2020

Inducing Compliance with Post-Market Studies for Drugs under FDA's Accelerated Approval Pathway

Liang Xu

Hui Zhao

Nicholas C. Petruzzi

Inducing Compliance with Post-Market Studies for Drugs under FDA's Accelerated Approval Pathway

Liang Xu
College of Business, University of Nebraska - Lincoln, liang.xu@unl.edu
Hui Zhao
Smeal College of Business, Pennsylvania State University, huz10@psu.edu
Nicholas C. Petruzzi
Smeal College of Business, Pennsylvania State University, ncp12@psu.edu

March 2019

Problem definition: In 1992, FDA instituted the accelerated approval pathway (AP) to allow promising drugs to enter the market based on limited evidence of efficacy, thereby permitting manufacturers to verify true clinical benefits through post-market studies. However, most post-market studies have not been completed as promised. We address this non-compliance problem.
Academic/Practical Relevance: The prevalence of this non-compliance problem poses considerable public health risk, thus compromising the original purpose of a well-intentioned AP initiative. We provide an internally consistent and implementable solution to the problem through a comprehensive analysis of the myriad complicating factors and tradeoffs facing FDA.
Methodology: We adopt a Stackelberg framework in which the regulator, which cannot observe the manufacturer's private cost information or level of effort, leads by imposing a post-market study deadline. The profit-maximizing manufacturer then follows by establishing its level of effort to invest in its post-market study. In establishing its deadline, the regulator optimizes the tradeoff between providing public access to potentially effective drugs and mitigating public health risks from ineffective drugs.
Results: We develop a deadline-dependent user fee menu as a screening mechanism that establishes an incentive for manufacturer compliance. We show that its effectiveness in inducing compliance depends fundamentally on the enforceability of sanction, a drug-specific measure that indicates how difficult it is to withdraw an unproven drug from the market, and the drug's success probability: The higher is either, the higher is the probability that the mechanism induces compliance.
Managerial Implications: We synthesize and distill the salient tradeoffs and nuances facing FDA's non-compliance problem and provide an implementable solution. We quantify the value of the solution as a function of a drug's success probability and enforceability. From public policy perspective, we provide guidance for FDA to increase the viability and effectiveness of AP.

Keywords: Public Policy, Compliance, Pharmaceutical Industry, Asymmetric Information, Moral Hazard

---

Accepted for publication as: Xu Liang, Zhao Hui, Petruzzi C. Nicholas. 2020. Inducing Compliance with Post-market Studies for Drugs under FDA's Accelerated Approval Pathway. Manufacturing & Service Operations Management, Articles in Advance.

1. Introduction

The typical drug development process is extremely costly and risky for drug developers (e.g., manufacturers). From 1960-2010, the average investment per new drug approved by the Federal Drug Administration (FDA) doubled every 7.5 years, reaching $2.6 billion per new drug (DiMasi et al., 2016). At the same time, only one in 5000 of the compounds tested in laboratory will ever be approved by FDA (FDA, 2015). Hence, the process of translating basic research into a viable treatment is often described as the "valley of death" (Butler, 2008). Meanwhile, the long durations of drug development further exacerbate the cost and risk involved. A typical drug goes through three phases of clinical trials before approval, taking 10-12 years on average to complete (Paul et al. 2010). This leaves the manufacturer only 8-10 years of patent protection in the market to recoup its development costs and earn a profit. Indeed, as the Cutting Edge Information (2004) reports, each additional day a drug spends in clinical trials costs manufacturers from $600,000 for niche drugs to $8 million for blockbuster drugs. Time is the critical determinant of the economic fate of a new drug. As can be seen, these huge investments, high risks, and long durations also result in negative public health implications. First, they discourage drug manufacturers from innovation. As a result, manufacturers tend to be very selective in their development efforts. The sobering reality is that, of the approximately 4,000 diseases for which molecular mechanisms are understood, only about 250 (or 6.25%) have available treatments (Collins, 2012). Second, they cause serious delays in patient access to promising treatments, especially for time-sensitive and life-threatening diseases such as cancer and HIV/AIDS.

The above no-win situation compels the regulator to find alternatives to expedite access to new drugs. Accelerated-approval pathway (AP), instituted by FDA in 1992, is one such regulatory alternative. Specifically, AP allows drugs targeted at serious diseases and unmet medical needs to enter the market with "accelerated approval" based on surrogate health improvement measures (such as tumor-shrinkage rates) instead of the true clinical benefits (such as survival rates) that are required for regular approval, given the former have proven to be reasonable predictors of the latter (Guide for Industry, 2014). Compared to the true clinical benefits, which typically take years of extensive study to determine accurately, surrogate measures can be observed more promptly and require smaller sample sizes to detect. Indeed, Moore and Furberg (2014) find that drugs approved through AP in 2008 spent on average only 5.1 years in pre-market clinical trials instead of the 10-12 years typically needed for regular-approval pathway (RP).

While allowing early access to new drugs, AP inevitably introduces new public health risks because surrogate measures, by definition, serve only as imperfect predictors of the intended clinical benefits (Fleming, 2005). Therefore, as a condition for AP, FDA requires a manufacturer to conduct a post-market study to verify whether or not its drug, indeed, provides the anticipated clinical benefits. If it does, then the accelerated approval is converted to regular approval; otherwise, the drug needs to be withdrawn from the market. By law, FDA is obligated to ensure the completion of post-market studies so that drugs entering the market through AP ultimately meet the same statutory standards for clinical benefits as those approved through RP (Guide for Industry, 2014). Thus, in principle, if a manufacturer does not complete its post-market study with due diligence, FDA must initiate a sanction process to withdraw the drug from the market.

Unfortunately, most post-market studies are not completed as required for drugs entering the market through AP (GAO, 2009). Take ProAmatine as one example. In 1996, ProAmatine entered the market through AP to treat low blood pressure. However, two decades later, the required post-market study has yet to be completed, and this remains true even after FDA offered the manufacturer a three-year patent extension as an added incentive to complete the study. This example is hardly an isolated case: Upon examining 90 drugs on the market under AP between 1992 and 2008, the Government Accountability Office (GAO) found that 36% of the required post-market studies remained uncompleted as of 2009 and 50% of the uncompleted studies took on average 5.5 years even to begin (GAO, 2009).

There are several important reasons for this non-compliance problem. First, manufacturers may have little incentive to conduct the post-market study after a drug is on the market, especially if the prospects of a successful outcome are uncertain and there is no significant improvement on profitability. Second, post-market studies incur high costs. According to Johnson et al. (2011), patient recruitment rates typically drop, sometimes by a dramatic degree, for post-market studies because patients have already gained access to the treatment. In fact, physicians often prefer prescribing these drugs directly to their patients rather than referring patients to a post-market study where they would risk receiving a placebo and would be subject to additional testing and reporting procedures. Hence, manufacturers typically must invest considerable effort and money to recruit patients for post-market studies. Third, the threat of having its drug withdrawn from the market for non-compliance does not necessarily compel a manufacturer to act due to the following extenuating circumstances faced by FDA:

- To initiate a drug withdrawal process, FDA is burdened with substantiating its position that the manufacturer failed to conduct a post-market study with due diligence. This burden is non-trivial because FDA typically cannot verify a manufacturer's efforts on conducting a post-market study (an issue of moral hazard), nor does FDA have knowledge of a manufacturer's costs of conducting such study (an issue of asymmetric information). Hence, to prove any lack of due diligence on the manufacturer's part, FDA is left to rely on observable outcomes such as whether or not the manufacturer meets a specified deadline. However, given the asymmetric information and moral hazard issues, coupled with the inherent tradeoff between providing patients access and safeguarding against the risk of ineffective drugs, setting such a deadline in the first place is a most challenging task. Indeed, as we show later, too long or too short a deadline would actually induce non-compliance. Yet, without such a well-considered deadline, "FDA could not determine when a manufacturer is taking too long" to complete its study (GAO, 2009).

- Even if FDA were able to establish a claim of non-compliance through the documented lapse of an appropriately set deadline, withdrawing a drug from the market may take an uncertain amount of time due to the strong resistance from patients and physicians (e.g., Mayo Clinic, 2010). Many times, FDA must follow complex legal procedures to complete the withdrawal process. As a result, the withdrawal of a drug under AP without definite evidence of ineffectiveness typically cannot be enforced immediately; rather, it takes an uncertain amount of time that is largely beyond FDA's control. We refer to this exogenous nature of the withdrawal process as *enforceability*. A drug with higher enforceability can be withdrawn, on average, in a relatively shorter period of time.

With this as our backdrop, the goal of our paper is to investigate the regulator's problem of manufacturers' non-compliance with post-market studies for drugs under AP by first capturing the various salient tradeoffs and complications described above and then deriving a potentially implementable solution. To do this, we develop a Stackelberg game framework in which the regulator (FDA), whose objective is to maximize patient welfare, first establishes, for a drug under AP, a deadline by which the manufacturer must complete its required post-market study to avoid potential withdrawal of its drug from the market. The profit-maximizing manufacturer then responds by establishing the level of effort to invest in its post-market study. In establishing its deadline, the regulator must weigh the tradeoff between increasing the public's access to potentially life-saving drugs against the associated risk of increasing the public's exposure to potentially ineffective drugs. And, it must weigh this delicate tradeoff while constrained by three inherent challenges that combine to fundamentally define its regulatory context, namely its enforceability challenge, asymmetric information challenge, and moral hazard challenge.

