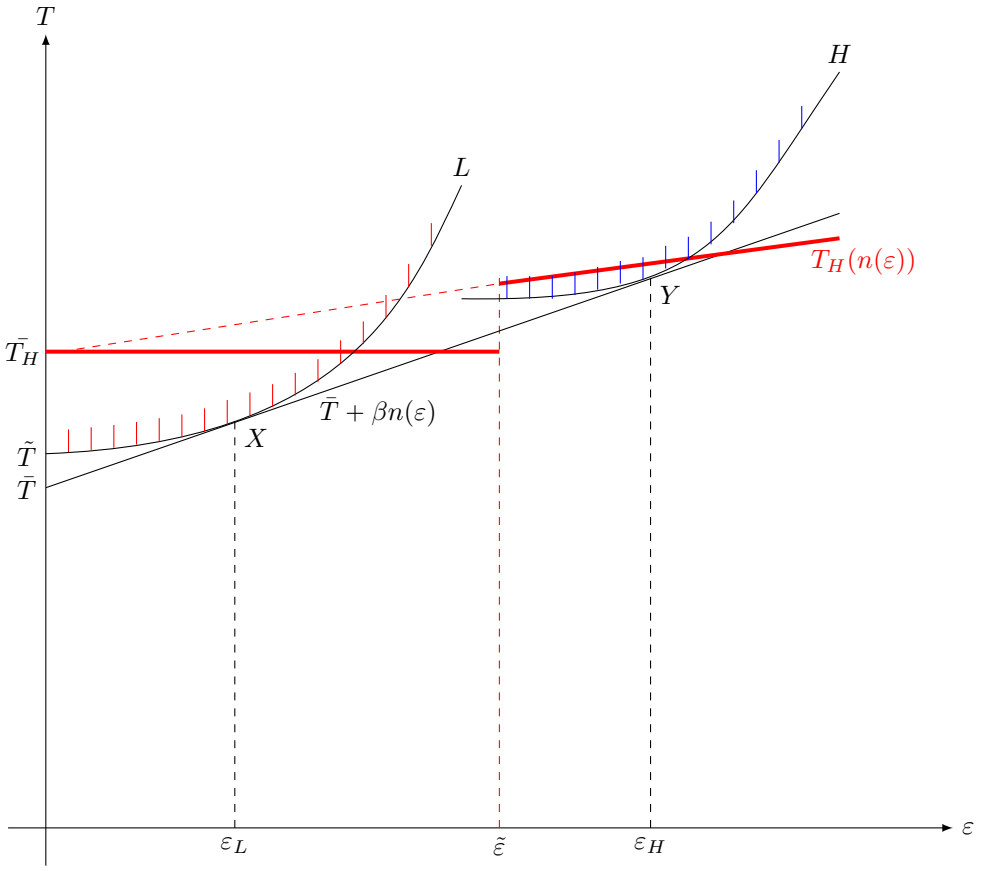

Figure A.5: Example of a non-linear contract of the form (A.34)

In that case, there is indeed a jump in the transfer (a bonus) received when the threshold effort level is reached, but the magnitude of this jump is determined by both $\tilde{\varepsilon}$ and β_H and does not give an additional degree of freedom to the planner.

We summarize our results in the next proposition.

Proposition A.2 *The planner can improve upon a pooling contract by offering type-H physicians a piecewise linear contract that includes a bonus once the threshold proportion of correctly treated patients, $n(\tilde{\varepsilon})$, is reached. This contract is incentive-compatible if the increase in transfer at the threshold is neither too small nor too large. However, the non-responsiveness result*

continues to hold for schedules without a discontinuity at this point, as well as for contracts restricted to the family described in (A.34). In these cases, the pooling contract remains the second-best welfare-maximizing outcome.

Although Proposition A.2 shows that the non-responsiveness result is not fully robust to certain extensions of the admissible contract space, several arguments still strongly support the persistence of non-responsiveness as a relevant outcome.

First, as the number of physician types increases, designing separating contracts becomes increasingly more complex, since each must include carefully calibrated incentives to maintain compatibility. At the same time, the informational burden of computing such contracts grows rapidly, whereas a pooling contract -depending only on average altruism- remains far more tractable and practical to implement. In the limit, we conjecture that the non-responsiveness property would continue to hold under a continuum of physician types.

