Optimal Costly Oversight with Investment and Cream-Skimming

By Sumeyra Akin and Tibor Heumann*

Transplant center oversight–conducted through costly flagging based on survival rates–is intended to improve outcomes. However, it has faced challenges because centers strategically select patients (and organs) to improve their reported outcomes. We study the optimal flagging policy using a mechanism-design model in which heterogeneous centers can improve their survival rates by investing, selecting patients, or both. We find that (a) the optimal mechanism never flags intrinsically good-enough center types, and (b) under the optimal mechanism, intrinsically worse centers select patients, while intrinsically better centers underinvest. These results apply to the regulation of charter schools and development banks.
JEL: D47, D82, I18
Keywords: Market design, transplant centers, mechanism design, flagging

> When a measure becomes a target, it ceases to be a good measure
>
> Goodhart's law

Over the past decade, the regulation of transplant centers has been the subject of a lively debate. Regulators aim to improve patient outcomes, for example, by encouraging greater investment, but face significant challenges. Oversight relies heavily on coarse aggregate performance measures, such as a center's average survival rate. These measures reflect not only a center's practices and investment choices but also (a) the pool of patients a center admits and (b) the transplant centers' intrinsic characteristics (such as access to the surgeon market). As a result, centers can improve their outcomes by selectively removing higher-risk patients, while the regulator faces the difficulty of determining how demanding to be with any center.

For concreteness, we focus on the regulation of kidney transplants. From 2007 to 2019, the Centers for Medicare and Medicaid Services (CMS) integrated survival rates into its Conditions of Participation (CoPs) for transplant centers.[1]

* Akin: Instituto de Economía, Pontificia Universidad Católica de Chile, sumeyra.akin@uc.cl. Heumann: Instituto de Economía, Pontificia Universidad Católica de Chile, tibor.heumann@uc.cl. Earlier versions circulated under the titles "Optimal Performance-Based Reviews" and "Optimal Flagging of Transplant Centers." We thank Dirk Bergemann, Nicolas Figueroa, Stephen Morris, Alexander Wolitzky, and three anonymous referees for helpful comments and suggestions.

[1] See Ho, Skaro and Abecassis (2015) for a detailed description of the regulatory environment and its

Centers with survival rates below a threshold were "flagged." These centers faced a rigorous review process that risked the loss of a Medicare contract and possible closure (Kasiske et al. 2016). The flagging mechanism improved outcomes, particularly for low-performing centers (Chandraker et al. 2019), but spurred strategic responses. Centers anticipating CMS sanctions delisted high-risk patients and rejected high-risk kidneys to increase survival rates (Stith and Hirth 2016, White et al. 2015, Chan 2021). These unintended consequences, along with the system's failure to provide actionable feedback and its cost burden on both centers and CMS, prompted its termination in 2019 (Andreoni 2020). While similar adverse outcomes are reported for other organ transplants, the discussion surrounding the regulation of kidney transplant centers is exceptionally stark and well-documented.[2] We return to the details of this market after introducing the model.

We develop a principal–agent model that incorporates strategic behavior, heterogeneity across centers, and enforcement costs to study the optimal design of a flagging policy. Although we frame the model with the regulation of transplant centers in mind, its structure also applies to other settings where regulators rely on coarse statistics to monitor institutions. In such environments, institutions may respond either by undertaking welfare-enhancing actions or by engaging in cream-skimming. We return to these alternative interpretations after presenting our results (see Section IV).

Principal-Agent Model of Flagging. — The model comprises a regulator that sets a flagging policy specifying the flagging probability as a function of survival rates. A center's average survival rate depends on its investment, the number of patients it removes from its waiting list, and its privately observed type, which captures characteristics that the regulator does not observe. Transplant centers maximize profits by choosing their investment and patient selection (the former increases costs; the latter decreases revenue). Investment improves overall survival rates, and the center can further increase its average survival rate by excluding patients with poorer characteristics (i.e., lower survival probabilities). For the regulator, patient removals are detrimental to welfare, but it cannot directly observe when the center removes patients.

We assume that flagging reduces both the center's profits and the regulator's welfare. The welfare cost of flagging, in reduced form, captures three effects. First, performance monitoring and regulatory follow-up impose administrative and compliance burdens on both transplant programs and oversight agencies.[3]

history.

[2]The reason it is better documented for kidney transplants is that patients listed for a kidney transplant can survive on dialysis if they are delisted, so patient selection is much more prevalent for kidneys than for other organs. For example, removing a patient from the liver transplant list typically results in the patient's death in a short period of time, and a sicker liver patient does not necessarily have worse post-transplant outcomes (Stith and Hirth 2016).

[3]For example, Jay and Schold (2017) discuss that performance measurement and oversight can increase documentation and monitoring requirements, diverting resources toward compliance.

Second, centers that receive low-performance evaluations (or are "flagged") often respond by reducing transplant activity, which is welfare-reducing.[4] Third, flagging can require waitlisted patients to transfer to another program (or otherwise face delays), which can adversely affect patient outcomes.[5] Our specification remains agnostic about the specific channels through which these costs affect welfare, beyond assuming that flagging is socially costly.

The regulator, therefore, designs a flagging policy that promotes investment, discourages patient selection, and limits the overall incidence of flagging, relying solely on the average survival rate as a coarse metric and flagging as a disciplinary instrument.

Optimal Flagging Policy. — We begin by studying the transplant center's problem of achieving a target survival rate under a given flagging probability. For each type, this induces a cost-minimization problem over investment and patient selection, thereby defining an indirect profit function in terms of survival rates and flagging probability. We show that the model satisfies the standard single-crossing condition, which guarantees that, under any flagging policy, higher types achieve higher survival rates and lower flagging probabilities. These incentive properties allow us to apply classical screening results from the literature to characterize implementable flagging policies using the indirect profit functions.

Our first main result on the optimal flagging policy (Proposition 2) shows that it never flags sufficiently high-type centers. The intuition relies on comparing the extensive margin of the costs and benefits of screening. Consider a flagging policy that screens in a neighborhood around the highest type (i.e., flagging is used with positive probability on types near the highest type). Screening at the top of the type distribution increases welfare generated by these centers by incentivizing them to invest more. However, doing so also requires increasing the flagging probability for all lower types. By being less demanding at the top (i.e., assigning a zero flagging likelihood to relatively low survival rates), the regulator can decrease the probability of costly flagging for (essentially) all types.

We then examine the distortions induced by the optimal flagging policy. To this end, we compare the average survival rate under the optimal flagging policy with that achieved if the regulator could observe center type but could not directly choose a center's investment and patient-selection decisions. This benchmark

[4]Schold et al. (2013) document that centers receiving low-performance evaluations subsequently experience declines in kidney transplant volume. Nevertheless, as shown by Stith and Hirth (2016), centers rarely shut down due to noncompliance. Instead, reduced volume is primarily associated with increased delistings and patient transfers to other centers. Moreover, they argue that this response is unlikely to be welfare-improving, as it reduces access to transplantation without corresponding improvements in patient survival.

[5]CMS guidance requires transplant programs facing inactivation or termination to inform waitlisted patients and provide assistance for transfer to another Medicare-approved program without loss of accrued waiting time (Centers for Medicare & Medicaid Services 2019, Centers for Medicare & Medicaid Services 2024). Such transfers may delay transplantation or prolong time on dialysis, which is associated with significantly worse survival and quality-of-life outcomes relative to transplantation; see Wolfe et al. (1999), Tonelli et al. (2011), and Chaudhry et al. (2022).

isolates the informational distortion arising from hidden types while preserving the moral-hazard friction associated with unobservable investment and patient selection. We use this benchmark for our analysis because the flagging policy screens center types but cannot directly correct the distortions arising from moral hazard. Using this comparison, we characterize how the optimal flagging policy distorts survival rates relative to the welfare-maximizing benchmark.

Our second main result (Proposition 4) shows that the optimal flagging policy inevitably introduces two types of distortion. The policy will be too demanding for low types; that is, the average survival rate implemented will be higher than it would be if types were observable. This will result in excessive patient selection by low types. The policy will be too lenient for high types; that is, the average survival rate implemented will be lower than it would be if types were observable. This will result in underinvestment by high types. Hence, distortions are inevitable and vary across types.

Related Literature. — Although there is a substantial literature on kidney allocation, the regulation of transplant centers has received much less attention; see, for example, Roth, Sönmez and Ünver (2004). Chan and Roth (2024) analyze the optimal regulation of transplant centers. They explore the possibility of jointly regulating transplant centers and organ procurement organizations (oversight is currently conducted independently). They perform a laboratory experiment demonstrating that joint regulation would increase kidney transplantation rates. Consistent with current practice, we focus on the optimal regulation of transplant centers.

Our model exhibits both moral hazard and adverse selection; however, because the regulator can observe a transplant center's average survival rate, the only relevant source of asymmetric information for regulatory design is adverse selection. In this sense, the model is similar to a regulator overseeing a firm that faces both forms of asymmetric information but can observe the firm's cost, as in Laffont and Tirole (1986). The main differences relative to their model are that our model addresses a multidimensional moral-hazard problem and employs flagging rather than transfers.

Our paper relates to work on regulation in the mechanism design literature, pioneered by Baron and Myerson (1982) and later studied in greater generality by Guesnerie and Laffont (1984). In contrast to these papers, in our model, the flagging probability reduces the principal's welfare (in the classic regulation literature, transfers are (weakly) beneficial for the principal). This modeling difference leads to substantial differences in predictions. First, our model predicts that the optimal mechanism never screens at the top of the type distribution.[6]

[6] In our model, bunching at the top of the type distribution arises not because the efficient allocation prescribes it, but because screening is too costly. For example, in a single-agent, single-good environment, the revenue-maximizing selling mechanism is to post a "take it or leave it" offer (Myerson 1981). The implemented allocation exhibits bunching at the top: all agents with sufficiently high valuation buy the object. In this case, bunching arises because the efficient allocation exhibits bunching.

Second, undistorted types occur only in the interior, in contrast to the ubiquitous "no distortion at the top" results. Third, distortions are not in the same direction across all types. Some types are distorted by receiving an allocation higher than the first-best allocation. At the same time, other types are distorted by receiving an allocation lower than the first-best allocation.[7]

Amador and Bagwell (2013), Ambrus and Egorov (2017), and Amador and Bagwell (2020) study the use of money burning in optimal delegation problems and characterize conditions under which it arises in optimal policies.[8] The flagging probability in our model plays a similar role, as it reduces both the agent's profit and the principal's welfare. However, an important distinction from the delegation setting is that, in our model, the agent's preferences are strictly monotonic with respect to the allocation (the survival rate). As a result, the regulator cannot screen types based on survival targets alone and must rely on an additional, costly instrument to maintain incentive compatibility.

Recent literature examines the use of ordeals to allocate goods, including Dworczak, Kominers and Akbarpour (2021), Akbarpour, Dworczak and Kominers (2024), Dworczak (2023), and Mookherjee and Png (1994). The flagging policy in our model can be viewed as an ordeal because it is costly for both parties. As in the literature on money burning, these papers focus on when and how ordeals are used to increase welfare. In contrast to this literature, we focus mainly on the qualitative properties of the optimal mechanism rather than on when such instruments are used.

In our model, payoffs are not quasi-linear; thus, the distinction between ordeals and money burning is not well-defined. For the same reason, the arguments are more basic and rely only on essential properties of the marginal rates of substitution. Hence, although similar results have appeared in the literature, our arguments provide a new perspective and demonstrate the generality of these results.[9]

In the mechanism design with costly verification literature, the principal can incur a cost to learn the agent's type; see, for example, Baron and Besanko (1984) and Border and Sobel (1987).[10] This screening tool is similar to the flagging probabilities in our model, since they are both screening instruments that are costly to the principal. However, in this literature, the agent does not bear the

[7]Bunching and distortions in both directions also arise in models with transfers when the agent's reservation utility depends on his type (see Lewis and Sappington (1989a; 1989b) for early contributions and Maggi and Rodriguez-Clare (1995) for a unified treatment). In our model, the participation constraint is trivially satisfied, so it plays no role in the analysis.

[8]Money burning has also appeared in multi-agent environments (Condorelli 2012, McAfee and McMillan 1992, Hartline and Roughgarden 2008, Chakravarty and Kaplan 2013).

[9]Employing variational arguments to provide necessary conditions for optimality in environments with non-quasi-linear preferences is not new to the literature; see, for example, Mirrlees (1971) and Guesnerie and Laffont (1984). Our contribution is to apply them to the specific model in which flagging is costly to both parties.

[10]There is also related literature studying costly verification in multi-agent settings, for example, see Ben-Porath, Dekel and Lipman (2014). There is also a broader literature on approval mechanisms when information acquisition is costly, see e.g., Henry and Ottaviani (2019).

direct verification cost.

The paper proceeds as follows. Section I introduces the model. The model analysis is presented in Sections II-III. In Section IV, we discuss additional applications of our model. Finally, Section V concludes. All proofs are collected in the Appendix.

I. Model

In Section I.A, we present the payoffs; in Section I.B, we explain how survival rates are determined; and in Section I.C, we present the regulator's problem. Section I.D discusses and interprets our model. Finally, Section I.E provides a preliminary overview of the analysis. Throughout the paper, all introduced functions are assumed to be three times differentiable.