We address the regulator's information asymmetry challenge by designing a deadline-dependent user fee to serve as a screening mechanism that induces the manufacturer to implicitly reveal its private information. We choose this mechanism in particular because of its potential practical appeal. Under PDUFA (Prescription Drug User Fee Act), FDA already requires a manufacturer to pay a fixed fee to fund its new drug application review (FDA, 2017). By replacing this fixed fee with one tied to a post-market study deadline that is acceptable to a given manufacturer, we leverage an existing FDA mechanism to induce the manufacturer to invest an appropriate level of effort for a post-market study. Given this screening mechanism, we then address the regulator's associated moral hazard challenge by restricting the available menu of deadline-dependent user fees to include only those that would guarantee the manufacturer's compliance in conducting its post-market study. The effectiveness and value of this menu ultimately depend not only on the probability of a given drug's success (which serves as a reward), but also on the enforceability of an unproven drug's market withdrawal (which serves as a penalty), such that the higher is either, the more likely is the manufacturer to comply with its required post-market study. Hence, a higher enforceability can be thought of as the functional equivalent of a higher probability of a drug's success in terms of inducing compliance. Accordingly, our modeling framework also enables us to investigate (1) the impact of enforceability and success probability on the manufacturer, the regulator, and patients, (2) the welfare loss due to the manufacturer's private information, (3) the value-added if the regulator could verify the manufacturer's effort on its post-market study, and (4) the value of the deadline-dependent user fee menu compared to an optimal single deadline, which provides valuable policy implications.

To complement our analytical results, we also compile and analyze data from the ProAmatine case to study the effectiveness and value of our deadline-dependent user fee. We show, for example, if the probability of the drug's clinical success is 85% and the enforceability of the drug is such that market withdrawal of the drug takes on average 2 years, then the proposed mechanism can increase the manufacturer's likelihood of compliance from 34.7% to 65.8% and induce the manufacturer to increase its effort by 103.3% as compared to the baseline situation in which withdrawal is not enforceable. In addition, we show that the higher is a drug's enforceability, the comparatively more valuable is it for the regulator to obtain the manufacturer's private information to set an appropriate deadline and deter non-compliance through ex-post sanction, whereas the lower is a drug's enforceability, the comparatively more valuable would it be for the regulator to verify the manufacturer's effort on its post-market study and prevent non-compliance through ex-ante monitoring.

Our paper makes several important contributions. First, from a public policy perspective, it addresses a critical non-compliance problem facing FDA in particular and the pharmaceutical industry in general that otherwise could jeopardize the effectiveness of a well-intentioned AP initiative. It also provides guidance for the regulator in managing its resources and priorities to ensure compliance. Second, from a modeling perspective, (1) it synthesizes and distills the salient trade-offs and nuances facing FDA and provides a potentially implementable deadline-dependent user fee solution accordingly; and (2) it isolates the effectiveness and quantifies the value of the solution as a function of enforceability and success probability of a drug. Third, from a literature perspective, it integrates three disparate streams of literature, namely, optimal effort allocation with time-cost tradeoffs (operations management), imperfect sanctioning of regulatory policy (economics), and AP non-compliance problem (health policy), as will be detailed in our literature review.

The remainder of this paper is organized as follows. In Section 2, we position our paper within the related literature. In Section 3, we develop our modeling framework. In Sections 4 and 5, we respectively solve the manufacturer's optimal effort problem and the regulator's optimal mechanism design problem. In Section 6, we examine the regulator's welfare loss attributed to the manufacturer's private information and its potential welfare gain if it could verify the manufacturer's effort. In Section 7, we conduct a numerical study using data compiled from the ProAmatine case to obtain additional insights to complement our analytical results. In Section 8, we explore three extensions to our model framework. Section 9 concludes with policy implications. Proofs of technical results are in Appendix.

2. Relation to Literature

Our paper is related to literature in health policy, operations management and economics. In this section, we review and position our paper in relation to these three areas.

Literature in health policy has touched the problem of non-compliance under AP. Some studies discuss possible reasons for non-compliance *without* providing solutions. For example, Dagher et al. (2004) attribute the non-compliance problem primarily to the effort required to recruit patient subjects into post-market studies, and Johnson et al. (2011) acknowledge that the manufacturer's lack of due diligence is a serious concern. Some other studies propose possible solutions to the non-compliance problem, but without comprehensive analysis. Gellad and Kesselheim (2017), for example, suggest a pricing scheme under which manufacturers are required to charge a comparatively low price before its post-market study is completed. However, as the authors recognize, the scheme is hard to implement because FDA cannot regulate prices. Alternatively, Wood (2006) suggests granting extended patents as a reward to manufacturers that complete their post-market studies, while GAO (2009) and Willyard (2014) both suggest increasing fines and warnings for manufacturers that do not complete post-market studies. However, as highlighted by the representative case of ProAmatine, without a carefully calibrated design, neither of these schemes necessarily will affect manufacturer's behavior. In contrast, our paper develops an internally consistent solution through comprehensive analysis of different factors and nuances that affect the different parties' decisions.

Literature in operations management has investigated, in various contexts, firms' optimal effort for meeting deterministic deadlines with fixed and known penalties imposed immediately upon missing the deadlines. These contexts include lead-time reduction (e.g., Shabtay and Steiner, 2007) , on-time delivery (e.g., Dai et al., 2016) and new product introduction (e.g., Cohen et al., 1996).

In our context, we also study an optimal-effort decision, but with an endogenized deadline set by the regulator. Accordingly, our work differentiates from this literature in two aspects. First, in our context, the penalty imposed upon missing the deadline is neither deterministic nor immediate because of the imperfect enforceability of sanction. Second, in our context, the deadline-setting the regulator faces both an asymmetric information challenge as well as a moral hazard challenge, where asymmetric information exists because the regulator does not know the manufacturer's cost efficiency and moral hazard exists because the regulator cannot verify manufacturer's effort. Although asymmetric information and moral hazard challenges are widely addressed in operations and supply chain management, fewer papers have both problems (e.g., Chick et al., 2016; Crocker and Letizia, 2014). And such challenges have not been addressed in the context of setting and meeting deadlines.

Our paper is also related to the literature in economics on sanction/penalty of non-compliance. Literature in economics has investigated non-compliance among firms when regulatory sanction is imperfect. In particular, this literature has shown that a firm engaged in environmentally hazardous activities will invest less precautionary effort than what is socially optimal if the courts are not able to hold the firm accountable for the damage (e.g., Kolstad et al., 1990) or if the firm can declare bankruptcy (e.g., Shavell, 1984). In our context, sanction for non-compliance also is imperfect, but the reason is due to the enforceability of drug withdrawal. Specifically, the drug withdrawal process will be initiated if manufacturer misses the required deadline, but it is uncertain how long it takes for that process to reach its completion, or in the extreme case, whether it will reach completion at all. To our knowledge, no previous studies have addressed issues associated with an uncertain regulatory sanction.

Related literature also has explored regulation and compliance issues in broader contexts. For example, studies of regulation on air pollution (Gray and Deily, 1996), on oil spill prevention (Gawande and Bohara, 2005) and on replacement of hazardous substances (Kraft et al., 2013) generally conclude that greater enforcement leads to better compliance. While these studies focus on compliance issues associated with exogenously given regulatory criteria, we in contrast endogenize the question of how to establish the regulatory criteria in the first place because such a question is pertinent in our context. Toward that end, we must incorporate the regulator's explicit tradeoff between risk and access, which is crucial for FDA to assure balanced public health benefits, in determining the required deadline for completing a post-market study. Indeed, GAO (2009) indicates that considerations of access to treatment give FDA pause when drug withdrawal becomes imminent. In the same spirit, Olson (2004) indicates that FDA may be willing to risk approving drugs with potentially adverse effects if the drugs are innovative in meeting patients' medical needs. Yet, no literature, to our knowledge, explicitly addresses this tradeoff between public health risk and access to treatment on the one hand, and the complications of compliance and enforceability on the other hand. Our study thus considers the manufacturer's compliance effort from the operations point of view, with enforceability of sanction and determination of regulatory criteria together, to provide a comprehensive view on the non-compliance problem that accompanies AP.

3. Model Setup

Consider a drug that enters the market under AP. Figure 1 summarizes the sequence of events. At time 0, the drug enters the market through AP with the manufacturer's promise to complete its required post-market study by a deadline, d, set by the regulator[1]. The manufacturer then chooses its effort, λ, to invest in completing its study. Then, if the manufacturer does not complete its study by d, the regulator imposes a sanction by initiating a process to withdraw the manufacturer's drug from the market. However, this process takes τ time to complete, where τ is a random variable that depends on the drug-specific enforceability of the withdrawal process (we will explain this notion of enforceability in detail later). Hence, the drug remains on the market until the completion of the withdrawal process at time $d+\tau$. If the manufacturer fails to complete the study by that time, then the drug exits the market at time $d+\tau$. If the manufacturer completes the study at any time $t \leq d+\tau$, then the results of the study are used to establish whether or not the drug, indeed, provides the intended clinical benefits. Due to the prohibitive legal consequences associated with the non-disclosure of clinical trial results, we assume the manufacturer truthfully and promptly reports the results of its post-market study upon the completion of its study. Therefore, at time t, the post-market study results either prove or disprove the drug's effectiveness in providing the true clinical benefits that it was designed to provide. Accordingly, with probability α (referred to as the success probability), the post-market study results confirm the drug's clinical benefits, in which case the regulator converts the AP approval to regular approval at time t and the drug continues to sell on the market; and with probability $1-\alpha$, the results do not confirm the intended clinical benefits, in which case the drug exits the market immediately at time t (withdrawal is immediate if the post-market study indicates the lack of clinical benefits because such a result suffices to provide definite evidence of ineffectiveness).