Second, in practice, most P4P schemes are intentionally simple, featuring only a limited number of thresholds or performance bands. Beyond the informational and incentive challenges just mentioned, the sophistication required to design and manage complex, highly non-linear contracts would be difficult to achieve and justify in real-world health systems. Consistent with this observation, large-scale P4P schemes appear closer to contracts of the form (A.34) than to (A.32). A salient example is the Hospital Value-Based Purchasing (VBP) program operated by the U.S. Centers for Medicare & Medicaid Services (CMS).[34] Under the VBP program, each quality measure is scored on a fixed 0–10 scale based on performance relative to an achievement threshold and a benchmark, with hospitals earning higher points only once they exceed the achievement threshold set in regulation for the formal scoring rules.[35] Hospitals' total performance scores are then transformed into value-based incentive payment percentages via a linear exchange function set in the regulations, so that there is no additional discretionary bonus at an aggregate threshold.[36] A similar structure arises in the UK's Quality and Outcomes Frame-

[34]Overview at https://www.cms.gov/medicare/quality/initiatives/hospital-quality-initiative/hospital-value-based-purchasing

[35]See https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-B/part-412/subpart-I/section-412.165

[36]https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-B/part-412/subpart-I/section-412.162

work (QOF) for general practitioners (GPs), under which practices accumulate achievement points across a menu of indicators that directly translate into payment adjustments via the GP contract.[37] In both cases, the implied jump in scoring at the achievement threshold follows mechanically from the scoring rules rather than from a freely chosen discontinuous bonus, and payments are then linked to those scores through essentially continuous schedules. Thus, while threshold-based discontinuities in scoring are commonly used in practice, they do not typically exploit the additional degree of freedom embodied in (A.32); instead, real-world P4P contracts tend to feature jumps that are tightly pinned down by the scoring rule, as in (A.34).

References

[1] Abaluck, J., Agha, L., Kabrhel, C., Raja, A. and Venkatesh, A. 2016. The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care, American Economic Review, 106 (12), pp. 3730-64.

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

[3] Akhmetov, I. and Bubnov, R., 2015. Assessing value of innovative molecular diagnostic tests in the concept of predictive, preventive, and personalized medicine, EPMA J, 6:19.

[4] 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.

[5] Bardey, D., De Donder, Ph. and Leroux, M.L. 2024. Incentivizing Physicians' Diagnostic Effort and Test with Moral Hazard and Adverse Selection. CEDE Working Paper, n. 24-46, Bogota. Available at `https://repositorio.uniandes.edu.co/bitstreams/8a706fd5-5e60-45f2-a3e8-b330540e6e3e/download`

[37] `https://en.wikipedia.org/wiki/Quality_and_Outcomes_Framework`; see also the online database and guidance at https://digital.nhs.uk/services/quality-and-outcomes-framework-qof-online-database.

[6] Bardey, D., Kembou Nzale, S., Ventelou, B. 2021, Physicians' incentives to adopt personalised medicine: an experimental approach, Journal of Economic Behavior and Organization, vol 191(C), p. 472-500.

[7] Barigozzi F., Canta C., and H. Cremer. 2025. Workers' Motivation and Quality of Services in Mission-Driven Sectors. TSE Working Paper, n. 25-1655.

[8] 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.

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

[10] 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

[11] 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.

[12] 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.

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

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

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

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

[17] Ergun-Sahin, B., Gunes, E. D., Kocabiyikoglu, A., and Keskin, A., 2022. How does workload affect test ordering behavior of physicians? An empirical investigation, Production and Operations Management, 31(6), 2664-2680.

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

[19] Galizzi, M.M., Godager, G., Li, J., Linnosmaa, I., Tammi, T., Wiesen, D., 2023, Economics of Healthcare Provider Altruism. In: Zimmermann, K.F. (eds) Handbook of Labor, Human Resources and Population Economics.

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

[21] Gertler, P. and Kwan, A., 2024, The Essential Role of Altruism in Medical Decision Making, NBER Working Paper 32151.

[22] 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.

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

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

[25] Jullien, B., 2000. Participation Constraint in Adverse-Selection Models, Journal of Economic Theory, 93, p.1-47.

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

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

[28] 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.

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

[30] Miller, G. and S. Babiarz, K., 2013. Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs. in Tony Cuyler (ed.), Encyclopedia of Health Economics, Elsevier.

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

[32] 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.

Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science 366 (6464), p. 447-453.

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

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

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

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

[37] 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.

[38] 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.