A. Payoffs

A transplant center serves a unit mass of potential patients. We refer to the center's intrinsic characteristic as its type θ and assume that it is drawn from an absolutely continuous distribution F with strictly positive density on a compact support $[\underline{\theta}, \overline{\theta}] \subset \mathbb{R}_+$. The center chooses an investment level $I \in \mathbb{R}_+$ and removes a mass $r \in [0,1]$ of patients from its waitlist, so that transplant volume equals $1-r$. A center may also be flagged; we let $f \in \{0,1\}$ denote its flagging status, with $f = 1$ indicating that it is flagged.

The center's expected profits, $\pi(r, I, f)$, are independent of center type. For any flagging probability $p \in [0,1]$, we define expected profits as

$$\pi(r, I, p) \triangleq (1-p)\pi(r, I, 0) + p\,\pi(r, I, 1). \tag{1}$$

We assume that the first-order partial derivatives of the profit function satisfy

$$\pi_I(r, I, p) < 0, \qquad \pi_r(r, I, p) < 0, \qquad \pi_p(r, I, p) < 0, \tag{2}$$

where subscripts denote partial derivatives. These assumptions imply that investment, patient removal, and flagging are all costly for the center. We further assume that π is weakly concave and that

$$\pi_{Ir}(r, I, p) \geq 0. \tag{3}$$

That is, investment and patient removal are weak complements. Economically, this captures the idea that investment increases per-patient marginal costs, so the marginal cost of investment is (weakly) lower when transplant volume is smaller.

We also assume that

$$\pi_{Ip}(r, I, p) \geq 0, \qquad \pi_{rp}(r, I, p) \geq 0. \tag{4}$$

Thus, the marginal cost of both investment and patient removal decreases with the probability of being flagged. Intuitively, regulatory scrutiny reduces transplant volume, lowering the marginal cost of patient-level investment, while a smaller patient pool limits the revenue loss from flagging.

Finally, we assume that a center that neither invests nor transplants any patients and is not flagged earns zero profits:

$$\pi(1,0,0) = 0. \tag{5}$$

This normalization will not play a substantive role in the analysis but will ensure that the participation constraint is always satisfied.

The welfare generated by a transplant center of type θ that removes r patients, invests I, and has flagging status f is denoted by $w(\theta, r, I, f)$. We assume that $w(\theta, r, I, f)$ is decreasing in r, increasing in I, and strictly decreasing in f. For any flagging probability $p \in [0,1]$, expected welfare is given by

$$w(\theta, r, I, p) \triangleq (1-p)w(\theta, r, I, 0) + p\, w(\theta, r, I, 1). \tag{6}$$

In summary, investment and patient removal are costly for the transplant center; investment raises social welfare, while patient removal lowers it. Flagging imposes costs on both the center and the regulator. Together, these forces define the model's central trade-offs. We introduce specific functional forms for profit and welfare in Section III.A.

The regulator does not directly observe the center's investment or patient-removal decisions. Instead, it observes a statistic that aggregates these actions with the center's intrinsic characteristics: the average patient survival rate, which we introduce next.

B. Survival Rates

Conditional on being listed at a center, a patient's survival probability s is strictly increasing in her health profile $\phi \in [0,1]$, the center type $\theta \in [\underline{\theta}, \overline{\theta}] \subset \mathbb{R}_+$, and the center's investment level $I \in \mathbb{R}_+$. Throughout, we assume without loss of generality that patient characteristics are uniformly distributed on $[0,1]$.[11] If a center removes all patients with a health profile below r, its average survival rate is

$$S(\theta, r, I) \triangleq \frac{\int_r^1 s(\phi, \theta, I)\, d\phi}{1-r}. \tag{7}$$

[11] Any non-uniform distribution can be absorbed into the definition of the survival function s.

A center's performance is evaluated by its average survival rate. Given transplant volume $(1-r)$ and investment I, the assumptions imply that the center optimally delists the sickest patients, that is, those with $\phi < r$. We assume that $S(\theta, 1, I) = 1$ for all (θ, I): a center that performs no transplants achieves a perfect survival rate and is therefore never flagged. Together with condition (5), this ensures that a center can always guarantee zero profits by shutting down.

We assume that survival probabilities are linear in center type,

$$s(\phi, \theta, I) = \theta\, \widehat{s}(\phi, I), \tag{8}$$

for some function $\widehat{s}(\phi, I)$. This assumption is made for analytical convenience but admits a natural interpretation: $(1-\theta)$ captures the probability that a center commits a preventable error leading to transplant failure.[12] Following the assumption made on s, we have that $\widehat{s}(\phi, I)$ is strictly increasing in both arguments, so that healthier patients and higher investment increase survival probabilities. Accordingly, define

$$\widehat{S}(r, I) \triangleq \frac{S(\theta, r, I)}{\theta}, \tag{9}$$

which corresponds to the average of $\widehat{s}(\phi, I)$ over patients with health profile $\phi \in [r, 1]$. We assume that the aggregate survival function $\widehat{S}(r, I)$ is concave and supermodular in (r, I); in the Appendix, we provide sufficient conditions for this in terms of the individual survival function $\widehat{s}$.

C. *Transplant Center's and Regulator's Problem*

A center observes its type, but the regulator only knows its distribution. The model's timing is as follows. Initially, the regulator announces the flagging probability R as a function of the center's average survival rate. The center then determines its investment level and patient selection. After all decisions are made, the average survival rate and profits are realized, and the flagging probability is determined by R.

The first step in describing the regulator's problem is to specify how a center achieves a given survival rate (analogous to solving a monopoly's cost-minimization problem). A center that wants to achieve an average survival rate q and faces a flagging probability p solves the following problem:

$$(r^c(\theta, q, p), I^c(\theta, q, p)) = \underset{\{(r,I)\in[0,1]\times \mathrm{R}_+ \,|\, S(\theta,r,I)\geq q\}}{\arg\max} \pi(r, I, p). \tag{10}$$

That is, (r^c, I^c) denotes the center's *c*hoices as a function of its type, the aver-

[12] See Ison, Holl and Ladner (2012) for a systematic review of preventable medical errors attributable to healthcare providers, including labeling, transcription, and communication errors.

age survival rate it targets, and the flagging probability it faces.[13] The indirect utilities are:

$$\Pi(\theta, q, p) \triangleq \pi(r^c(\theta, q, p), I^c(\theta, q, p), p); \tag{11}$$

$$W(\theta, q, p) \triangleq w(\theta, r^c(\theta, q, p), I^c(\theta, q, p), p). \tag{12}$$

Hence, indirect utilities capture the payoffs of the transplant center and the regulator.

The regulator's problem is to find the welfare-maximizing flagging-probability function:

$$R^{**} \in \underset{R:[0,1]\to[0,1]}{\arg\max} \int W(\theta, Q(\theta), R(Q(\theta)))dF(\theta) \tag{13a}$$

$$\text{subject to } Q(\theta) \in \underset{q\in[0,1]}{\arg\max}\ \Pi(\theta, q, R(q)) \tag{13b}$$

Following results in the literature, we know that an optimal mechanism exists; we relegate the formal statement and proof to the Appendix. The survival rate chosen by the center under the optimal flagging probability function is denoted by:

$$Q^{**}(\theta) \in \underset{q\in[0,1]}{\arg\max}\ \Pi(\theta, q, R^{**}(q)).$$

The function $Q^{**}(\theta)$ solves (13b) when the flagging probability function is R^{**}. We define

$$P^{**}(\theta) \triangleq R^{**}(Q^{**}(\theta)).$$

That is, P^{**} is the flagging probability incurred by each type under the optimal flagging mechanism. In cases with multiple optimal policies, our results apply to all of them.

D. Discussions About the Model

Survival rates. — We assume that centers can improve survival rates either by investing in patient care or by delisting the sickest patients. In our model, investment refers to any costly action that increases survival rates and welfare, such as improving administrative infrastructure or investing in capital to improve waitlist management.[14] Assuming centers can remove patients is reasonable, as this

[13] We assume that centers can costlessly lower their average survival rate, so the constraint is written as $S(\theta, r, I) \geq q$ rather than as an equality, since any center that can achieve survival above q can also adjust downward to meet q, without affecting the results.

[14] For instance, transplant centers may invest in better kidney preservation, which affects transplant success. There are two main preservation methods: machine perfusion and cold storage (Jochmans et al. 2010). Machine perfusion, a newer and more expensive technology, reduces the risk of returning to dialysis after transplantation.

practice remains central to debates about the effects of flagging. Kidney transplant centers have discretion over waitlist management because dialysis remains a treatment option. They can also clearly identify which patients to remove, since sicker patients have poorer transplant outcomes.

Transplant Center Heterogeneity. — The second aspect of the problem is the center's intrinsic characteristics, specifically its type. Transplant centers are heterogeneous in many ways that affect their outcomes; therefore, achieving a specific survival rate is significantly easier for some centers than for others. For example, transplant centers are influenced by regional referral practices and their affiliation with health networks (Chandraker et al. 2019), and they differ in their access to specialists and in the surgeon market.[15]

Flagging as a Policy Instrument. — There are three relevant aspects regarding the regulator's policy instrument (flagging) that are worth discussing.

First, we assume that the flagging probability depends only on the survival rate. We implicitly assume that the regulator cannot directly observe patient removals (modeled by the parameter r in our model) nor infer delisting practices indirectly from the number of transplants performed. This captures a limitation regulators face in detecting this kind of behavior.[16] The medical literature has also discussed why the regulator does not have access to better measures of patient care.[17]

Second, in our model, the regulator lacks other instruments to influence transplant centers. For example, organ allocation criteria are independent of centers'

[15]The following quote from NBC News coverage from 2007 regarding a transplant program at UCI that was on probation shows that centers may differ in human resource management and other administration-related issues:

> A review of 50,000 pages of subpoenaed hospital records indicated that in some cases, donor organs were available, but patients didn't receive them because there was no surgeon available . . . An independent review by a committee appointed by the university's chancellor blamed the program's failures on high turnover among medical staff, a lack of interest and oversight by campus leaders, and the fact that the two transplant surgeons lived in San Diego, some 90 miles away. The program also failed to recruit new surgeons effectively, the panel found.

See Associated Press (2007) for further details.

[16]Of course, delisting patients is a normal practice that is not necessarily associated with strategic manipulation of average survival rates. In particular, centers might take these actions for logistical or medical reasons. For example, a center might deny a donor for poor biopsy results or delist a patient for worsening health conditions while waiting for the transplant. A rather convoluted scheme to detect delisting would be to compare the number of transplants a center performs against a benchmark reflecting the expected number it should perform. However, computing such a statistic is infeasible because the patient's doctor or the dialysis center generally refers the patient for transplantation, and the decision to waitlist or remove a patient from a waitlist rests with the transplant center. Moreover, there is no national referral data tracking patients with kidney disease to identify the set of viable transplant candidates, making it impossible for the regulator to detect or infer the center's strategic decisions– such as rejecting a referred patient (Patzer and Pastan 2020).

[17]Schold (2020) discusses various measures of patient care and their limitations, offering insight into why survival rates have remained the primary performance metric.

survival rates (so this aspect of the market is beyond the scope of our paper).[18] The regulator also cannot impose fines. Hence, we believe it is a realistic assumption that the regulator is constrained to using only flagging as an instrument.

Third, in our model, flagging does not provide intrinsic benefits, underscoring its primarily disciplinary role. Although flagging may still help centers evaluate and improve their practices, the medical literature suggests that oversight was not a significant channel for improvement. For example, Woodside and Sung (2016) report that although some centers improved under CMS oversight– due in part to patient selection– the authors could not pinpoint specific mechanisms by which the review process enhanced clinical practices.

E. Outline of the Analysis

This section presents the model's components and outlines the analysis. The framework operates at two levels. First, we analyze the transplant center's patient selection and investment decisions. Second, we examine the average survival rate resulting from these choices under a given flagging policy.

For the purpose of designing the optimal flagging policy, it is sufficient to work with the indirect profit and welfare functions defined in (11)-(12). From the regulator's perspective, this formulation reduces the problem to pure adverse selection. However, these profit and welfare functions are endogenously determined by patient selection and investment decisions. Consequently, characterizing the center's investment and selection behavior is essential, as these choices shape the properties of the indirect payoff functions. Once these properties are established, the detailed mechanics of investment and selection no longer play an active role in the analysis, although we return to them later to interpret the results.[19]

The analysis proceeds as follows. Section II examines the center's incentives and their implications for the optimal flagging policy. We prove that the optimal policy assigns zero flagging probability to a positive measure of centers, based on two assumptions: flagging is detrimental to welfare and properties derived from the center's incentives. Section III then studies the regulator's indirect welfare function, showing that the optimal policy distorts both the top and bottom of the type distribution.

II. Implementability and Optimality of Pooling at the Top

We begin this section by examining the center's optimal investment and patient-selection decisions. We then show that the marginal rate of substitution between

[18] The time a patient has been on a waiting list and the proximity of the organ to different transplant centers are among these criteria.