It is important to note that drugs typically face different demand rates and potentially different prices once converted to regular approval. This is because the regulator requires a label for a drug under AP to reflect the fact that the drug's clinical benefits remain uncertain. For example, labels must include statements such as "An improvement in clinical benefit has not been established. Continued approval for this indication may be contingent upon confirmatory trials" (Guide for

[1]As a matter of current practice, FDA has been trying to facilitate the completion of post-market studies by setting timelines for manufacturers, however, to date, the timelines have been non-binding.

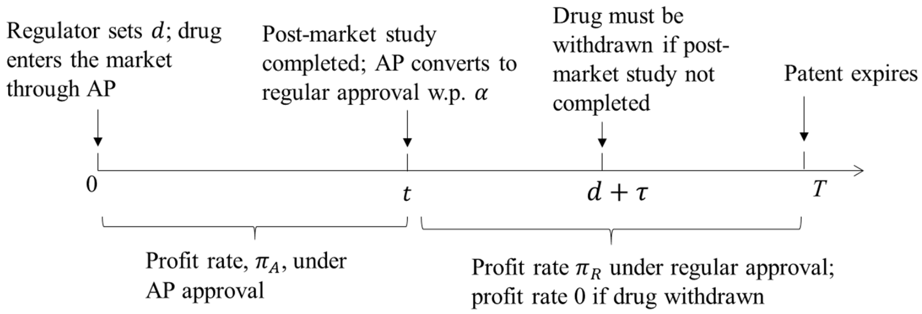

Figure 1: Sequence of events

Industry, 2014). As a result, patients have a higher reservation price for a drug under regular approval (Gellad and Kesselheim, 2017). In addition, the market size for the drug under regular approval also typically goes up because more patients and physicians will be informed about the drug. Therefore, the manufacturer typically earns a higher profit under regular approval. To capture this difference in profit, we assume a linear demand function $q_i(p) = a_i - b_i p$, where $i = A, R$ indicates AP approval and regular approval, respectively[2], $a_R \geq a_A$ represents a higher market size under regular approval, and $b_R \leq b_A$ represents a higher reservation price under regular approval. We assume the marginal production cost of the drug is negligible compared with the drug's price. Hence, the manufacturer's profit rate can be written as $\pi_i = \max_p q_i(p)p$. Let $p_i^* = \arg\max_p q_i(p)p$ and $q_i^* = q_i(p_i^*)$. Then, patient surplus from purchasing the drug is given by $\mu_i = \int_0^{q_i^*}(p_i(q) - p_i^*)dq$.

Define T as the time when the drug's patent expires. We designate T as the planning horizon because, after patent expiration, sales of brand drugs drop dramatically due to the entry of generics. From the regulator's perspective, the post-market study needs to be completed before patent expiration because the entry of generics would open a bigger market and, thus, introduces risk to more patients.

We now explain the withdrawal process. The withdrawal of a drug under AP is a process that takes an uncertain amount of time depending on the enforceability of the drug's withdrawal. To model this, we assume the process of withdrawing a drug with an overdue post-market study takes $\tau(s)$ time, where $\tau(s)$ follows an exponential distribution with parameter s denoting the enforceability: a higher s indicates higher enforceability and thus a shorter time to withdraw the drug on average. Hence, the expected time to carry the withdrawal process through to completion takes $1/s$ time. At one extreme, $s = \infty$ means the regulator is able to complete the withdrawal process immediately, i.e., $\tau(\infty) = 0$. We refer to this as the immediate-sanction case. At the other extreme, $s = 0$ means the regulator is not able to complete the withdrawal process at all, i.e., $\tau(0) = \infty$. We refer to this as the nonenforceable-sanction case. We assume s is exogenously

[2]Although we stipulate a linear demand function as our representative case, other tractable downward sloping demand functions (e.g., $q(p) = ae^{-bp}$) can be adopted similarly.

given and drug-specific due to the fact that withdrawing an AP drug without definite evidence of ineffectiveness requires a complex legal procedure (starting with, for example, assembling open hearings), which involves many factors beyond the regulator's control and for which the resistance from patients and physicians typically vary from drug to drug. From a practical standpoint, the regulator could arrive at such an estimate of s, for example, from its past experience trying to withdraw drugs with similar usage characteristics.

Given the deadline d, the manufacturer chooses effort λ for its post-market study to maximize its profit. Recall that one critical feature of a post-market study is the increased difficulty in recruiting patients as compared to the difficulty of recruiting patients for a pre-market clinical study (Johnson et al., 2011). As a result, the manufacturer must invest costly effort to complete the post-market study in time. Such effort may include opening multiple clinical sites in different locations to make the study more accessible to patients, providing more education and advertisement to physicians and patients to encourage scientific contribution, offering more funding to principal investigators, and providing better patient care and additional reimbursement to encourage patient participation. Thus, with increased effort, the manufacturer can increase the patient recruitment rate. However, such effort does not necessarily guarantee the completion of the study in time because the completion time also depends on stochastic patient recruitment, which we model as a Possion process. Since the time to test a drug on a patient subject is usually fixed and negligible compared to the time it takes to recruit patients, we approximate the time to complete a post-market study by the time it takes to recruit patients. Accordingly, let n be the required sample size for deriving statistically meaningful results from a given post-market study based on exogenously determined medical requirements. Then, the time to complete the study, t, which is equivalent to the time to recruit n patients, follows a gamma distribution with a probability density function (pdf) $g(t, \lambda, n) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}$ and a cumulative distribution function (cdf) $G(t, \lambda, n)$. Given $g(t, \lambda, n)$, then, the recruitment rate for the study, λ, can be interpreted as the manufacturer's effort. Accordingly, we define $\lambda = 0$ to mean non-compliance and $\lambda > 0$ to mean compliance. As mentioned, the regulator cannot observe λ because the regulator typically cannot verify the manufacturer's progress on a post-market study unless a detailed monitoring system is first designed and put in place.

Since patient recruiting essentially drives how fast a post-market study can be accomplished, we focus on the cost of effort in recruiting patients and do not consider fixed costs associated with the study (e.g., medications costs). Specifically, we assume the manufacturer incurs cost $C(\lambda, \theta) = \theta\lambda^k$ for investing effort λ, where $k \geq 1$ indicates that the cost is convex in effort invested and θ indicates how efficient the manufacturer is in expending a unit of effort on the given post-market study. We refer to θ as the manufacturer's cost type such that a lower θ denotes a higher cost efficiency of the manufacturer. The regulator does not know the manufacturer's cost type. The manufacturer, in contrast, knows its cost type because of its prior experience and knowledge. In some instances, the manufacturer may even influence a physician's choice between referring patients to post-market study and prescribing the drug already on the market under AP. To capture this information asymmetry, we assume the manufacturer knows θ but the regulator is limited to the belief that θ follows a distribution characterized by pdf $\psi(\theta)$ and cdf $\Psi(\theta)$ for $\theta \in [\underline{\theta}, \bar{\theta}]$.

Given that the regulator must set deadline d without knowing the manufacturer's cost type θ, we develop an incentive compatible, direct revelation mechanism, which, according to the revelation principle (Fudenberg and Tirole, 1991), optimizes the regulator's objective. Specifically, we develop a contract menu consisting of an upfront payment $F(\theta)$ and a corresponding deadline $d(\theta)$ such that the manufacturer implicitly reveals its cost type by choosing the pair that maximizes its profit. As mentioned, PDUFA enables FDA to charge the manufacturer a user fee to fund the review of a new drug application. Hence, by converting FDA's current fixed fee to a deadline-dependent fee, we essentially provide the regulator the ability to tailor its post-market study deadline in accordance with the manufacturer's efficiency in conducting post-market studies. As a result, if a manufacturer has a lower cost type (higher efficiency) in conducting its post-market study, it can commit to a shorter deadline in exchange for a lower user fee; otherwise, it can pay a higher user fee for a longer deadline. Note that it is still possible that the manufacturer accepts a shorter d (and hence a lower user fee) without actually investing the corresponding effort to meet it because the regulator typically cannot verify the manufacturer's effort. Therefore, we must design $(d(\theta), F(\theta))$ to induce the manufacturer not only to reveal its true cost type, but also to comply with the requirement to complete its study.

We are now ready to solve the problem by backward induction. The next section will present the analysis of the manufacturer's optimal effort given a regulatory menu and Section 5 will be devoted to designing the optimal menu.

4. The Manufacturer's Effort on Post-market Study

In this section, we first solve the manufacturer's problem of choosing the optimal effort for its post-market study given regulatory requirement $d(\theta)$ and $F(\theta)$, and we then study how the manufacturer's cost type (θ), enforceability (s), and success probability (α) impact its compliance.

Let $\tilde{d} = \min(d + \tau, T)$ denote the time when the drug ceases to sell on the market due to either the completion of the regulatory withdrawal process or the expiration of the drug's patent. If the manufacturer completes its post-market study before $\tilde{d}$, i.e., if $t \leq \tilde{d}$, the drug sells under AP for t time and then the drug either sells under regular approval for $T - t$ time (which corresponds to the case in which the results of the post-market study confirm the clinical benefits) or exits the market (which corresponds to the case in which the results of the study do not confirm the clinical benefits). If the manufacturer does not complete its post-market study before $\tilde{d}$, i.e., if $t > \tilde{d}$, the drug sells under AP until $\tilde{d}$ and then exits the market at $\tilde{d}$. Accordingly, for a given realization of t and τ, the manufacturer's total revenue is,

$$\phi(t,\tau) = \begin{cases} t\pi_A + (T-t)\alpha\pi_R, & \text{if } t \leq \tilde{d}, \\ \tilde{d}\pi_A, & \text{if } t > \tilde{d}. \end{cases}$$

Since the manufacturer knows its cost type θ, we suppress θ from $d(\theta)$ and $F(\theta)$ in this section. The manufacturer's problem, for a given regulatory requirement (d, F), is to determine its effort λ to maximize its expected profit, $\Pi(\lambda, d, \theta)$, where

$$\Pi(\lambda, d, \theta) \quad = \quad \int_0^\infty E_\tau(\phi(t,\tau))g(t,\lambda,n)dt - C(\lambda,\theta).$$

Note that we do not include the user fee F in $\Pi(\lambda, d, \theta)$ because it is a fixed upfront payment, hence it has no impact on the manufacturer's choice of λ.

The manufacturer has two incentives to comply: the market incentive of potentially converting to a higher profitability under regular approval after completing its post-market study and the regulatory incentive of trying to avoid the withdrawal of the drug from the market due to an overdue post market study. Thus, in effect, the manufacturer's problem boils down to a time-cost tradeoff: On the one hand, investing effort increases the manufacturer's likelihood of completing its post-market study early enough to avoid sanction and potentially increase its profitability; on the other hand, investing effort increases its cost. Given this tradeoff, Lemma 1 characterizes the manufacturer's optimal effort $\lambda^*(d)$ as a function of its deadline.

Lemma 1 *Given deadline d, the manufacturer's optimal effort is $\lambda^*(d) = \arg\max_{\lambda=\{0,\bar{\lambda}\}} \Pi(\lambda, d, \theta)$, where $\bar{\lambda} := \bar{\lambda}(d, \theta)$ satisfies the first-order-condition $\frac{\partial \Pi(\lambda,d,\theta)}{\partial \lambda} = 0$. If $\alpha\pi_R \geq \pi_A$, $\bar{\lambda}$ is the larger of two positive roots.*

Lemma 1 shows that given a deadline d, the manufacturer's optimal effort is either 0 or $\bar{\lambda}$, which implies that the manufacturer does not necessarily comply. This result follows because the manufacturer's marginal benefit of effort first increases and then decreases as a function of λ. Specifically, when λ is small, the manufacturer has a low chance of completing the post-market study by d. Hence, the manufacturer benefits from increased effort λ: its chance of completing post-market study by d increases in λ. As λ continues to increase, however, the manufacturer's chance of completing its post-market study by d starts to increase at a decreasing rate. As a result, the benefits of additional effort are limited. Therefore, when λ is small or λ is sufficiently large, the marginal cost of effort may outweigh its marginal benefit. The former results in non-compliance ($\lambda = 0$) as one local maximum of $\Pi(\lambda, d, \theta)$, and the latter results in $\bar{\lambda}$ as another local maximum.

Accordingly, let $\Pi_C(d,\theta) = \Pi(\bar{\lambda}, d, \theta)$ denote the manufacturer's profit if it does comply and let $\Pi_N(d) = \Pi(0, d, \theta)$ denote the manufacturer's profit if it does not comply. Then, the manufacturer will comply if and only if $\Pi_C(d,\theta) \geq \Pi_N(d)$. Given Lemma 1, we next examine under what condition the manufacturer would comply.

Proposition 1 *Given enforceability s and success probability α, there exists a cost type threshold $\hat{\theta}$ such that:*

(a) If $\theta > \hat{\theta}$, then $\lambda^*(d) = 0$ for all d.

(b) If $\theta \leq \hat{\theta}$, then there exists $d_1(\theta) \leq d_2(\theta)$ such that $\lambda^*(d) > 0$ if and only if $d \in [d_1(\theta), d_2(\theta)]$, where $d_1(\theta)$ and $d_2(\theta)$ correspond to the two solutions to $\Pi_C(d,\theta) = \Pi_N(d)$. Moreover, $d_1(\theta)$ increases with θ while $d_2(\theta)$ decreases with θ.

Proposition 1 reveals that, to induce the manufacturer to conduct its post-market study, the regulator must set a deadline that falls within an interval that depends on the manufacturer's private cost type. Intuitively, if the regulator undershoots and sets the deadline too short, the manufacturer would not comply because its cost to complete the study in such a short time would be prohibitively high; if the regulator overshoots and sets the deadline too long, the manufacturer would not comply because the benefit of converting to regular approval is limited by the imminent expiration of its patent. (Drug sales after patent expiration essentially drop to zero, hence the invested effort can no longer be capitalized once the patent expires.) In this spirit, not only would $d_2(\theta)$ answer FDA's practical question of "how long is too long?" when it comes to setting a deadline (GAO, 2009), but also would $d_1(\theta)$ answer the unasked analogous question of "how short is too short?" Further, note that the deadline zone that establishes the manufacturer's compliance region shrinks for decreased cost efficiencies, to the extent that the zone disappears altogether if the manufacturer's cost type exceeds the threshold $\hat{\theta}$. Thus, if the manufacturer's cost type is sufficiently high, namely if $\theta > \hat{\theta}$, the manufacturer will not comply with any deadline because conducting its post-market study simply becomes cost prohibitive. Figure 2 illustrates how the compliance region changes as a function of the manufacturer's cost type θ, where the lower bound of the shaded region represents $d_1(\theta)$ and the upper bound represents $d_2(\theta)$.

We next examine how drug enforceability, s, and post-market study success probability, α, impact the manufacturer's compliance behavior.

Proposition 2 *The cost type threshold $\hat{\theta}(s,\alpha)$ increases with enforceability s and it increases with success probability α.*

Proposition 2 indicates that, with either a higher enforceability, a higher success probability, or both, the regulator could induce compliance from a less efficient manufacturer that otherwise would

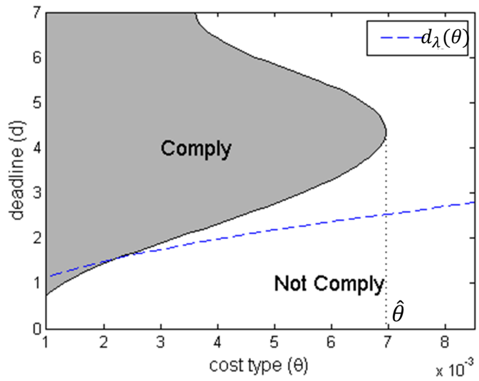

Figure 2: Compliance region ($s = 1, \alpha = 0.85, \pi_A = 19.8, \pi_R = 37, n = 405, T = 7, k = 2$)

not comply under a lower enforceability or success probability. This is true because, recall, a manufacturer will not comply regardless of what deadline the regulator sets if its cost type is above $\hat{\theta}(s, \alpha)$. Hence, for any given success probability, with everything else equal, the probability of compliance increases with a drug's enforceability to the extent that if enforceability exceeds the threshold $\hat{s} \equiv \hat{\theta}^{-1}(\bar{\theta}; \alpha)$, then compliance can be guaranteed by imposing an appropriate deadline. And by the same token, compliance can likewise be guaranteed with an appropriate deadline if, for a given drug's enforceability, with everything else equal, the success probability exceeds the threshold $\hat{\alpha} \equiv \hat{\theta}^{-1}(\bar{\theta}; s)$. Therefore, in essence, higher enforceability can be thought of as the functional equivalent of a higher success probability.

While Propositions 1 and 2 together establish the range of deadlines under which the manufacturer complies, Proposition 3 next indicates how the manufacturer's effort under compliance ($\bar{\lambda}$) is impacted by the deadline d.

Proposition 3 *Given cost type θ, there exists $d_\lambda(\theta)$ such that $\bar{\lambda}$ decreases with d if and only if $d \geq d_\lambda(\theta)$. And, $d_\lambda(\theta)$ increases with θ.*

Proposition 3 provides the condition under which the regulator can use a deadline as a lever to induce effort. When $d \geq d_\lambda(\theta)$, the regulator can induce the manufacturer to invest higher effort by imposing a shorter deadline d. However, similar to Proposition 1, if the required deadline already is too short, i.e., if $d < d_\lambda(\theta)$, then decreasing the deadline actually would discourage the manufacturer from investing effort because the manufacturer would have too low a chance of completing its study in time. Hence, Proposition 3 implies that $\bar{\lambda}$ reaches its maximum at $d_\lambda(\theta)$. Furthermore, Proposition 3 indicates that $d_\lambda(\theta)$ increases with θ, implying that the range of deadlines that the regulator could use to induce effort is more limited for a less efficient manufacturer. The dashed line in Figure 2 illustrates $d_\lambda(\theta)$, which, for the most part, is less than $d_1(\theta)$.