[19] This is similar to Mirrlees (1971), where workers have private information (ability) and are subject to moral hazard (labor supply). In Mirrlees (1971), the principal can observe total income, so the problem reduces to pure adverse selection, much like our model. The main difference is that, in our model, screening is implemented via a costly instrument (a flagging policy), and moral hazard is multidimensional, thereby affecting indirect payoff functions.

the attained average survival rate and the flagging probability is monotonic in the center type. We then study the implementation of an optimal policy and show that the monotonicity of the marginal rate of substitution implies that higher center types will attain higher survival rates. Finally, we show that the optimal policy will always have an open set of center types that is never flagged.

A. Center Incentives and Monotonicity

We begin by analyzing the center's optimal choices of investment and patient selection, denoted by $(r^c(\theta, q, p), I^c(\theta, q, p))$. We fix (θ, p) and study how center choices vary with the targeted survival rate q. At the optimum, the marginal rate of substitution in the survival function equals that in the profit function:

$$\frac{S_r(\theta, r, I)}{S_I(\theta, r, I)} = \frac{\pi_r(r, I, p)}{\pi_I(r, I, p)}. \tag{14}$$

Equation (14) characterizes the center's optimal trade-off between investment and patient selection in achieving a given average survival rate. Figure 1 provides a schematic illustration of this condition.

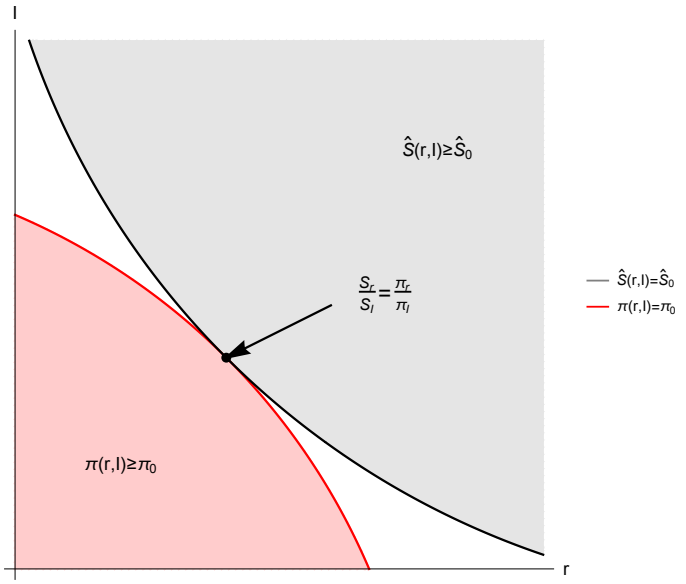


Figure 1. Optimal choices of patient selection and investment. The black curve shows combinations of (r, I) that deliver a given survival rate; the red curve shows combinations of (r, I) that deliver a given level of profit. The optimal choice equates the marginal rate of substitution in the survival function with that in the profit function.

We next establish monotone comparative statics with respect to (θ, q). For a fixed type, achieving a higher survival target requires weakly higher investment and patient selection. Conversely, holding q fixed, higher-type centers can attain it with weakly lower levels of both decisions. These properties are summarized in the following lemma.

LEMMA 1 (Monotonicity of Decisions):
The functions $(r^c(\theta, q, p), I^c(\theta, q, p))$ are weakly increasing in q and weakly decreasing in θ.

The result follows from the supermodularity of the aggregate survival function $\widehat{S}(r, I)$ and the profit function π, which delivers monotone comparative statics in the survival constraint. A tighter constraint– higher q or lower θ– induces greater reliance on both instruments. Although it may seem intuitive that investment and patient selection are monotonic, this is not straightforward and depends critically on $\widehat{S}(r, I)$ being supermodular. Without this restriction, investment and patient selection could respond non-monotonically to changes in regulatory requirements. To illustrate this, consider a scenario where profits are linear in terms of (r, I). Then the center's problem becomes mathematically equivalent to that of a firm choosing conditional input demands. In this analogy, S plays the role of the production function, while $|\pi_r|$ and $|\pi_I|$ correspond to the input prices for patient selection and investment, respectively. As in standard production theory, conditional input demands need not be monotonic in total output—inputs can be inferior factors. Hence, the supermodularity condition ensures that such nonmonotonicities cannot arise in this context.

Although investment and patient selection are the center's primitive choices, the optimal flagging-probability function depends only on the resulting indirect profit function. Recall that $\Pi(\theta, q, p)$ denotes the maximum profit attainable by a type-θ center that targets survival rate q under flagging probability p. This representation summarizes how each type trades off survival rate against flagging probability, abstracting from the underlying actions. The key object is the marginal rate of substitution between the target survival rate and the flagging probability:

$$M(\theta, q, p) \equiv -\frac{\Pi_q(\theta, q, p)}{\Pi_p(\theta, q, p)}.$$

The following proposition shows that this marginal rate of substitution is monotone in the center's type.

PROPOSITION 1 (Monotonic Marginal Rate of Substitution):
The marginal rate of substitution is increasing in the center's type, that is,

$$\frac{\partial M(\theta, q, p)}{\partial \theta} \geq 0. \tag{15}$$

Proposition 1 implies that higher-type centers are more willing to increase their

survival rate to reduce the probability of being flagged by a given amount. Two features of the model drive this result. First, center type θ affects the problem only through the survival constraint and not directly through profits, so higher types face a more relaxed problem and a lower shadow cost of improving survival rates. Second, supermodularity of $\widehat{S}$ and π delivers monotone comparative statics in the survival constraint, as shown in Lemma 1. As type increases, centers rely less on investment and exclude fewer patients to attain a given survival rate. Because flagging is less costly at higher levels of investment or patient selection, as imposed by assumption (4), its marginal impact on profit increases with type. Together, these forces imply that higher-type centers face a lower marginal cost of raising survival relative to the cost of flagging.

B. Implementation

Proposition 1 implies that higher center types have flatter indifference curves in the (q, p) plane. Standard incentive-compatibility arguments (Theorems 1 and 2 of Guesnerie and Laffont (1984)) therefore imply that any implementable flagging policy must assign weakly higher survival rates to higher types. A key difference between our environment and Guesnerie and Laffont (1984) is that transfers increase the regulator's payoff in their model while flagging probabilities reduce it in ours; this difference is irrelevant for implementability and matters only for optimality. We briefly sketch the intuition here and refer the reader to Guesnerie and Laffont (1984) for the proofs.

Figure 2 depicts the feasible set of survival–flagging pairs (q, p) together with the indifference curves of types $\theta \leq \theta'$ passing through $(Q^{**}(\theta), P^{**}(\theta))$. By Proposition 1, the indifference curve of the higher type θ' is locally flatter. Since profits are decreasing in both q and p, each type strictly prefers allocations located southwest of its indifference curve.

The blue-shaded region in the figure represents allocations strictly preferred by the higher type θ'. The red-shaded region corresponds to allocations strictly preferred by the lower type θ, and is therefore excluded by incentive compatibility. The set of incentive-compatible allocations for type θ' is the portion of the blue region that lies outside the red region (this is the blue hatched region). As illustrated, any such allocation involves a weakly higher survival rate and a weakly lower flagging probability than $(Q^{**}(\theta), P^{**}(\theta))$. The figure thus illustrates that incentive compatibility requires survival rates to be nondecreasing in type. We now formally state the result.

LEMMA 2 (Monotonicity):
Any flagging probability function induces a non-decreasing survival-rate function $Q(\theta)$.

This result indicates that, for any flagging-probability function, higher types must be assigned weakly higher survival rates. We next characterize flagging probability functions that implement a given nondecreasing survival-rate function.

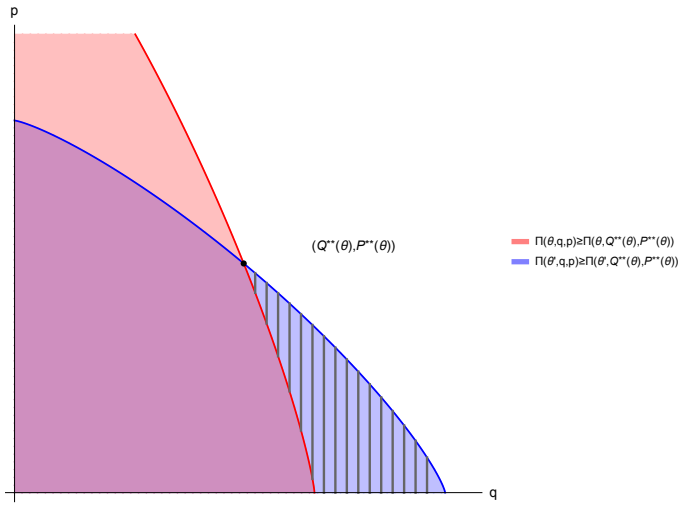


Figure 2. Indifference curves for a low-type and a high-type center in the (q, p) plane. The shaded region contains allocations preferred by the high type but not by the low type, implying that incentive-compatible contracts assign weakly higher survival rates and weakly lower flagging probabilities to higher types.

To build intuition, consider the center's problem under the optimal flagging policy,

$$\max_{q\in[0,1]} \Pi(\theta, q, R^{**}(q)).$$

Suppose that R^{**} is differentiable. At the optimal choice $q = Q^{**}(\theta)$, the first-order condition can be written as

$$R_q^{**}(Q^{**}(\theta)) = M(\theta, Q^{**}(\theta), P^{**}(\theta))\,.$$

This condition links the slope of the flagging-probability function to the center's marginal rate of substitution, thereby disciplining the growth of R^{**} along the equilibrium survival rates. Since flagging is costly, the optimal mechanism satisfies the boundary condition $R(Q(\bar{\theta})) = 0$, and the full flagging schedule is obtained by integrating this relationship backward from the top.

To formalize this observation, let Θ denote the generalized inverse of the survival-rate function Q:

$$\Theta(q) \triangleq \inf\{\theta \in [\underline{\theta}, \overline{\theta}] : q \leq Q(\theta)\}. \tag{16}$$

We can now characterize the flagging-probability functions that implement a given nondecreasing survival-rate function.

LEMMA 3 (Implementation):
If a nondecreasing survival-rate function Q is implementable, then the corresponding flagging-probability function R satisfies

$$R(q) = d - \int_q^{Q(\bar{\theta})} M(\Theta(x), x, R(x))\, dx, \tag{17}$$

for some nonnegative constant d. Furthermore, a nondecreasing survival-rate function Q is implementable if there exists R satisfying (17).

Lemma 3 provides a necessary condition characterizing the flagging-probability functions that implement a given nondecreasing survival-rate function. We use this characterization to establish our first main result (Proposition 2). The substantive part of the lemma is the necessary condition for optimality. The sufficient condition is less directly useful, because verifying it requires showing that the integral equation (17) admits a solution satisfying the constraint that flagging probabilities lie in $[0, 1]$. Thus, not every nondecreasing survival-rate function Q is implementable. However, if (17) can be satisfied (with the domain restriction), then the condition is also sufficient.[20] The constant d shifts the flagging-probability function uniformly. Because higher flagging probabilities reduce both the regulator's welfare and the center's profit, any optimal mechanism must set $d = 0$.

C. No Flagging at the Top of the Distribution

We present our first main result on the optimal flagging policy. It states that an optimal mechanism never flags high enough types.

PROPOSITION 2 (No Flagging at the Top):
There exists θ_f, with $\theta_f < \bar{\theta}$, such that $P^{**}(\theta) = 0$ if and only if $\theta \geq \theta_f$.

This proposition shows that in an optimal mechanism, there is a non-negligible set of types that are never flagged. The key aspect of the model that yields this result is that screening is conducted with a costly flagging policy. Because it is costly, any given survival-rate function $Q(\theta)$ is implemented with the minimum possible flagging probability, so naturally the highest type $\overline{\theta}$ is assigned zero flagging probability. Then, flagging probabilities of a given type $\widehat{\theta}$ are determined by

[20] Guesnerie and Laffont (1984) provides necessary and sufficient conditions for implementability in a standard screening environment. In our setting, implementability is further constrained by the requirement that flagging probabilities lie in $[0, 1]$. The necessary conditions remain applicable without modification; in particular, the integral equation (17) corresponds to equation (2.11) in Guesnerie and Laffont (1984). The domain restriction is not, in itself, substantive, as one can reparameterize– for example, by setting $\widetilde{p} = \log(p/(1-p))$, which maps $[0, 1]$ to $\mathbb{R}$. To obtain sufficiency, Guesnerie and Laffont (1984) additionally imposes condition (B), which ensures the existence of an R satisfying (17). If, instead, the existence of a solution to (17) is assumed directly (as in Lemma 3), then Q is implementable, and the argument follows as in Guesnerie and Laffont (1984).

the amount of screening there is between this type and the highest type (that is, it is determined by the whole path $\{Q(\theta)\}_{\theta\in[\widehat{\theta},\bar{\theta}]}$). This leads to an extensive-margin difference in the cost-benefit analysis, as we explain next.