Taken together, Propositions 1 and 3 indicate, first, that the regulator must set a deadline $d \geq \max(d_1(\theta), d_\lambda(\theta))$ to induce the manufacturer to comply and, second, that the manufacturer's optimal effort $\bar{\lambda}$ increases as d decreases. Hence, if the regulator imposed the deadline $d(\theta) = \max(d_1(\theta), d_\lambda(\theta))$, then the manufacturer would invest the maximum effort it would be willing to invest in its post-market study. However, such a deadline may not be possible because the regulator does not know the manufacturer's private cost type. Moreover, such a deadline may not be desirable because the regulator needs to make a tradeoff between drug risk and drug access when determining its deadline. Thus, we next solve the regulator's optimal deadline problem, considering the asymmetric information and the tradeoff between risk and access.

5. The Regulator's Mechanism Design Solution

The regulator's objective is to maximize the expected patient welfare, which, recall, involves a tradeoff between access and risk: On the one hand, a drug under AP provides patients access to a potentially life-saving treatment sooner than what could have been possible under RP; on the other hand, drugs on the market without confirmed clinical benefits expose patients to negative health effects with probability $1 - \alpha$. Accordingly, for any d and given realizations of t and τ, patient welfare is

$$\nu(t, \tau; d) = \begin{cases} t\mu_A + (T - t)\alpha\mu_R - w(1 - \alpha)t & \text{if } t \leq \tilde{d}, \\ \tilde{d}\mu_A - w(1 - \alpha)\tilde{d} & \text{if } t > \tilde{d}, \end{cases}$$

where μ_A and μ_R indicate the patient surplus from purchasing the drug under AP and regular approval, respectively, as defined in Section 3, and w represents the regulator's weight on risk such that the higher is w, the higher is the priority the regulator puts on the potential negative effects that patients may experience from a drug without confirmed clinical benefits. Given the manufacturer's optimal effort response $\lambda^*(d)$, expected patient welfare can be written as

$$U(d) = \int_0^\infty E_\tau\big(\nu(t, \tau; d)\big) g(t, \lambda^*(d), n) dt$$

Therefore, the regulator's objective is to maximize $U(d)$.

Given that the regulator does not know the manufacturer's cost type, it essentially must solve a mechanism design problem in which the manufacturer is provided a deadline-dependent user fee menu to reveal the manufacturer's cost type and, thus, to maximize $U(d)$. Formally, to determine the optimal deadline-dependent user fee menu $(d(\theta), F(\theta))$, the regulator must solve the following problem:

$$\max_{d(\theta),F(\theta)} \int_{\underline{\theta}}^{\bar{\theta}} [U(d(\theta)) + F(\theta)]\psi(\theta)d\theta \tag{1}$$
$$\text{s.t. } \Pi_C(d(\theta),\theta) - F(\theta) \geq \Pi_C(d(\tilde{\theta}),\theta) - F(\tilde{\theta}), \forall\tilde{\theta} \in [\underline{\theta},\bar{\theta}]$$
$$\Pi_C(d(\theta),\theta) - F(\theta) \geq \Pi_0,$$
$$\Pi_C(d(\theta),\theta) \geq \Pi_N(d(\theta)), \forall\theta \leq \hat{\theta}.$$

The first constraint in Problem (1) is the standard incentive compatibility constraint, which ensures that the manufacturer chooses $(d(\theta), F(\theta))$ according to its true cost type. To satisfy this constraint, $d(\theta)$ and $F(\theta)$ should be non-decreasing in θ. The second constraint is the standard individual rationality constraint, which guarantees the manufacturer obtains at least its reservation profit Π_0. The third constraint, which is unique to our regulatory context, is the compliance constraint. This constraint requires the manufacturer to invest effort to complete its post-market study, thereby allowing the regulator to fulfill its legal obligation as a gatekeeper to confirm an approved drug's intended clinical benefits (Guide for Industry, 2014). Notably, however, as Proposition 1 establishes, the manufacturer with cost type $\theta > \hat{\theta}$ will not comply regardless of what the deadline is. Therefore, despite the compliance constraint, the regulator must bear the risk of non-compliance with probability $1 - \Psi(\hat{\theta})$.

Based on Fudenberg and Tirole (1991), the solution to Problem (1) must satisfy the well-known single-crossing requirement to guarantee incentive compatibility. Whereas many mechanism design studies in operations management stipulate that the single-crossing requirement is satisfied (e.g., Iyer et al., 2005), we explicitly consider the condition under which the requirement is satisfied in our context. In our context, the single-crossing requirement is satisfied when the manufacturer's optimal effort increases as d decreases, which, according to Proposition 3, is not necessarily guaranteed. Nevertheless, Proposition 3 establishes that the requisite condition is satisfied if $d \geq d_\lambda(\theta)$. This condition, together with the compliance constraint, thus restricts the regulator to choose deadline $d(\theta) \geq \max(d_1(\theta), d_\lambda(\theta))$. Accordingly, Proposition 4 characterizes the regulator's optimal deadline-dependent user fee menu, where the superscript "AI" represents asymmetric information.

Proposition 4 *(a) For $\theta \in [\underline{\theta}, \hat{\theta}]$, the regulator's optimal deadline-dependent user fee menu is characterized as follows:*

$$d^{AI}(\theta) = \min\left(d_2(\hat{\theta}), \max\left(d^*(\theta), d_1(\theta), d_\lambda(\theta)\right)\right)$$

$$F^{AI}(\theta) = \Pi_C(d^{AI}(\theta),\theta) - \Pi_0 - \int_{\theta}^{\hat{\theta}} \left(\bar{\lambda}(d^{AI}(\tilde{\theta}))\right)^k d\tilde{\theta},$$

where $d^*(\theta)$ solves $\frac{3}{2}\frac{\partial\Pi}{\partial d}\left(\bar{\lambda}, d, \theta\right) + \frac{C'}{2}\frac{\partial\bar{\lambda}}{\partial d} - w(1-\alpha)\frac{\partial E_{t,\tau}(\min(\tilde{d},t))}{\partial d} - k\bar{\lambda}^{k-1}\frac{\partial\bar{\lambda}}{\partial d}\frac{\Psi(\theta)}{\psi(\theta)} = 0.$

(b) For $\theta > \hat{\theta}$, the regulator's optimal deadline and corresponding user fee are as follows:

$$\begin{aligned} \bar{d}^{AI} &= \max\left(d^{AI}(\hat{\theta}), \frac{1}{s}\ln(1 - \frac{2w(1-\alpha)}{3\pi_A}) + T\right), \\ \bar{F}^{AI} &= \Pi_N(\bar{d}) - \Pi_0. \end{aligned}$$

Proposition 4 indicates that the regulator's optimal menu consists of two parts: $(d^{AI}(\theta), F^{AI}(\theta))$ for $\theta \leq \hat{\theta}$ and $(\bar{d}^{AI}, \bar{F}^{AI})$ for $\theta > \hat{\theta}$. We now provide some intuition on how the menu induces the manufacturer to respond. First, if the manufacturer's cost type is below the threshold $\hat{\theta}$ (i.e., if $\theta \leq \hat{\theta}$), then the regulator's menu will induce the manufacturer to choose a deadline and to comply accordingly. In the optimal menu, the regulator imposes a shorter deadline and charges a lower fee for a manufacturer with a lower cost type. The user fee equals the manufacturer's profit less the sum of its reservation profit Π_0 and its *information rent*, which essentially refers to the amount of user fee that the regulator must forgo to induce the manufacturer to reveal its cost type. This information rent $\int_{\theta}^{\hat{\theta}} \left(\bar{\lambda}(d^{AI}(\tilde{\theta}))\right)^k d\tilde{\theta})$ increases with a shorter deadline because the manufacturer has to commit to a higher effort. Notably, to ensure compliance, not only must $d^{AI}(\theta) \geq \max(d_1(\theta), d_\lambda(\theta))$ as discussed previously, but also must $d^{AI}(\theta) \leq d_2(\hat{\theta})$. Intuitively, this is true because, from Proposition 1, if $\theta \leq \hat{\theta}$, then $d^{AI}(\theta)$ must be no greater than $d_2(\theta)$ to induce compliance while $d^{AI}(\theta)$ also must be non-decreasing in θ to remain incentive compatible.

Second, if the manufacturer's cost type is above the threshold $\hat{\theta}$ (i.e., if $\theta > \hat{\theta}$), then the regulator's menu will not induce compliance because, as Proposition 1 establishes, no deadline would compel a manufacturer with such a high cost type to comply. Rather, the regulator's menu will induce such a manufacturer to choose $\bar{d}^{AI}$, which is the deadline that optimizes the regulator's risk versus access tradeoff given that the drug withdrawal process inevitably will be initiated. To be incentive compatible, the regulator's optimal deadline-dependent user fee $\bar{F}^{AI}$ in this situation must not only induce the manufacturer to choose $(\bar{d}^{AI}, \bar{F}^{AI})$ if its cost type is greater than $\hat{\theta}$, but also must prevent the manufacturer from choosing $(\bar{d}^{AI}, \bar{F}^{AI})$ if its cost type is less than $\hat{\theta}$. As a result, $(\bar{d}^{AI}, \bar{F}^{AI})$ is designed such that a non-complying manufacturer is indifferent between choosing $(\bar{d}^{AI}, \bar{F}^{AI})$ and $(d^{AI}(\hat{\theta}), F^{AI}(\hat{\theta}))$, while $(\bar{d}^{AI}, \bar{F}^{AI})$, if chosen, would optimize the regulator's tradeoff between risk and access.

Figure 3 illustrates the optimal deadline-dependent user fee menu specified in Proposition 4, where Figure 3(a) depicts the deadline $d(\theta)$, Figure 3(b) depicts the corresponding user fee $F(\theta)$, and Figure 3(c) provides the resulting deadline-dependent user fee menu (d, F) presented to the manufacturer. Figure 3 summarizes three key points we explained above. First, the regulator requires a shorter deadline for a lower cost manufacturer and charges a lower user fee to induce the manufacturer to reveal its true cost type. Second, the regulator can only set a deadline to ensure compliance for a manufacturer with $\theta \leq \hat{\theta}$, with a maximum deadline of $d^{AI}(\hat{\theta}) = d_1(\hat{\theta})$. For a

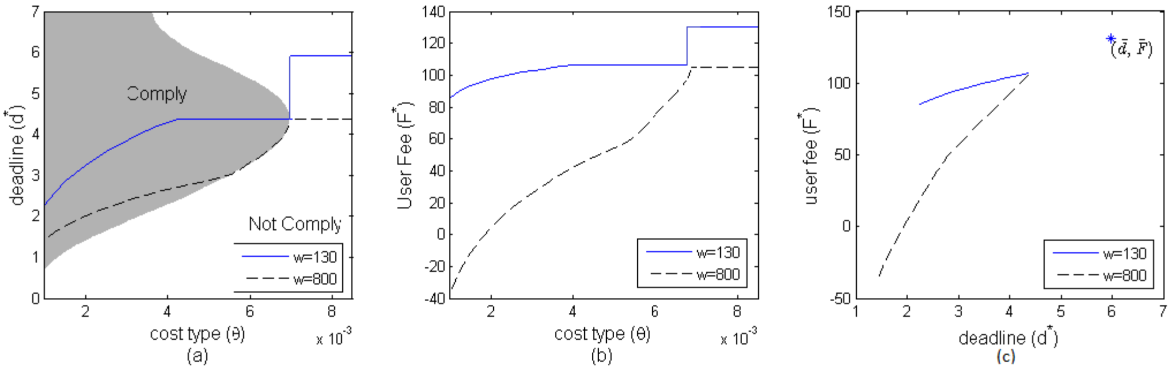

Figure 3: Optimal deadline and corresponding user fee ($s = 1, \alpha = 0.85, \pi_A = 19.8, \pi_R = 37, n = 405, T = 7, k = 2$)

manufacturer with $\theta > \hat{\theta}$, the regulator's separate pair of $(\bar{d}^{AI}, \bar{F}^{AI})$ optimizes its risk versus access tradeoff given that the manufacturer is destined not to comply. Third, everything else being equal, the higher is the regulator's weight on risk, w, the shorter should be the deadline imposed so as to induce a higher effort from the manufacturer, thereby reducing the length of time patients are exposed to potentially ineffective or harmful drugs. Given $\bar{d}^{AI} \geq d^{AI}(\hat{\theta})$ according to Proposition 4(b), we see for this example that when $w = 130$, $\bar{d} > d^{AI}(\hat{\theta})$ but when $w = 800$, $\bar{d} = d^{AI}(\hat{\theta})$. Note that, if w is especially high (e.g., if $w = 800$), then the user fee can be negative, which can be interpreted to mean that the regulator may even consider "paying" an extremely efficient manufacturer to complete its post-market study quickly, thereby expeditiously removing all doubt one way or the other regarding the drug's clinical benefits.

6. Loss from Information Asymmetry and Gain from Verifying Effort

Although the deadline-dependent user fee menu described in Proposition 4 effectively solves the regulator's problem, it also highlights the inefficiency in regulating the post-market study that arises from the regulator's inability to observe the manufacturer's cost type and effort invested. In particular, Proposition 4 highlights that the regulator bears a risk of non-compliance that cannot be eliminated by imposing a deadline if the manufacturer has a high cost type. Additionally, Proposition 4 highlights that the regulator essentially must pay an information rent when it can induce compliance, where, recall, such information rent increases with shorter deadlines. As a result, the regulator must impose a comparatively longer deadline, which results in a lower manufacturer effort than otherwise would be the case without the information asymmetry. Therefore, in this section, we study the following two questions: 1) What if the regulator had complete information with regard to the manufacturer's cost type? And 2) What if the regulator could verify the manufacturer's effort? The answers to these two questions provide insights into the regulator's welfare loss from asymmetric information and also the potential welfare gain if the regulator could verify the manufacturer's effort.

6.1 Loss from Asymmetric Information

In this subsection, we examine the regulator's welfare loss from not knowing the manufacturer's cost type. Toward this end, we first solve the regulator's optimal mechanism if the regulator had complete information on the manufacturer's cost type and then compare the solution to that of Problem (1).

Formally, if the regulator knew the manufacturer's cost type θ, for $\theta > \hat{\theta}$, the regulator may consider not even granting AP for the drug since the regulator would know ex-ante that the manufacturer will not comply (see Proposition 1). This will eliminate non-compliance in the first place. For $\theta \leq \hat{\theta}$, on the other hand, the regulator would choose a deadline $d(\theta)$ to solve the following problem:

$$\begin{aligned} &\max_{d} U(d) \\ &\text{s.t. } \Pi_C(d,\theta) \geq \Pi_0, \\ &\qquad \Pi_C(d,\theta) \geq \Pi_N(d). \end{aligned} \tag{2}$$

Analogous to Problem (1), the two constraints here ensure the manufacturer's participation and compliance, respectively. Unlike Problem (1), no incentive compatibility constraint is required here because, by definition, the regulator already knows θ. Let $d^{CI}(\theta)$ denote the optimal solution to this problem, where the superscript stands for complete information.

Proposition 5 *If the regulator knew θ, for $\theta \in [\underline{\theta}, \hat{\theta}]$, its optimal deadline would be $d^{CI}(\theta) = \min\left(d_2(\theta), \max(d^0(\theta), d_1(\theta))\right)$, where $d^0(\theta)$ solves $\frac{1}{2}\frac{\partial \Pi}{\partial d}(\bar{\lambda}, d, \theta) + \frac{C'}{2}\frac{\partial \bar{\lambda}}{\partial d} - w(1-\alpha)\frac{\partial E_{t,\tau}(\min(\tilde{d},t))}{\partial d} = 0$.*

Proposition 5 shows that, if the regulator knows θ, it can impose a deadline without having to account for any information rent, which is a nuance reflected by the implicit definition of $d^0(\theta)$. In addition, Proposition 5 further shows that while the regulator still must impose d within a certain interval lest the manufacturer would not comply, because $d^{CI}(\theta)$ is not required to be incentive compatible, $d^{CI}(\theta)$ can take any value in $[d_1(\theta), d_2(\theta)]$ rather than be bounded by $d_1(\hat{\theta})$ or $d_\lambda(\theta)$ as is the case in Proposition 4.

Corollary 1 next compares $d^0(\theta)$ from Proposition 5 to $d^*(\theta)$ from Proposition 4 to shed light on the impact of asymmetric information on the regulator's optimal deadline.

Corollary 1 *If $d^*(\theta) \geq d_\lambda(\theta)$, then $d^*(\theta) > d^0(\theta)$.*

According to Corollary 1, if $d^*(\theta) \geq d_\lambda(\theta)$, which from Proposition 3 would imply that the manufacturer's effort decreases with its deadline, then knowing the manufacturer's cost type would allow the regulator to impose a more restrictive deadline, which, correspondingly, would induce the manufacturer to increase its effort in compliance.

6.2 Gain from Verifying Effort

In this subsection, we examine the value-added to the regulator if it could somehow verify the manufacturer's effort. While the manufacturer's effort is typically non-verifiable, FDA has been taking initiatives towards verifying the manufacturer's effort in some circumstances (GAO, 2009). Take ProAmatine as an example. In 2012, the regulator and the manufacturer reached an agreement that specified a set of detailed milestones that the manufacturer needed to meet. In principle, with unambiguous milestones, the regulator could verify the manufacturer's effort by diligently monitoring the progress of the post-market study. However, such monitoring would require significant resources from the regulator. Therefore, exploring the potential gain from verifying the manufacturer's effort is critical for guiding the regulator in decisions regarding whether and how to deploy monitoring resources.