An incentive-compatible mechanism that separates types in the neighborhood $(\bar{\theta}-\varepsilon,\bar{\theta})$ requires assigning distinct survival rates to these types. Because the flagging schedule is constructed recursively from the top (via (17)), such separation necessarily raises the flagging probabilities of all types below $\bar{\theta}-\varepsilon$. Thus, screening high types generates a first-order welfare loss affecting a large set of types, rather than a distortion localized near the top. By contrast, the benefits of screening at the top are second-order: they affect only the behavior of types in a small neighborhood near $\bar{\theta}$. Pooling these types slightly distorts their survival choices but strictly reduces the flagging probabilities for all lower types. Because welfare losses from flagging operate on the extensive margin, whereas gains from improved investment at the top operate on a small neighborhood, the former dominates when ε is small.

There are two technical aspects of the analysis that are needed to derive Proposition 2. First, the type space is compact, and the profit and survival functions are continuous, which, in turn, implies that the marginal rate of substitution is bounded. When type space is unbounded, the gains at the top of the distribution may become arbitrarily larger than the losses at the bottom of the distribution, so differences in the extensive margin are insufficient to guarantee pooling at the top of the distribution. Second, the indirect profit function exhibits a monotonic marginal rate of substitution (that is, (15)), which ensures that such variations can be made without violating the incentive compatibility constraints globally. All the assumptions we made about the model in Section I are used only to prove that (15) is satisfied (and that the marginal rate of substitution is bounded). Hence, the results extend to any model that yields an indirect profit function with a monotonic, bounded marginal rate of substitution, in which screening is conducted with a costly instrument.

Proposition 2 contrasts with classic screening models, which use transfers and in which screening typically occurs across the entire distribution. In those classic models, transfers are constructed using the Envelope theorem, starting from the lowest type. One might therefore expect a similar extensive-margin difference at the lowest part of the type distribution in this context. However, in the classic screening models, the transfer for the lowest type is not fixed. Instead, it is adjusted to leave the lowest type with zero informational rents. As a result, small changes to the screening policy at the bottom of the distribution yield only second-order effects on transfers paid by higher types. This second-order effect occurs because policy adjustments under a variational argument are small, and these are further weighted by the difference in marginal utility of the allocation, which is also small (rather than by a fixed transfer for low types).

III. Distortions

In this section, we examine the direction of the distortions induced by the optimal flagging policy. In particular, we examine when investment is insufficient or when patient selection is excessive. Here, the terms "insufficient" and "excessive" are relative to a benchmark in which the regulator can observe the center's type but cannot directly control its practices (i.e., investment and patient selection). We begin this section by making additional assumptions about our model; these assumptions are maintained throughout. We then provide properties of the regulator's indirect welfare function and formally introduce our benchmark. Finally, we provide properties of the distortions induced by the optimal flagging policy.

A. Additional Parametric Assumptions

We introduce additional assumptions regarding the profit, welfare, and survival functions.

We make the following assumption on the welfare function. There exist positive constants b and $c_\omega \geq 0$ such that:

$$w(\theta, r, I, p) = \left(\int_r^1 s(\phi, \theta, I) d\phi + \int_0^r b \, d\phi \right) - c_\omega p. \tag{18}$$

The interpretation is as follows. When a center is flagged, the regulator must conduct a review, incurring a pecuniary cost of c_ω. A mass $(1 - r)$ of patients are transplanted, with each patient generating a welfare of 1 if they survive and 0 if they do not survive.[21] A mass r of patients remain on dialysis and generate a welfare of b. Thus, the constant b represents a baseline welfare for removed patients, where we assume that $b \leq s(\phi, \theta, I)$ for all (ϕ, θ, I).[22] This functional form is employed to demonstrate the existence of a unique optimal survival rate that is independent of the flagging probability a center faces (Proposition 3).

We assume the profit function is as follows:

$$\pi(I, r, p) = \ (1 - r) - I - c_\pi p. \tag{19}$$

Here, the terms are interpreted as follows. The term $(1 - r)$ is the center's revenue; the term I is the center's investment cost; the term c_π captures the direct pecuniary cost of flagging. Note that it is without loss of generality to assume that profits decrease linearly with investment, as any nonlinearity can be

[21] Assuming welfare values of 1 and 0 for these outcomes is without loss of generality, because it corresponds to a normalization of the Bernoulli utility function.

[22] For simplicity, we assume b is identical for all patients (independent of ϕ). While we make this assumption for analytical convenience, it is a reasonable assumption since many key transplant risk factors– such as immunological sensitization (e.g., antibodies) and HLA compatibility– primarily affect post-transplant rather than life-quality under dialysis.

incorporated into the survival function. Importantly, we obtain that the center's choices do not depend on p.

LEMMA 4 (Center Choices):
Under (19), *the center choices* $(r^c(\theta,q,p), I^c(\theta,q,p))$ *are constant in* p.

We thus omit the argument p in the choices and write more compactly

$$(r^c(\theta,q), I^c(\theta,q)).$$

While we maintain (19) to fix ideas, the analysis in this section relies only on the assumption that profits are linear in investment I and patient selection r. The profit function need not be linear in the flagging probability p, and the marginal effects π_I and π_r may depend on p as long as the ratio π_r/π_I does not depend on p. What is important for the analysis is that Lemma 4 holds, as this is necessary for a uniquely defined optimal survival rate.

Finally, we assume that the iso-curves of the marginal rate of substitution of $\widehat{S}$ are quasi-convex, that is,

$$\left\{(r,I) : \frac{\widehat{S}_r(r,I)}{\widehat{S}_I(r,I)} \leq \alpha\right\} \text{ is convex.} \tag{20}$$

for every $\alpha \in \mathbb{R}_+$. That is, as survival requirements become more demanding, patient selection becomes increasingly effective relative to the investment required. This assumption is the most difficult to interpret, so we provide a more extended discussion in the following subsection. While it is difficult to pin down what constitutes a natural assumption on the survival function, we argue that the implications of this assumption are consistent with the available empirical evidence.

Throughout this section, we maintain these assumptions (that is, (18), (19), (20)).

B. Quasi-Convexity of Marginal Rate of Substitution

We can provide an alternative interpretation of (20) by considering the locus of optimal investment as a function of optimal patient selection. For this we implicitly define $\widehat{I}$ as follows:

$$\frac{\widehat{S}_r(r,\widehat{I}(r))}{\widehat{S}_I(r,\widehat{I}(r))} = \frac{\pi_r}{\pi_I}. \tag{21}$$

This is the first-order condition (14) (we recall that S is linear in θ so we can replace it in the first-order condition with $\widehat{S}$, and in this section, we assume that π is linear in (r,I) so we can omit the arguments on the right-hand side of (21)).

LEMMA 5 (Concavity of Locus Points):
The investment $\widehat{I}(r)$ is concave in r (for every π_r/π_I) if and only if (20) *is satisfied.*

This lemma provides a geometric interpretation of how the marginal rate of substitution for $\widehat{S}(r, I)$ changes with (r, I) when (20) holds. In Figure 3a, we plot the iso-curves of $\widehat{S}$ when it is a homothetic function. When $\widehat{S}$ is homothetic, the contour of the set in (20) is linear, and hence, the marginal rate of substitution stays constant along any ray starting at the origin. In Figure 3b, we plot the isocurves of $\widehat{S}$ when the inequality in (20) is strict. In this case, the curve capturing the constant marginal rate of substitution points of each iso-curve is concave.

The key implication of (20) is that, as the center is incentivized to select combinations of (r, I) that yield a higher survival rate $\widehat{S}$, patient selection becomes relatively more effective than investment in improving survival. Consequently, along the optimal locus, patient selection increases faster than investment.

LEMMA 6 (Convexity of Patient Selection):
If (20) *is satisfied then:*

$$\frac{\partial^2 r^c(\theta, q)}{\partial q^2} \geq 0. \tag{22}$$

Hence, the condition guarantees that the patient selection is convex with respect to the target survival rate. One way to interpret this condition is that there are decreasing returns to demanding a higher survival rate for any given type, because reliance on patient selection increases sufficiently quickly. The convexity of patient selection is the critical property we use in the analysis that follows. In fact, we could have omitted (20) as an assumption and assumed by fiat that (22) is satisfied. While it is difficult to evaluate what is a natural functional form for the survival rate, the empirical evidence suggests that (22) is satisfied.[23]

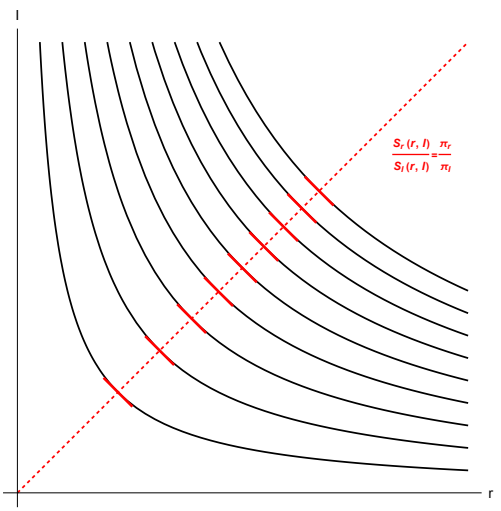
C. Properties of the Indirect Welfare Function

To continue our analysis, we define a benchmark welfare-maximizing survival rate:

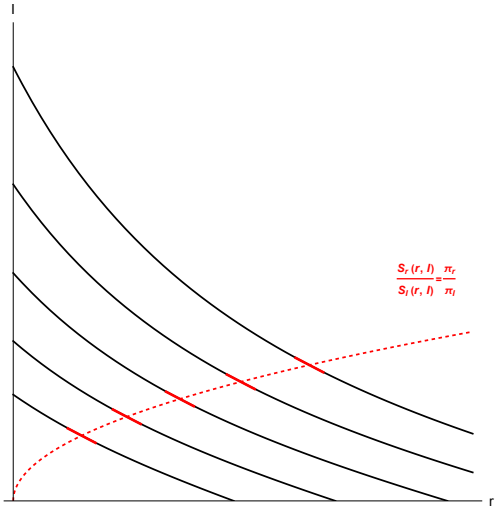
$$Q^*(\theta) \triangleq \arg\max_{q \in [0,1]} W(\theta, q, p),$$

where $W(\theta, q, p)$ is the indirect welfare function defined in (12). Here, we omit p as an argument of Q^* because we will show that the optimum is independent of p. The benchmark Q^* is the optimal survival rate the regulator would impose

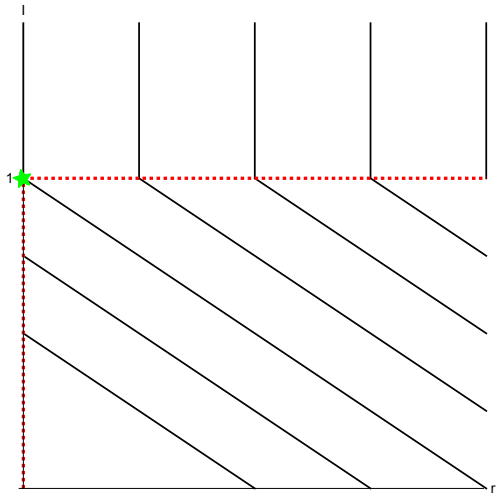
[23] White et al. (2015) and Bowring et al. (2018) document that transplant programs facing sustained or repeated regulatory scrutiny exhibit sharp reductions in transplant volume and more restrictive use of higher-risk organs relative to centers that receive only a single regulatory flag.



(a) Iso-survival curves for a homothetic survival function. Along each ray from the origin, the marginal rate of substitution between investment and patient selection remains constant.



(b) Iso-survival curves for a general survival function satisfying condition (20). The locus of points with a constant marginal rate of substitution is concave.



(c) Iso-survival curves for the survival function $s(\phi, I) = \phi + \min\{I, 1\}$. Investment increases survival up to a maximum of 1; beyond this point, further survival is achieved only through patient selection.

Figure 3. Iso-survival curves for three specifications of the survival function.

on the transplant center if the center's type were publicly known. However, the regulator could not observe patient selection and investment separately (hence, there is still room for moral hazard). We refer to this as the *constrained* first-best outcome; the qualifier "constrained" refers to the fact that in this benchmark, the regulator cannot optimally choose the center's actions but can observe its type.

PROPOSITION 3 (Properties of Welfare Function):
The welfare function $W(\theta, q, p)$ is strictly concave in q with a unique maximum at $Q^*(\theta)$, which is increasing in θ and constant in p.

Proposition 3 captures the essential elements of the model. Demanding a moderate survival rate is beneficial for welfare. After all, it drives investment, whereas being overly demanding is detrimental because it leads to patient removal. Hence, the regulator has an "ideal" survival rate for each center type, denoted $Q^*(\theta)$. Furthermore, the regulator seeks a higher survival rate for higher types. The final part of the theorem is not straightforward because low-type centers select more patients; it is natural to be more demanding of them, since they transplant patients with better characteristics (for example, if the welfare function is as in (18)). However, this effect is never dominant.