Toward that end, we next solve the regulator's optimal mechanism if the regulator does not know θ but could verify the manufacturer's effort. We then compare the solution to that of Problem (1). If the regulator could verify effort, then the regulator's problem would be to determine the effort to require from the manufacturer rather than to determine the deadline to impose on the manufacturer. Accordingly, we again focus on a direct revelation mechanism that allows the regulator to infer the manufacturer's cost type and determine the optimal effort requirement. Specifically, we introduce a menu of *effort*-dependent user fees, i.e., $(\lambda(\theta), F(\theta))$, where $\lambda(\theta)$ is the required effort and $F(\theta)$ is the upfront user fee. In this case, a higher effort requirement is coupled with a lower user fee such that a manufacturer with a lower cost type would prefer to commit to a higher effort in exchange for a smaller fee payment, while a manufacturer with a higher cost type would prefer the opposite.

Unlike in Problem (1), if the regulator can verify the manufacturer's effort, then it would not need to impose a deadline for completing the post-market study. Thus, we assume here that non-compliance is prevented through ex-ante monitoring and that the manufacturer will invest based on the effort required. Therefore, given an effort λ required by the regulator, the manufacturer's expected profit is $\Pi(\lambda, \theta) = \int_0^\infty \phi(t, \tau = \infty) g(t, \lambda, n) dt - C(\lambda, \theta)$. Correspondingly, the regulator's expected patient welfare is $U(\lambda) = \int_0^\infty \nu(t, \tau = \infty) g(t, \lambda, n) dt$. Accordingly, to determine the optimal effort-dependent user fee menu $(\lambda(\theta), F(\theta))$, the regulator solves the following problem:

$$\begin{aligned} \max_{\lambda(\theta),F(\theta)} & \int_{\underline{\theta}}^{\bar{\theta}} [U(\lambda(\theta)) + F(\theta)]\psi(\theta)d\theta \\ \text{s.t. } & \Pi(\lambda(\theta),\theta) - F(\theta) \geq \Pi(\lambda(\tilde{\theta}),\theta) - F(\tilde{\theta}), \forall \tilde{\theta} \in [\underline{\theta},\bar{\theta}] \\ & \Pi(\lambda(\theta),\theta) - F(\theta) \geq \Pi_0, \end{aligned} \tag{3}$$

where the first constraint is the incentive compatibility constraint that ensures the manufacturer with cost type θ chooses $(\lambda(\theta), F(\theta))$ and the second constraint is the individual rationality constraint. In Problem (3), since the regulator eliminates non-compliance by explicitly imposing and then verifying an effort requirement, the compliance constraint from Problem (1) is not needed.

Proposition 6 provides the optimal effort-dependent user fee menu, where superscript "VE" indicates verifiable effort.

Proposition 6 *For $\theta \in [\underline{\theta},\bar{\theta}]$, the regulator's optimal effort-dependent user fee menu is characterized as follows:*

$$F^{VE}(\theta) = \Pi(\lambda^{VE}(\theta),\theta) - \Pi_0 - \int_{\theta}^{\bar{\theta}} \left(\lambda^{VE}(\tilde{\theta})\right)^k d\tilde{\theta},$$

where $\lambda^{VE}(\theta)$ solves the following first-order condition: $\left(\frac{3}{2}(\alpha\pi_R - \pi_A) + w(1-\alpha)\right)G(T,\lambda,n+1)\frac{n}{\lambda^2} - k\lambda^{k-1}\left(\theta + \frac{\Psi(\theta)}{\psi(\theta)}\right) = 0.$

According to Proposition 6, any non-increasing $\lambda(\theta)$ coupled with a corresponding $F(\theta)$ would reveal the manufacturer's cost type. However, analogous to Proposition 4, the regulator would not require an arbitrarily high effort because a higher effort requirement would imply a higher information rent. Consequently, the regulator chooses the effort level $\lambda^{VE}(\theta)$ that optimizes its access versus risk tradeoff subject to the manufacturer's participation constraint. Notably, in this case, the regulator also does not need to be concerned about the enforceability of withdrawing a drug for an overdue post-market study because non-compliance is prevented through ex-ante monitoring. Thus, the effort-dependent user fee menu can be applied even to a manufacturer with a cost type above the threshold $\hat{\theta}$.

Figure 4 illustrates the optimal effort-dependent user fee menu described in Proposition 6 under two different weights on risk ($w = 130, 800$). Given w, the regulator requires a higher effort (see Figure 4(a)) but a lower user fee (see Figure 4(b)) for a lower cost (i.e., more efficient) manufacturer to induce the manufacturer to reveal its cost type. Figure 4(a) also suggests that the regulator would impose a higher effort requirement if its weight on risk associated with the drug is higher. Thus, similar to the deadline-dependent user fee menu from Figure 3, an especially high w can yield a negative effort-dependent user fee.

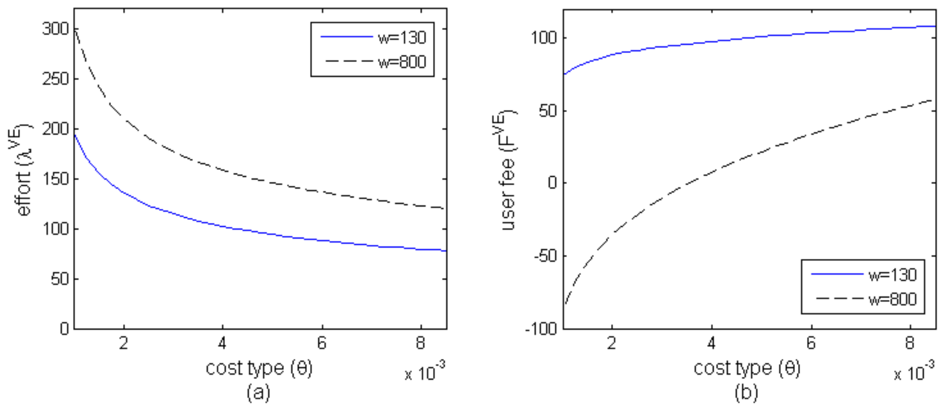

Figure 4: Optimal effort and corresponding user fee ($s = 1, \alpha = 0.85, \pi_A = 19.8, \pi_R = 37, n = 405, T = 7, k = 2$)

7. Numerical Study

In this section, we study three questions numerically, based on parameters of a real drug, to provide additional insights that complement our analytical results. First, what is the impact of a higher enforceability (or, equivalently, a higher success probability) on the manufacturer, the regulator and the patients? Second, what is the patient welfare loss resulting from the regulator not knowing the manufacturer's private cost type θ? Third, what is the welfare gain of verifying the manufacturer's effort on post-market study? To answer these questions, we numerically evaluate and compare the optimal solutions to Problem (1), Problem (2) and Problem (3).

To conduct the study based on realistic parameters, we choose one specific drug as the base for the numerical analysis, thereby providing calibrated answers to all three of the above questions. We also vary different parameters for robustness. Specifically, we parameterize our model using the data from ProAmatine, a representative case of a drug that was approved through AP but has yet to have its post-market study completed. Recall from Introduction that ProAmatine entered the market through AP in September 1996 to treat low standing blood pressure; yet, two decades later, its post-market study remains uncompleted.

Parameter Estimation

We collected and compiled data from several sources including:

1. FDA Orange Book: Patent information of all brand drugs including ProAmatine[3].

2. 1999 Annual report of Shire (the manufacturer of ProAmatine): Sales of ProAmatine upon its initial approval[4].

[3]FDA Orange Book: `https://www.accessdata.fda.gov/scripts/cder/ob/default.cfm`.

[4]1999 Annual Report of Shire: `http://investors.shire.com/~/media/Files/S/Shire-IR/annual-interim-reports/archive/shire99.pdf`.

3. The website *clinicaltrial.gov*: Information regarding clinical trials, including all post-market studies.

4. The medical literature and interactions with FDA: Information regarding parameters such as the average success rate of post-market study.

Specifically, from the FDA Orange Book, we estimate that ProAmatine had $T = 7$ years of patent remaining when it entered the market through AP. From the 1999 annual report of Shire, we approximate the sales of ProAmatine under AP with its early sales in 1999 as $\pi_A = \$19.8$ million. The sales ProAmatine would obtain if it were converted to regular approval is approximated by the combined sales in 1999 of ProAmatine and Florinef as $\pi_R = \$37$ million, where Florinef is the only alternative that was prescribed off-label for treating this disease (i.e., Florinef is not FDA approved for this disease). According to *clinicaltrial.gov*, five post-market studies are registered under ProAmatine (i.e., Study 401, 403, 404, 405, 406), from which we estimate the total sample size required for the post-market study of ProAmatine to be $n = 405$ patients. We normalize the manufacturer's reservation profit Π_0 to zero, which is common in the mechanism design literature (e.g., Iyer et al., 2005; Chick et al., 2016). In our context, this also indicates that participation does not imply compliance, which is usually the case observed in practice.