The sharpest intuition of Proposition 3 is obtained when the survival function is given by:[24]

$$\widehat{s}(\phi, I) = \phi + \min\{I, 1\}, \tag{24}$$

Hence, the average survival function becomes linear in (r, I) but I is capped at 1. In this case, the center invests when the target survival rate is relatively low; however, the investment is capped, after which the center begins selecting patients. This represents the starkest case, in which $Q^*(\theta)$– represented by a green star in Figure 3c– defines the critical threshold average survival rate. Beyond this point, the center raises its average survival rate through patient selection alone; below it, the average survival rate increases only through investment. More generally, under the constrained first-best survival rate, different types are induced to choose different combinations of patient selection and investment.

We now turn to the analysis of distortions at the boundaries of the type distribution. In this setting, (22) plays a key role in establishing that $Q^*(\theta)$ is increasing in the center type (Proposition 3). However, the key property for determining the direction of the distortions induced by the flagging probability function (Proposition 4) is the conclusion of Proposition 3 itself.

[24] This survival function is not differentiable, but we can obtain it as a limit of differentiable functions. A particular class of survival functions that satisfies (20) is of the form:

$$\widehat{s}(\phi, I) = \phi + (1 + I^{-\sigma})^{\frac{-1}{\sigma}}. \tag{23}$$

In the limit $\sigma \to \infty$ we obtain (24).

D. Direction of the Distortions

We now examine the distortions in an optimal mechanism; this will inform us which types have survival rates that are too high or too low relative to the welfare-maximizing rate $Q^*(\theta)$.

PROPOSITION 4 (Distortions in Survival Probability):
There exists $\theta_{d_1}, \theta_{d_2} \in (\underline{\theta}, \bar{\theta})$, *with* $\theta_{d_1} \leq \theta_{d_2}$, *such that* $Q^{**}(\theta) > Q^*(\theta)$ *for all* $\theta \in (\underline{\theta}, \theta_{d_1})$ *and* $Q^{**}(\theta) < Q^*(\theta)$ *for all* $\theta \in (\theta_{d_2}, \bar{\theta})$.

This proposition states that the optimal mechanism can distort the center's survival rate in two ways and determines the direction of the distortion based on the threshold type. We remark that we do not have a result for distortions for types in $(\theta_{d_1}, \theta_{d_2})$. If $\theta_{d_1} < \theta_{d_2}$, we might have that the types in $[\theta_{d_1}, \theta_{d_2}]$ are not distorted or $Q^{**}(\theta)$ cross $Q^*(\theta)$ multiple times. We present a representative optimal policy in Figure 4, which illustrates the trade-offs we have introduced thus far.

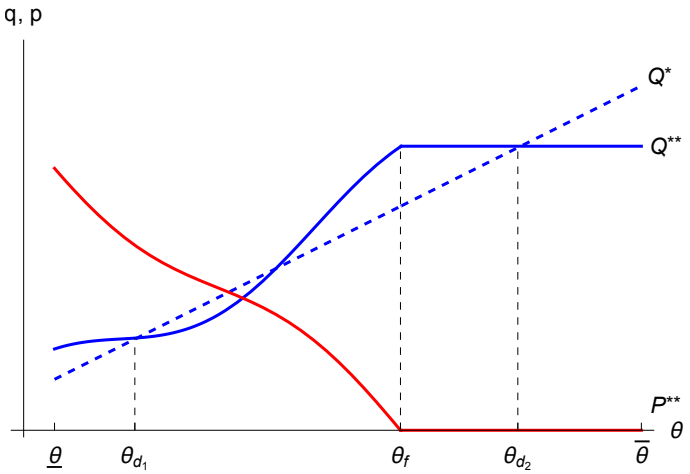


Figure 4. Schematic representation of the optimal mechanism. The constrained first-best survival schedule $Q^*(\theta)$ is strictly increasing in type. The optimal second-best schedule $Q^{**}(\theta)$ is flatter: it lies above $Q^*(\theta)$ for low types, reflecting upward distortion, and below $Q^*(\theta)$ for high types, reflecting downward distortion due to the high cost of screening at the top. The associated flagging probability $P^{**}(\theta)$ is positive only for a subset of types and becomes zero above the threshold θ_f.

To build intuition for Proposition 4, it is helpful to start from the constrained first-best benchmark, in which the regulator implements the welfare-maximizing survival schedule $Q^*(\theta)$. At $Q^*(\theta)$, each type chooses the optimal combination of

investment and patient selection. However, enforcing Q^* requires strong screening incentives: the regulator must rely heavily on flagging, which is costly in welfare terms. The regulator, therefore, has an incentive to relax screening by flattening the survival schedule.

A flatter schedule $Q^{**}(\theta)$ necessarily crosses $Q^*(\theta)$, implying upward distortions for some types and downward distortions for others. The proof shows that such distortions must occur near the extremes of the type distribution. At the top, Proposition 2 implies that high types are not flagged. Since $Q^*(\theta)$ is strictly increasing, the survival rates for types near $\bar{\theta}$ cannot coincide with $Q^*(\theta)$ without inducing positive flagging. Increasing survival targets for high types would only intensify screening and raise flagging probabilities, so high types are optimally distorted downward.

At the bottom of the type distribution, downward distortions are never optimal: lowering survival rates reduces welfare and does not relax the incentive constraints that bind at the top. In contrast, upward distortions can be beneficial. Increasing survival targets for low types makes their survival rates more similar to those of high types, thereby relaxing incentive constraints and allowing the regulator to reduce the costly flagging probability. At $Q^*(\theta)$, such upward deviations generate no first-order welfare loss, while the reduction in flagging yields a first-order welfare gain. Consequently, low types are distorted upward in the optimal mechanism.

We can then describe the implications of Proposition 4 for the center's choices. The optimal mechanism distorts types above the higher threshold (θ_{d_2}) downward; that is, it assigns survival rates below the constrained first-best benchmark to these types. These center types will be under-invested under the optimal flagging policy. However, it distorts types below the lower threshold (θ_{d_1}) upward; that is, it yields survival rates higher than the constrained first-best rates. These center types will select patients under the optimal flagging policy. To formalize this, we define:

$$(r^*(\theta), I^*(\theta)) \triangleq (r^c(\theta, Q^*(\theta)), I^c(\theta, Q^*(\theta)))$$
$$(r^{**}(\theta), I^{**}(\theta)) \triangleq (r^c(\theta, Q^{**}(\theta)), I^c(\theta, Q^{**}(\theta)))$$

where we omit the dependency on p on the right-hand side of the definition following Lemma 4. That is, $(r^*(\theta), I^*(\theta))$ and $(r^{**}(\theta), I^{**}(\theta))$ are the patient selection and investment under the first- and second-best policies. We now formalize the distortions under the optimal policy.

COROLLARY 1 (Distortions in Investment and Patient Selection):
For all $\theta < \theta_{d_1}$, $r^{**}(\theta) > r^*(\theta)$; for all $\theta > \theta_{d_2}$, $I^{**}(\theta) < I^*(\theta)$.

An immediate implication of Proposition 4 is that types in $(\bar{\theta} - \varepsilon, \bar{\theta}]$ must be distorted downward. This contrasts with classic mechanism-design environments with transfers; see, e.g., (Baron and Myerson 1982). In those environments,

distortions are monotone, and the highest type is undistorted ("no distortion at the top"). In our environment, flagging is costly for both parties. As a result, the optimal survival schedule is significantly flatter than the constrained first-best schedule: low types attain survival rates that are too high relative to Q^*, and high types attain survival rates that are too low.

Propositions 2–4 establish two qualitative features of the optimal mechanism: (i) pooling at the top of the type distribution, and (ii) distortions in survival rates in both directions. These results, however, do not quantify the extent of pooling or the magnitude of the distortions. We now briefly discuss these magnitudes and explain how they depend on the relative costs of flagging to the center and to the regulator. The relative cost of flagging for the center versus the regulator plays a central role in shaping the optimal mechanism. If flagging is very costly for the regulator (e.g., $c_\omega \gg 0$), the mechanism minimizes flagging, leading to substantial pooling. Conversely, if flagging is very costly for the transplant center (e.g., $c_\pi \gg 0$), the regulator relies heavily on screening incentives, and the resulting survival schedule is close to the constrained first-best. Thus, the magnitude of pooling and distortion is determined by the cost of flagging for each party.

Finally, we now relate our results to those in the literature. In Condorelli (2012), the optimal mechanism is complete pooling when the value distribution has an increasing hazard rate, thereby leading to upward and downward distortions. However, if the hazard rate is decreasing, the optimal mechanism is efficient. Hence, the type of distortion critically depends on the distribution of values (see the unified treatment in Kleiner, Moldovanu and Strack (2021), which finds the same pattern across a variety of settings). In our setting, the support of the distribution is compact, so the hazard rate is always increasing at the top. While they do not emphasize this, standard ironing techniques would allow one to show that pooling at the top always arises when the support of values is compact. Hence, our results are consistent with the ones in the literature. Here, our contribution is a novel variational argument that applies beyond quasilinear environments. Our result on distortion at the bottom of the distribution is incompatible with their corresponding result because we find there are always distortions at the bottom of the distribution, while they find that it is possible to have an efficient allocation at the bottom of the type distribution[25] The difference arises because in our model, the welfare function is concave and differentiable, so we can show that the loss due to distortions is always of second-order relative to the cost of additional screening. Hence, we can always sign the distortions when the effects of the distribution are negligible (at the lower part of the support) or when they are dominant (at the upper part of the support). Of course, our results are weaker than those in the literature, as they apply only to neighborhoods near the edge of support.

[25]Here, the decreasing hazard rate (which is necessary to get an efficient allocation in their setting) does not conflict with our assumption that the type space is compact.

IV. Application to Other Economic Environments

We have developed a model in which centers can improve their observed average survival rate either through welfare-enhancing investment or through patient selection. At the same time, the regulator has a costly oversight instrument– namely, flagging– to discipline the center's behavior. We now discuss additional applications of the model and interpret our results in these settings. Although the model does not apply to all cases in which agents manipulate benchmarks,[26] we argue that there are various instances where the three critical ingredients of our model are present: (i) institutions can influence measured performance through both welfare-improving and welfare-reducing actions, (ii) regulators rely on a costly oversight tool, and (iii) institutions are heterogeneous. We argue that these conditions characterize the oversight of charter schools and financial aid institutions, and that our framework captures the central trade-offs in these contexts.

A. Evaluation of Charter Schools

Most states have enacted laws requiring charter schools to close if they fail to meet specified performance goals (e.g., California, Texas, and New York). In most cases, a school is deemed a low performer if its accountability index (e.g., based on test scores and safety) falls below a specified quantile. However, closures are also costly for the regulator. In addition to the obvious bureaucratic and political costs, students displaced by closures may not end up in better-performing schools.[27] For this reason, typically, authorizers (who ultimately decide whether to close a school) are conservative in their closure decisions. Indeed, guidelines for authorizers in many states, e.g., Florida and North Carolina, require that they grant charter schools a reasonable 3- to 5-year period to remediate performance deficiencies. As a result, only a minimal number of low-performing schools are closed.[28]

A critical concern for regulators is that charter schools improve their performance records by engaging in various admission strategies.[29] Holding charter schools to high standards provides them with more substantial gaming incentives, which is undesirable from a social standpoint. Hence, the regulator must carefully balance the incentives provided by its performance standards.

We can thus conclude that the problem of evaluating charter schools shares

[26] Such behavior is often discussed under Goodhart's Law and, more broadly, relates to the Lucas critique.

[27] Han et al. (2017) find that more than half of the displaced students are transferred to a (weakly) worse-performing school.

[28] Han et al. (2017) find that on average, approximately $5 - 6\%$ of low-performing schools are closed.

[29] Lacireno-Paquet et al. (2002) show that charter schools may engage in selective admission to achieve higher test scores. Anecdotal evidence from the New York Times article "Bronx Charter School Disciplined over Admissions Methods" (July 20, 2011), for example, suggests that charter schools may strategically admit students to maintain strong academic records.

the same elements with the regulation of transplant centers: (a) closing and reviewing charter schools is costly for the regulator and charter schools, (b) being too demanding with charter schools is undesirable, and (c) charter schools are heterogeneous.

B. Development Banks and Financial Aid Institutions

Development banks and international financial aid institutions often condition their lending on performance metrics such as repayment rates, compliance with fiscal targets, and the speed of fund disbursement. Borrowers who fail to meet these targets may face sanctions, reduced future lending, or costly monitoring procedures. However, such monitoring is itself costly for the lender, both administratively and politically, and can reduce the flow of resources to intended development projects.

A critical concern for regulators is that development banks and aid institutions can improve their repayment records, compliance indicators, or disbursement times not by genuinely fostering better projects or strengthening institutional capacity, but by selecting safer borrowers or countries with lower default risk.[30] Holding development banks to stringent repayment, compliance, or disbursement standards provides them with stronger incentives to engage in such selection, which is undesirable from a global development perspective. Hence, the oversight must carefully balance the incentives provided by performance standards.

We can thus conclude that the problem of evaluating development banks shares the same elements with the regulation of transplant centers: (a) monitoring and sanctioning borrowers is costly for the institution, (b) being too demanding with performance standards can distort lending behavior in undesirable ways, and (c) borrowers are heterogeneous.

V. Concluding Remarks

We analyzed a model of a transplant center that can improve its survival rate through investment or patient selection. We derived qualitative predictions for the optimal flagging policy. The predictions are consistent with the observed practices.

First, the model explains why oversight regimes often feature relatively lenient performance thresholds. Proposition 2 shows that setting higher benchmarks for top performers inevitably increases the probability of flagging among lower types, thereby reducing overall welfare. This logic is consistent with the empirical

[30] For example, Dreher, Sturm and Vreeland (2009) provide evidence that IMF lending decisions are influenced by political considerations and borrower selection, rather than pure development needs. Anecdotal evidence also suggests that development banks may favor middle-income or geopolitically aligned countries with strong repayment prospects and fast disbursement capabilities, at the expense of riskier but needier borrowers.

observation that, during 2007–2015, roughly 40% of transplant centers faced zero probability of flagging, as documented by Kasiske et al. (2016).

Second, the model captures how strategic behavior emerges under oversight. Proposition 4 demonstrates that lower-type institutions "game the system" by engaging in practices that inflate reported outcomes while reducing welfare. By contrast, higher-type institutions do not resort to such practices, but they underinvest. These theoretical predictions align with evidence from the medical literature: Schold et al. (2016) and White et al. (2015) document that low-performing centers respond to oversight by becoming more selective, while high-performing centers neither alter patient selection nor change practices following performance-based reviews.

Finally, the model highlights the potential of random mechanisms in regulatory design. While not yet fully implemented, the OPTN has considered tiered review systems incorporating random audits of intermediate-performing centers (Kasiske et al. 2016). Such designs address the shortcomings of deterministic thresholds, which create strong incentives for gaming. Our analysis shows that randomization can encourage investment without inducing widespread cream-skimming, though at the cost of occasionally flagging high-performing institutions.[31]

Although our primary motivation is the oversight of transplant centers, the analysis highlights trade-offs that arise more broadly in regulatory settings. The common guiding principle is the use of costly institutional oversight to improve performance through investment or cream-skimming. Taken together, these findings underscore the importance of designing oversight policies that balance deterrence against poor performance, avoidance of perverse incentives, and promotion of genuine investment. The framework developed here provides both a rationale for existing practices and a guide for future reforms in transplant oversight and beyond.

REFERENCES

Akbarpour, Mohammad, Piotr Dworczak, and Scott Duke Kominers. 2024. "Redistributive allocation mechanisms." *Journal of Political Economy*, 132(6): 1831–1875.

Amador, Manuel, and Kyle Bagwell. 2013. "The theory of optimal delegation with an application to tariff caps." *Econometrica*, 81(4): 1541–1599.

Amador, Manuel, and Kyle Bagwell. 2020. "Money burning in the theory of delegation." *Games and Economic Behavior*.

Ambrus, Attila, and Georgy Egorov. 2017. "Delegation and nonmonetary incentives." *Journal of Economic Theory*, 171: 101 – 135.

[31] According to the OPTN proposal, the highest-tier institutions are never reviewed, the lowest-tier institutions are always reviewed, and the two middle tiers are randomly reviewed with 10% and 50% probability, respectively.

Andreoni, Kenneth A. 2020. "Now is the time for the Organ Procurement and Transplantation Network to change regulatory policy to effectively increase transplantation in the United States; Carpe Diem." *American Journal of Transplantation*, 20(8): 2026–2029.

Associated Press. 2007. "Calif. Hospital Settles Transplant Suit after 33 Die." *https://www.nbcnews.com/health/health-news/calif-hospital-settles-transplant-suit-after-33-die-flna1C9469999*, NBC News, July 26.

Baron, David P, and David Besanko. 1984. "Regulation, asymmetric information, and auditing." *The RAND Journal of Economics*, 447–470.

Baron, David P., and Roger B. Myerson. 1982. "Regulating a Monopolist with Unknown Costs." *Econometrica*, 50(4): 911–930.

Ben-Porath, Elchanan, Eddie Dekel, and Barton L Lipman. 2014. "Optimal allocation with costly verification." *American Economic Review*, 104(12): 3779–3813.

Border, Kim C, and Joel Sobel. 1987. "Samurai accountant: A theory of auditing and plunder." *The Review of economic studies*, 54(4): 525–540.

Bowring, Mary G, Allan B Massie, Rebecca Craig-Schapiro, Dorry L Segev, and Lauren Hersch Nicholas. 2018. "Kidney offer acceptance at programs undergoing a Systems Improvement Agreement." *American Journal of Transplantation*, 18(9): 2182–2188.

Carlier, Guillaume, and Kelvin Shuangjian Zhang. 2020. "Existence of solutions to principal-agent problems with adverse selection under minimal assumptions." *Journal of Mathematical Economics*.

Centers for Medicare & Medicaid Services. 2019. "Advance copy: Revisions to transplant center oversight." *Quality, Safety & Oversight Memorandum QSO-19-11-Transplant*, U.S. Department of Health and Human Services.

Centers for Medicare & Medicaid Services. 2024. "Revisions to transplant center oversight and conditions of participation." *Quality, Safety & Oversight Memorandum QSO-25-03-Transplant*, U.S. Department of Health and Human Services.

Chakravarty, Surajeet, and Todd R Kaplan. 2013. "Optimal allocation without transfer payments." *Games and Economic Behavior*, 77(1): 1–20.

Chan, Alex. 2021. "Cherry Picking and Lemon Dropping to Meet Quality Standards." *Unpublished Working Paper*.

Chan, Alex, and Alvin E Roth. 2024. "Regulation of Organ Transplantation and Procurement: A Market-Design Lab Experiment." *Journal of Political Economy*, 132(11): 3827–3866.

Chandraker, Anil, Kenneth A Andreoni, Robert S Gaston, John Gill, Jayme E Locke, Amit K Mathur, Douglas J Norman, Rachel E Patzer, Abbas Rana, Lloyd E Ratner, et al. 2019. "Time for reform in transplant

program–specific reporting: AST/ASTS transplant metrics taskforce." *American Journal of Transplantation*, 19(7): 1888–1895.

Chaudhry, Daoud, Abdullah Chaudhry, Javeria Peracha, and Adnan Sharif. 2022. "Survival for waitlisted kidney failure patients receiving transplantation versus remaining on waiting list: systematic review and meta-analysis." *Bmj*, 376.

Condorelli, Daniele. 2012. "What money can't buy: Efficient mechanism design with costly signals." *Games and Economic Behavior*, 75(2): 613–624.

Dreher, Axel, Jan-Egbert Sturm, and James Raymond Vreeland. 2009. "Development aid and international politics: Does membership on the UN Security Council influence World Bank decisions?" *Journal of Development Economics*, 88(1): 1–18.

Dworczak, Piotr. 2023. "Equity-efficiency trade-off in quasi-linear environments." *Unpublished working paper. https://bpb-us-e1. wpmucdn. com/sites. northwestern. edu/dist/1/2465/files/2023/08/Equity-efficiency-trade-off. pdf.*

Dworczak, Piotr, Scott Duke Kominers, and Mohammad Akbarpour. 2021. "Redistribution through markets." *Econometrica*, 89(4): 1665–1698.

Guesnerie, Roger, and Jean-Jacques Laffont. 1984. "A Complete Solution to a Class of Principal-Agent Problems with an Application to the Control of a Self-Managed Firm." *Journal of public Economics*, 25(3): 329–369.

Han, C, ME Raymond, JL Woodworth, Y Negassi, DP Richardson, and W Snow. 2017. "Lights off: Practice and impact of closing low-performing schools, 2017, Vol. I." *Stanford, CA: CREDO*.

Hartline, Jason D, and Tim Roughgarden. 2008. "Optimal mechanism design and money burning." 75–84.

Henry, Emeric, and Marco Ottaviani. 2019. "Research and the approval process: The organization of persuasion." *American Economic Review*, 109(3): 911–55.

Ho, Bing, Anton I Skaro, and Michael M Abecassis. 2015. "Should both UNOS and CMS provide regulatory oversight in kidney transplantation?" *Current Transplantation Reports*, 2(2): 127–134.

Ison, Michael G, Jane L Holl, and Daniela Ladner. 2012. "Preventable errors in organ transplantation: an emerging patient safety issue?" *American Journal of Transplantation*, 12(9): 2307–2312.

Jay, Colleen, and Jesse D Schold. 2017. "Measuring transplant center performance: the goals are not controversial but the methods and consequences can be." *Current transplantation reports*, 4: 52–58.

Jochmans, Ina, Cyril Moers, Jacqueline M Smits, Henri GD Leuvenink, Jürgen Treckmann, Andreas Paul, Axel Rahmel, Jean-Paul Squifflet, Ernest van Heurn, Diethard Monbaliu, et al. 2010. "Machine perfusion versus cold storage for the preservation of kidneys donated after car-

diac death: a multicenter, randomized, controlled trial." *Annals of surgery*, 252(5): 756–764.

Kasiske, BL, N Salkowski, A Wey, AK Israni, and JJ Snyder. 2016. "Potential implications of recent and proposed changes in the regulatory oversight of solid organ transplantation in the United States." *American Journal of Transplantation*, 16(12): 3371–3377.

Kleiner, Andreas, Benny Moldovanu, and Philipp Strack. 2021. "Extreme points and majorization: Economic applications." *Econometrica*, 89(4): 1557–1593.

Lacireno-Paquet, Natalie, Thomas T. Holyoke, Michele Moser, and Jeffrey R. Henig. 2002. "Creaming Versus Cropping: Charter School Enrollment Practices in Response to Market Incentives." *Educational Evaluation and Policy Analysis*, 24(2): 145–158.

Laffont, Jean-Jacques, and Jean Tirole. 1986. "Using cost observation to regulate firms." *Journal of political Economy*, 94(3, Part 1): 614–641.

Lewis, Tracy R, and David EM Sappington. 1989*a*. "Countervailing incentives in agency problems." *Journal of economic theory*, 49(2): 294–313.

Lewis, Tracy R, and David EM Sappington. 1989*b*. "Inflexible rules in incentive problems." *The American Economic Review*, 69–84.

Maggi, Giovanni, and Andres Rodriguez-Clare. 1995. "On countervailing incentives." *Journal of Economic Theory*, 66(1): 238–263.

McAfee, R Preston, and John McMillan. 1992. "Bidding rings." *The American Economic Review*, 579–599.

Mirrlees, James A. 1971. "An exploration in the theory of optimum income taxation." *The Review of Economic Studies*, 38(2): 175–208.

Mookherjee, Dilip, and Ivan PL Png. 1994. "Marginal deterrence in enforcement of law." *Journal of Political Economy*, 102(5): 1039–1066.

Myerson, Roger B. 1981. "Optimal auction design." *Mathematics of operations research*, 6(1): 58–73.

Patzer, Rachel E, and Stephen O Pastan. 2020. "Policies to promote timely referral for kidney transplantation." Vol. 33, 58–67, Wiley Online Library.

Phillips, Anna M. 2011. "Bronx Charter School Disciplined over Admissions Methods." *New York Times*, July 20. `https://www.nytimes.com/2011/07/21/nyregion/bronx-charter-school-disciplined-over-admissions.html`.

Roth, Alvin E, Tayfun Sönmez, and M Utku Ünver. 2004. "Kidney exchange." *The Quarterly journal of economics*, 119(2): 457–488.

Schold, JD, LD Buccini, ED Poggio, SM Flechner, and DA Goldfarb. 2016. "Association of candidate removals from the kidney transplant waiting list and center performance oversight." *American Journal of transplantation*, 16(4): 1276–1284.

Schold, Jesse D. 2020. "The evolving role of regulatory reporting on patient

and donor selection in organ transplantation." *Current Opinion in Organ Transplantation*, 25(2): 158–162.

Schold, Jesse D, Laura D Buccini, Tulasi R Srinivas, Emilio D Poggio, David A Goldfarb, and Stuart M Flechner. 2013. "The association of center performance evaluations and kidney transplant volume in the United States." *American Journal of Transplantation*, 13(1): 67–75.

Stith, Sarah S, and Richard A Hirth. 2016. "The effect of performance standards on health care provider behavior: evidence from kidney transplantation." *Journal of Economics & Management Strategy*, 25(4): 789–825.

Tonelli, M, N Wiebe, G Knoll, A Bello, S Browne, D Jadhav, S Klarenbach, and J Gill. 2011. "Systematic review: kidney transplantation compared with dialysis in clinically relevant outcomes." *American journal of transplantation*, 11(10): 2093–2109.

White, Sarah L, Dawn M Zinsser, Matthew Paul, Gregory N Levine, Tempie Shearon, Valarie B Ashby, John C Magee, Yi Li, and Alan B Leichtman. 2015. "Patient selection and volume in the era surrounding implementation of Medicare conditions of participation for transplant programs." *Health services research*, 50(2): 330–350.

Wolfe, Robert A, Valarie B Ashby, Edgar L Milford, Akinlolu O Ojo, Robert E Ettenger, Lawrence YC Agodoa, Philip J Held, and Friedrich K Port. 1999. "Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant." *New England journal of medicine*, 341(23): 1725–1730.

Woodside, Kenneth J, and Randall S Sung. 2016. "Do federal regulations have an impact on kidney transplant outcomes?" *Advances in Chronic Kidney Disease*, 23(5): 332–339.

Mathematical Appendix

We begin by establishing and proving the existence of an optimal policy.

LEMMA A1 (Existence):
An optimal flagging-probability function R^{**} exists.

PROOF OF LEMMA A1:

In our model, there are no individual rationality constraints, and the set of feasible contracts $[\underline{q}, \bar{q}] \times [0, 1]$ is compact. Hence, we can apply Theorem 1 in (Carlier and Zhang 2020) to conclude that an optimal flagging probability function exists.[32]

PROOF OF LEMMA 1:

[32]To verify that this result carries through, one only needs to check that the coercivity assumption holds. However, because in our model there are no outside options, we can take $K = [\underline{q}, \bar{q}] \times [0, 1]$ (with K defined as in Remark 3.2 therein).

We first note that the profit function is strictly decreasing in (r, I), while the average survival rate is strictly increasing in the same variables. Hence, the constraint will be binding whenever $(r, I) \neq 0$. The Lagrangean of the maximization problem (10) is given by:

$$L(r, I, \lambda) \triangleq \pi(r, I, p) + \lambda \left(\widehat{S}(r, I) - \frac{q}{\theta} \right).$$

The first-order conditions are thus given by:

$$\pi_I(r^c, I^c, p) + \lambda^* \widehat{S}_I(r^c, I^c) = 0; \tag{A1}$$

$$\pi_r(r^c, I^c, p) + \lambda^* \widehat{S}_r(r^c, I^c) = 0; \tag{A2}$$

$$\widehat{S}(r^c, I^c) - \frac{q}{\theta} = 0 \tag{A3}$$

for some $\lambda^* \geq 0$. Here we write (r^c, I^c) as shorthand notation for $(r^c(\theta, q, p), I^c(\theta, q, p))$, that is, we omit the argument of variables to make the notation more compact.

Since $\pi, \widehat{S}$ are concave and supermodular in (r, I) we have that L is also concave and supermodular for any $\lambda \geq 0$. We denote by $(\widehat{r}(\lambda), \widehat{I}(\lambda))$ the solution of (A1)-(A2) with λ being a free variable, that is,

$$L_I(\widehat{r}(\lambda), \widehat{I}(\lambda), \lambda) = 0;$$
$$L_r(\widehat{r}(\lambda), \widehat{I}(\lambda), \lambda) = 0.$$

Taking the derivatives of (A1)-(A2) with respect to λ we obtain:

$$\begin{pmatrix} \frac{d\widehat{r}(\lambda)}{d\lambda} \\ \frac{d\widehat{I}(\lambda)}{d\lambda} \end{pmatrix} = \frac{1}{L_{II}L_{rr} - L_{Ir}^2} \begin{pmatrix} -L_{II} & L_{Ir} \\ L_{Ir} & -L_{rr} \end{pmatrix} \cdot \begin{pmatrix} \widehat{S}_r(r(\lambda), I(\lambda)) \\ \widehat{S}_I(r(\lambda), I(\lambda)) \end{pmatrix} \tag{A4}$$

Since L is concave, we have that the diagonal elements of the matrix are positive and:

$$L_{II}L_{rr} - L_{Ir}^2 \geq 0.$$

Since L is supermodular, the off-diagonal elements of the matrix are also positive. Finally, by assumption, $\widehat{S}$ is increasing in both arguments, so the first partial derivatives are positive. Hence,

$$\frac{d\widehat{r}(\lambda)}{d\lambda}, \frac{d\widehat{I}(\lambda)}{d\lambda} \geq 0.$$

Since both $\widehat{r}(\lambda)$ and $\widehat{I}(\lambda)$ are increasing in λ, the effect of q on $(r^c(\theta, q, p), I^c(\theta, q, p))$

is determined by the sign of how λ^* changes with q, that is, $d\lambda^*/dq$. In principle, both components could move in either direction with q. However, the constraint requires at least one of them to increase with q; otherwise, the constraint would either be violated or fail to bind. It follows that

$$\frac{d\lambda^*}{dq} \geq 0 \quad \text{and} \quad \frac{d\lambda^*}{d\theta} \leq 0. \tag{A5}$$

Therefore, $(r^c(\theta,q,p), I^c(\theta,q,p))$ are both increasing in q (and, consequently, also decreasing in θ).

PROOF OF PROPOSITION 1:

We compute how the marginal rate of substitution changes with θ:

$$\frac{\partial M(\theta,q,p)}{\partial \theta} = \frac{-\Pi_{q\theta}(\theta,q,p)\Pi_p(\theta,q,p) + \Pi_q(\theta,q,p)\Pi_{p\theta}(\theta,q,p)}{\Pi_p^2(\theta,q,p)}. \tag{A6}$$

which we now show is monotonic in θ. For this, we use the Envelope theorem to find that:

$$\begin{aligned}\Pi_q(\theta,q,p) =& -\frac{\lambda^*}{\theta} \leq 0;\\ \Pi_p(\theta,q,p) =& \pi_p(r^c(\theta,q,p), I^c(\theta,q,p), p) < 0.\end{aligned}$$

where the first inequality follows from the fact that the Lagrange multiplier is always positive, while the second one follows from (2). For the second derivatives, we obtain:

$$\begin{aligned}\Pi_{q\theta}(\theta,q,p) =& -\frac{1}{\theta}\frac{\partial \lambda^*}{\partial \theta} + \frac{\lambda^*}{\theta^2} \geq 0;\\ \Pi_{p\theta}(\theta,q,p) =& \pi_{pr}(r^c, I^c)\frac{\partial r^c}{\partial \theta} + \pi_{pI}(r^c, I^c)\frac{\partial I^c}{\partial \theta} \leq 0\end{aligned}$$

We obtain the first inequality using (A5); the second inequality follows from (4) and Lemma 1. Hence, both terms in the numerator of (A6) are positive, which proves the result.

Before proceeding with the proofs, we provide sufficient conditions under which a nondecreasing survival-rate function Q is implementable. These conditions apply to variations of an already implementable function Q, and ensure that the resulting survival-rate function remains implementable.

LEMMA A2 (Implementability of Truncations):

If Q is implementable, then any truncation of Q to the domain $[q_1, q_2]$ is implementable, that is, the survival-rate function:

$$\widehat{Q}(\theta) = \min\{\max\{Q(\theta), q_1\}, q_2\}$$

is implementable for any q_1, q_2.

PROOF OF LEMMA A2:

Consider any Q that is implementable, and hence, nondecreasing. We denote by $P(\theta)$ the corresponding probabilities that implement Q, and hence, $((Q(\theta), P(\theta))$ is an incentive compatible mechanism. That is, by definition,

$$\theta \in \arg\max_{\theta'} \ \Pi(\theta, Q(\theta'), P(\theta')).$$

We denote by (θ_1, θ_2) the maximal domain such that $Q(\theta) \in [q_1, q_2]$, that is, $Q(\theta) \in [q_1, q_2]$ if and only if $\theta \in [\theta_1, \theta_2]$. We then consider the following mechanism:

$$(\widehat{Q}(\theta), \widehat{P}(\theta)) = \begin{cases} (Q(\theta_1), P(\theta_1)) & \text{if } \theta < \theta_1 \\ (Q(\theta), P(\theta)) & \text{if } \theta \in [\theta_1, \theta_2] \\ (Q(\theta_2), P(\theta_2)) & \text{if } \theta > \theta_2 \end{cases}$$

For every $\theta \in [\theta_1, \theta_2]$, $(\widehat{Q}, \widehat{P})$ is trivially incentive compatible because $(Q(\theta), P(\theta))$ is incentive compatible. For $\theta \notin [\theta_1, \theta_2]$, incentive compatibility follows from the fact that the marginal rate of substitution is monotonic in type (see Proposition 1). Hence, $(\widehat{Q}(\theta), \widehat{P}(\theta))$ is also an incentive-compatible mechanism.

PROOF OF PROPOSITION 2:

We fix an optimal mechanism denoted by $Q^{**}(\theta)$ and denote by θ_f the minimum type that is never flagged:

$$\theta_f = \min\{\theta \in [\underline{\theta}, \bar{\theta}] : P^{**}(\theta) = 0\}.$$

If $P^{**}(\theta)$ is strictly increasing in a neighborhood around $\bar{\theta}$ (that is, $\theta_f = \bar{\theta}$) then we must have that $Q^{**}(\theta)$ is strictly increasing in some neighborhood $(\bar{\theta} - \epsilon, \bar{\theta})$.

Consider the following class of survival-rate functions parameterized by $z \in [Q^{**}(\underline{\theta}), Q^{**}(\overline{\theta})]$:

$$\widehat{Q}(\theta) \triangleq \min\{Q^{**}(\theta), z\} \tag{A7}$$

In other words, we take the minimum between $Q^{**}(\theta)$ and z. Because Q^{**} is nondecreasing, $\widehat{Q}$ is also nondecreasing. We remark that $\widehat{Q}$ is a truncation of Q^{**}, so following Lemma A2 it is implementable. We define:

$$\theta_f(z) \triangleq \Theta^{**}(z).$$

with $\Theta^{**}(\theta)$ defined as in (16).

We denote by $\widehat{R}$ the corresponding flagging-probability function defined as in (17) with $d = 0$ and the corresponding incentive compatible mechanism $(\widehat{Q}(\theta), \widehat{P}(\theta))$. Of course, $(\widehat{Q}(\theta), \widehat{P}(\theta))$ depend on z but we omit an explicit reference to z in the

notation to make this more compact. We can write the expected welfare generated by $\widehat{Q}$ as follows:

$$\mathbb{E}[W] \triangleq \int_{\underline{\theta}}^{\theta_f(z)} W(x, Q^{**}(x), \widehat{P}(x))dF(x) + \int_{\theta_f(z)}^{\bar{\theta}} W(x, z, 0)dF(x). \tag{A8}$$

By taking the derivative with respect to z and evaluating at $z = Q^{**}(\bar{\theta})$, we obtain:

$$\begin{aligned}\frac{d\mathbb{E}[W]}{dz} \mid_{z=Q^{**}(\bar{\theta})} = &\int_{\underline{\theta}}^{\theta_f} W_p(x, Q^{**}(x), P^{**}(x)) \frac{d\widehat{P}(x)}{dz} \mid_{z=Q^{**}(\bar{\theta})} dF(x) \\ &+ \int_{\theta_f}^{\bar{\theta}} W_q(x, Q^{**}(\bar{\theta}), 0)dF(x).\end{aligned} \tag{A9}$$

To obtain this expression, it is important to observe that the derivatives of the limits of integration are either zero or cancel out; we now explain the reason for this. By definition, $Q^{**}(\theta) < Q^{**}(\bar{\theta})$ for all $\theta \in [\underline{\theta}, \theta_f)$. That is, Q^{**} is strictly increasing or discontinuous at θ_f. If Q^{**} is discontinuous at θ_f, then $\theta_f(z) = \theta_f$ for all z in a neighborhood around $Q^{**}(\bar{\theta})$. Thus, in this case, the derivative of $\theta_f(z)$ with respect to z is zero at $z = Q^{**}(\bar{\theta})$. If Q^{**} is continuous at θ_f, then $\theta_f(q)$ is continuous for all q in a neighborhood around $Q^{**}(\bar{\theta})$. Still, since $(Q^{**}(\theta), P^{**}(\theta))$ is continuous, the derivatives of the limits in both integrals cancel out.

Now, as mentioned above, $\widehat{Q}$ is a truncation, so by Lemma A2 it is implementable. Hence, the corresponding first-order condition is:

$$\frac{d\mathbb{E}[W]}{dz}\Big|_{z=Q^{**}(\bar{\theta})} \geq 0 \quad \text{if } Q^{**}(\bar{\theta}) \in (0, 1]. \tag{A10}$$

The inequality (rather than equality) arises because the admissible variation is one-sided: Lemma A2 applies only when $\widehat{Q}$ is a truncation of Q^{**}, which requires $z \leq Q^{**}(\bar{\theta})$. Thus, only downward deviations are feasible at this point. If instead $Q^{**}(\bar{\theta}) = 0$, no such variation is admissible, implying that the survival rate function must be identically zero. We will use this condition to show that $\theta_f = \bar{\theta}$ cannot be optimal.

If $\theta_f = \bar{\theta}$, then we have the following:

$$\frac{d\mathbb{E}[W]}{dz} \mid_{z=Q^{**}(\bar{\theta})} = \int_{\underline{\theta}}^{\bar{\theta}} W_p(x, Q^{**}(x), P^{**}(x)) \frac{d\widehat{P}(x)}{dz} dF(x) < 0,$$

where the inequality follows from the fact that $W_p(\theta, q, p) < 0$ for all (θ, q, p) and

following (17), the flagging-probability function can be written as follows:

$$\widehat{R}(q) = -\int_{q}^{z} M(\theta_f(x), x, R(x))dx,$$

We thus have that

$$\frac{d\widehat{P}(x)}{dz} \mid_{z=Q^{**}(\bar{\theta})} > 0.$$

Hence, if $\theta_f = \bar{\theta}$, optimality condition (A10) requires that $Q(\bar{\theta}) = 0$. However, in this case, the mechanism is full bunching, so $\theta_f = \underline{\theta}$. Thus, we reach a contradiction, so $\theta_f = \bar{\theta}$ is never optimal.

PROOF OF LEMMA 4:

The proof follows directly from the fact that the ratio π_r/π_I does not depend on p.

PROOF OF LEMMA 5:

This result follows directly from finding the second derivative of $\widehat{I}(r)$ using (21) and the implicit function theorem for the second derivative.

PROOF OF LEMMA 6:

For any q we must have that:

$$S(\theta, r^c, I^c) = q,$$

where we omit the arguments of $r^c,\ I^c$ to make the notation more compact. We can write this as follows:

$$S(\theta, r^c, \widehat{I}(r^c)) = q$$

The second derivative with respect to q of this equation can be written as follows:

$$\left(\frac{\partial r^c}{\partial q}\right)^2 \left(\begin{pmatrix} 1 & \widehat{I}' \end{pmatrix} \cdot \begin{pmatrix} S_{rr} & S_{rI} \\ S_{rI} & S_{II} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \widehat{I}' \end{pmatrix} \right) + (S_r + S_I \widehat{I}') \frac{\partial^2 r^c}{\partial q^2} + S_I \widehat{I}'' \left(\frac{\partial r^c}{\partial q}\right)^2 = 0.$$

The first term is a quadratic form, and S is concave, so it is negative. The third term is negative because $\widehat{I}'' < 0$ by Lemma 5. Hence, the second term must be positive. However, we know that $(S_r + S_I \widehat{I}') \geq 0$, so we must have that:

$$\frac{\partial^2 r^c}{\partial q^2} > 0,$$

which concludes the proof.

PROOF OF PROPOSITION 3:

We can write the indirect welfare function as follows:

$$W(\theta, q, p) = (1 - r^c(\theta, q))(q - b) - c_\omega p + b,$$

where we omit the dependency of r^c on p following Lemma 4. We first check that W is concave in q and then show it is supermodular in (θ, q). The derivatives are given by:

$$\begin{aligned} W_{qq}(\theta, q, p) =& -2r^c_q(\theta, q) - (q-b)r^c_{qq}(\theta, q) < 0; \\ W_{q\theta}(\theta, q, p) =& -r^c_\theta(\theta, q) - (q-b)r^c_{q\theta}(\theta, q) > 0. \end{aligned}$$

For the first inequality, we use Lemma 1 and Lemma 6. For the second inequality, we use Lemma 1 and the fact that the constraint in (10) depends on q/θ so Lemma 6 implies that $r^c(\theta, q)$ is submodular in (q, θ). Hence, Q^* is increasing in θ. Finally, note that the choices (r^c, I^c) do not depend on p and the welfare is linear in p. Hence, Q^* does not depend on p.

PROOF OF PROPOSITION 4:

We first prove there exists an interval $[\bar{\theta} - \epsilon, \bar{\theta}]$ such that $Q^{**}(\theta) < Q^*(\theta)$. To prove this, we consider the same variation $\widehat{Q}$ as in (A7). If $Q^{**}(\theta) \geq Q^*(\theta)$ for all $\theta \in [\bar{\theta} - \epsilon, \bar{\theta}]$, we have that $W_q(\theta, Q^{**}(\bar{\theta}), 0) \leq 0$ for all $\theta \in [\bar{\theta} - \epsilon, \bar{\theta}]$. In this case, both terms on the right-hand side of (A9) are weakly negative, with at least one term being strictly negative. However, the optimality condition then requires that $Q^{**}(\bar{\theta}) = 0$. Hence, we reach a contradiction. We thus have $Q^{**}(\theta) < Q^*(\theta)$ for all $\theta \in [\bar{\theta} - \epsilon, \bar{\theta}]$.

We now prove there exists an interval $(\underline{\theta}, \underline{\theta} + \epsilon)$ such that $Q^{**}(\theta) > Q^*(\theta)$ for all $\theta \in (\underline{\theta}, \underline{\theta} + \epsilon)$. For this, we define:

$$\widetilde{Q}(\theta) \triangleq \max\{Q^{**}(\theta), Q^{**}(\underline{\theta}) + \epsilon\}.$$

We note that $\widetilde{Q}$ is a truncation of Q^{**} so, following Lemma A2, it is implementable. We prove that, if $Q^{**}(\underline{\theta}) \leq Q^*(\underline{\theta})$, then $\widetilde{Q}$ generates higher welfare, thus reaching a contradiction.

The expected welfare generated by this policy is:

$$\mathbb{E}[W] \triangleq \int_{\underline{\theta}}^{\bar{\theta}} W(x, \widetilde{Q}(x), \widetilde{P}(x)) dF(x).$$

We define $\widehat{\theta}$ as follows:

$$\widehat{\theta} \triangleq \max\{\theta \in \Theta : Q^{**}(\theta') < \widetilde{Q}(\underline{\theta}) \text{ for all } \theta' < \theta\}.$$

That is, $\widetilde{Q}(\theta) \neq Q^{**}(\theta)$ if and only if $\theta \in [\underline{\theta}, \widehat{\theta}]$. To compute the derivative, we write the expected welfare as follows:

$$\mathbb{E}[W] = \int_{\underline{\theta}}^{\widehat{\theta}} W(x, \widetilde{Q}(x), \widetilde{P}(x)) dF(x) + \int_{\widehat{\theta}}^{\bar{\theta}} W(x, Q^{**}(x), P^{**}(x)) dF(x).$$

Thus, the derivative is:

$$\frac{d\mathbb{E}[W]}{d\epsilon} = \int_{\underline{\theta}}^{\widehat{\theta}} \frac{\partial W(x, \widetilde{Q}(x), \widetilde{P}(x))}{\partial q} \frac{\partial \widetilde{Q}(x)}{\partial \epsilon} + \frac{\partial W(x, \widetilde{Q}(x), \widetilde{P}(x))}{\partial p} \frac{\partial \widetilde{P}(x)}{\partial \epsilon} dF(x).$$

Here we used that, if $Q^*(x)$ is continuous at $x = \widehat{\theta}$, then the derivative of the limits of integration with respect to $\widehat{\theta}$ cancel out; if $Q^*(x)$ is discontinuous at $x = \widehat{\theta}$, then the derivative of $\widehat{\theta}$ with respect to ϵ is zero. In both cases, the derivative of $\widehat{\theta}$ with respect to ϵ does not appear in the derivative.

We first note that if $Q^{**}(\underline{\theta}) < Q^*(\underline{\theta})$, then both terms in the integrand are positive for a small enough ϵ. This follows from the fact that $\partial W/\partial q$ is positive for a small enough ϵ because by assumption $Q^{**}(\underline{\theta}) < Q^*(\underline{\theta})$ and $\partial \widetilde{P}/\partial \epsilon$ is negative (which is multiplied times a negative term).

We now assume that $Q^{**}(\underline{\theta}) = Q^*(\underline{\theta})$ and also reach a contradiction. We denote by $\widetilde{\theta}$ the minimum θ such that $\widetilde{Q}(\theta) < Q^*(\theta)$:

$$\widetilde{\theta} = \min\{\theta \in [\underline{\theta}, \widehat{\theta}] : \widetilde{Q}(\theta) < Q^*(\theta)\}.$$

We now note that $\widetilde{Q}(\widetilde{\theta}) = Q^*(\widetilde{\theta})$, so the integrand is strictly positive for all $\theta \in (\widetilde{\theta}, \widehat{\theta}]$ (which follows the same logic as the case $Q^{**}(\underline{\theta}) < Q^*(\underline{\theta})$). We thus have that:

(A11)
$$\frac{d\mathbb{E}[W]}{d\epsilon} \geq \int_{\underline{\theta}}^{\widetilde{\theta}} \frac{\partial W(x, \widetilde{Q}(x), \widetilde{P}(x))}{\partial q} \frac{\partial \widetilde{Q}(x)}{\partial \epsilon} + \frac{\partial W(x, \widetilde{Q}(x), \widetilde{P}(x))}{\partial p} \frac{\partial \widetilde{P}(x)}{\partial \epsilon} dF(x) \triangleq \kappa.$$

We now prove that the right-hand side of the inequality, denoted by κ, is positive for ϵ small enough. We have $\widetilde{\theta} \to \underline{\theta}$ as $\epsilon \to 0$ so we have that:

$$\kappa \mid_{\epsilon=0} = 0.$$

Hence, the first order vanishes. We now compute the second order of the term:

$$\frac{d\kappa}{d\epsilon} \mid_{\epsilon=0} = f(\underline{\theta}) \frac{\partial \widetilde{\theta}}{\partial \epsilon} \left(\frac{\partial W(\underline{\theta}, Q^{**}(\underline{\theta}), P^{**}(\underline{\theta}))}{\partial q} + \frac{\partial W(\underline{\theta}, Q^{**}(\underline{\theta}), P^{**}(\underline{\theta}))}{\partial p} \frac{\partial \widetilde{P}(x)}{\partial \epsilon} \right) > 0.$$

The inequality follows from the fact that:

$$\frac{\partial W(\underline{\theta}, Q^{**}(\underline{\theta}), P^{**}(\underline{\theta}))}{\partial q} = \frac{\partial W(\underline{\theta}, Q^{*}(\underline{\theta}), P^{**}(\underline{\theta}))}{\partial q} = 0;$$
$$\frac{\partial W(\underline{\theta}, Q^{**}(\underline{\theta}), P^{**}(\underline{\theta}))}{\partial p} \frac{\partial \widetilde{P}(\underline{\theta})}{\partial \epsilon} > 0.$$

The equality comes from the fact that by definition Q^* satisfies the first-order condition; the inequality comes from the fact that increasing ϵ decreases the flagging probability (that is, $\partial \widetilde{P}(\underline{\theta})/\partial \epsilon < 0$). Hence, we prove the result.

Finally, we provide sufficient conditions on $\widehat{s}$ such that $\widehat{S}(r, I)$ is concave and supermodular. For this, we further assume that $\widehat{s}(\phi, I)$ is weakly concave and weakly supermodular in (ϕ, I), but not "excessively so". To formalize this last restriction, define the extreme values of its second derivatives:

(A12)
$$d_{II} \triangleq \min_{\phi, I} \left| \widehat{s}_{II}(\phi, I) \right|, \qquad d_{I\phi} \triangleq \max_{\phi, I} \left| \widehat{s}_{I\phi}(\phi, I) \right|, \qquad d_{\phi\phi} \triangleq \min_{\phi, I} \left| \widehat{s}_{\phi\phi}(\phi, I) \right|.$$

We impose the following dominance condition:

(A13)
$$\frac{d_{I\phi}}{2} \leq \min \left\{ d_{II}, \frac{d_{\phi\phi}}{3} \right\}.$$

This condition ensures that the degree of complementarity between patient health and investment does not dominate curvature in either dimension.

LEMMA A3 (Properties of Average Survival Rate):

If the survival rate $\widehat{s}$ satisfies (A12)-(A13)*, then the average survival function $\widehat{S}(r, I)$ is concave and supermodular.*

PROOF OF LEMMA A3:

Given the definition (7), we have that:

$$\begin{aligned}
\widehat{S}_I =& \frac{1}{1-r}\int_r^1 \widehat{s}_I(\phi, I)\, d\phi,\\
\widehat{S}_r =& \frac{1}{(1-r)^2}\int_r^1 (\widehat{s}(\phi, I) - \widehat{s}(r, I))\, d\phi.\\
\widehat{S}_{II} =& \frac{1}{1-r}\int_r^1 \widehat{s}_{II}(\phi, I)\, d\phi,\\
\widehat{S}_{rr} =& \frac{2}{(1-r)^3}\int_r^1 (\widehat{s}_\phi(\phi, I) - \widehat{s}_\phi(r, I))(1-\phi)\, d\phi.\\
\widehat{S}_{rI} =& \frac{1}{(1-r)^2}\int_r^1 \widehat{s}_{\phi I}(\phi, I)\,(1-\phi)\, d\phi,
\end{aligned}$$

We thus have that the first two second derivatives are negative and the third is positive, which follows from the fact that $\widehat{s}(\phi, I)$ is supermodular and concave. We thus have that $\widehat{S}$ is supermodular. To prove its concavity, we need to prove that:

$$\widehat{S}_{II}\widehat{S}_{rr} - \widehat{S}_{rI}^2 \geq 0.$$

To prove this, we prove that condition (A13) implies that

$$|\widehat{S}_{rI}(r, I)| \leq \min\left\{|\widehat{S}_{rr}(r, I)|,\, |\widehat{S}_{II}(r, I)|\right\} \tag{A14}$$

for all (r, I). We note that:

$$|\widehat{S}_{II}| \geq d_{II}$$
$$|\widehat{S}_{rI}| \leq \frac{d_{I\phi}}{2}$$

$$|\widehat{S}_{rr}| \geq \frac{2d_{\phi\phi}}{(1-r)^3}\int_r^1 (\phi - r)(1-\phi)\, d\phi = \frac{d_{\phi\phi}}{3}.$$

We thus have that, if (A13) is satisfied, then (A14) is also satisfied. This proves the result.