Based on DiMasi et al. (2010), the probability of success for a new drug varies from 66.7% to 100% for different diseases. Note that, as a general rule, α cannot be too low because, if so, the regulator would not have approved the drug initially under the AP pathway. Calibrating these percentages with the feedback we received from one FDA official, we estimate the probability that a post-market study will confirm the clinical benefits of ProAmatine to be approximately $75\% - 85\%$. Accordingly, we adjust the value of $\alpha \in \{0.75, 0.8, 0.85\}$ to see its impact on our results. In a similar vein, since the estimation of the regulator's weight on risk towards ProAmatine, w, is not straightforward, as a starting point, we approximate the risk measured in monetary terms as equivalent to the manufacturer's sales under AP, i.e., $w = \pi_A = 20$. However, we also consider $w = 130$ to see its effect on our results. A complete list of estimated parameters and brief rationale of the estimation is summarized in Table 1.

Table 1: Model parameters in the numerical analysis

Parameters	Estimation Rationale	Values
T	Remaining patent length upon AP from FDA patent data	7 years
π_A	Estimated as sales in 1999, i.e., three years after AP	\$19.8M
π_R	Market size combining ProAmatine and Florinef in 1999	\$37M
n	Total sample size required for the post-market study	405
α	Avg. success probability of new drug applications	{0.75, 0.8, 0.85}
w	Regulator's weight on risk of approving ineffective drug	{\$20M, \$130M}

Given that information on the cost of a post-market study is not available, we stipulate the quadratic form for $C(\theta, \lambda)$ to capture its convexity (i.e., we set $k = 2$) and set the bounds for the uniformly distributed θ as $\underline{\theta} = 0.001$ and $\bar{\theta} = 0.0085$. We set these bounds to be internally consistent with the parameter data in Table 1. Indeed, upon applying the data from Table 1, we found, through preliminary analysis of the regulator's optimal deadline-dependent user fee menu, that a manufacturer with a cost type below 0.001 would always complete its post-market study regardless of enforceability s or success probability α, whereas the manufacturer with a cost type above 0.0085 would never complete its post-market study regardless of enforceability s or success probability α. However, for a manufacturer with $\theta \in [0.001, 0.0085]$, the probability of compliance varies with s and α, thus setting the stage for interpretable results.

7.1 Compliance Probability and Expected Patient Welfare

Because the combination of enforceability s and success probability α plays an important role in inducing the manufacturer's compliance, in this subsection, we explore the impact of the optimal deadline-dependent user fee menu on the manufacturer, the regulator and patients as a function of s and α. Table 2 summarizes the results for $w = 20$ and $w = 130$.

Consistent with Proposition 2, Table 2 illustrates how compliance probability increases with both enforceability and success probability. Indeed, as Table 2 shows, if it is impossible to enforce withdrawal of the drug under AP for an overdue post-market study (i.e., if $s = 0$) when the success probability also is at the low end of its range (i.e., $\alpha = 0.75$), then the likelihood of the manufacturer complying with its post-market study requirement is a mere 19.5%, regardless of the regulator's weight on risk associated with the drug. With such a low probability of compliance, this drug would pose substantial risk to exposed patients, hence reducing patient welfare dramatically. However, if enforceability were such that the expected time for withdrawal were 2 years (i.e., if $s = 0.5$), the manufacturer's effort invested in its post-market study would increase considerably and, as a result, the likelihood of compliance would increase to 51.3%-65.8%, depending on the success probability of the drug. Continuing this trend, if enforceability were such that the expected time for withdrawal were only 1 year (i.e., if $s = 1$) when the success probability is at the high end of its range, then the likelihood of compliance would further increase to 79.3%. Correspondingly, the risk associated with AP would drop substantially and patient welfare would increase accordingly. Moreover, the higher is the regulator's weight on risk associated with the drug, the more pronounced is the improvement of patient welfare associated with this increased compliance probability. Yet, notably, Table 2 also shows that if the manufacturer does comply, the average time to complete its study is relatively insensitive to different levels of enforceability and success probability. Hence, the key for the regulator boils down simply to inducing compliance.

Table 2: Comparison among different levels of enforceability and success probabilities

	$w = 20$				$w = 130$			
	$s = 0$	$s = 0.5$	$s = 1$	$s = \infty$	$s = 0$	$s = 0.5$	$s = 1$	$s = \infty$
$\alpha = 0.75$								
Mfg's Compliance Prob.	19.5%	51.3%	64.9%	84.3%	19.5%	51.3%	64.9%	84.3%
Manufacturer's Effort	19.62	55.14	67.48	83.39	19.62	66.33	79.58	98.12
Time-to-Complete (yrs)	4.03	3.82	3.91	4.13	4.03	3.28	3.43	3.63
Benefit of Access ($M)	71.58	69.22	70.65	75.00	71.58	70.10	71.67	76.46
Disutility of Risk ($M)	32.13	23.62	21.89	21.16	208.82	144.21	132.07	124.11
Patient Welfare ($M)	39.46	45.60	48.76	53.85	-137.23	-74.11	-60.40	-47.65
$\alpha = 0.80$								
Mfg's Compliance Prob.	27.1%	58.5%	72.1%	91.7%	27.1%	58.5%	72.1%	91.7%
Manufacturer's Effort λ	29.72	61.99	74.33	90.21	29.72	71.05	84.00	101.87
Time-to-Complete (yrs)	4.01	3.85	3.92	4.11	4.01	3.44	3.56	3.75
Benefit of Access($M)	73.55	72.72	74.70	79.62	73.55	73.67	75.76	81.08
Disutility of Risk ($M)	24.53	18.52	17.22	16.72	159.43	114.32	105.44	100.24
Patient Welfare ($M)	49.03	54.20	57.48	62.90	-85.88	-40.65	-29.69	-19.16
$\alpha = 0.85$								
Mfg's Compliance Prob.	34.7%	65.8%	79.3%	99.0%	34.7%	65.8%	79.3%	99.0%
Manufacturer's Effort λ	37.22	68.90	81.23	97.09	37.22	75.66	88.38	105.65
Time-to-Complete[a] (yrs)	3.93	3.85	3.92	4.10	3.93	3.58	3.68	3.86
Benefit of Access[b] ($M)	75.65	76.52	79.07	84.59	75.65	77.40	80.02	85.87
Disutility of Risk ($M)	17.73	13.58	12.68	12.38	115.23	84.89	78.83	75.84
Patient Welfare[c] ($M)	57.92	62.94	66.38	72.21	-39.58	-7.48	1.19	10.03

Note: [a]Time-to-complete is the average time to complete a study *if the manufacturer were to comply.* [b]Benefit of Access is computed as the expected patient surplus from purchasing a drug. [c]Patient Welfare = Benefit of Access - Disutility of Risk.

Given Table 2, let $E(U_{AI,NonVE})$ denote the expected patient welfare across all possible values of θ under the optimal deadline-dependent user fee menu from Proposition 4. Then Figure 5, which plots $E(U_{AI,NonVE})$ as a function of enforceability s and success probability α for two different values of w, highlights two important insights. First, the expected patient welfare increases with enforceability but the increase soon levels off. Specifically, Figure 5 shows if the withdrawal of a drug can be implemented within, say, 1-2 years (i.e., if $s \in [0.5, 1]$), the expected patient welfare can potentially increase substantially, but not much beyond that. Second, the benefit of increased enforceability is higher for either a higher α or a higher w. Intuitively, this is true because the higher is w, the more risky is exposing patients to a potentially ineffective drug. Hence, the ability to withdraw the drug in a timely fashion is especially important for patient welfare when w is high. Similarly, when α is lower, the disincentive for compliance is greater and the regulatory incentive from penalizing non-compliance is especially important.

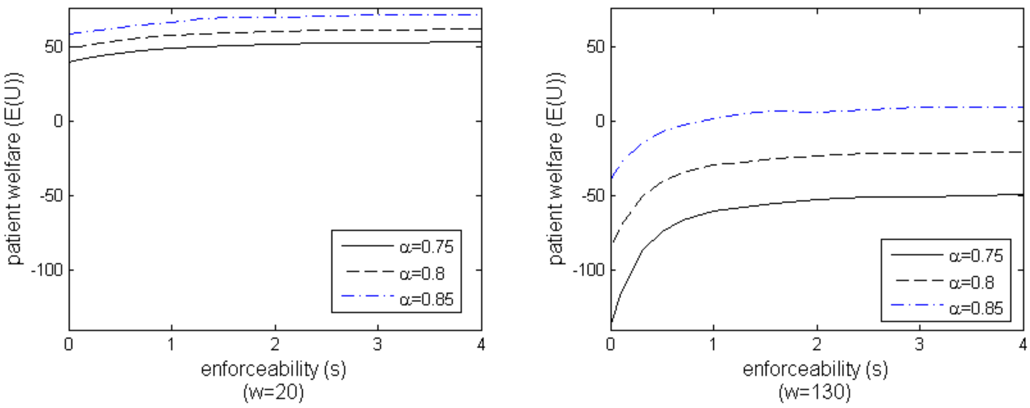

Figure 5: Impacts of enforceability of sanction and drug success probability on patient welfare

7.2 Welfare Loss from Asymmetric Information and Welfare Gain From Verifying Effort

In this subsection, we first calculate expected patient welfare loss from asymmetric information and then we calculate expected patient welfare gain from verifying effort. Not knowing the manufacturer's cost type in advance means the regulator essentially pays an information rent to induce the manufacturer to choose the optimal deadline. Moreover, as Corollary 1 indicates, the regulator's consideration of such information rent under asymmetric information may result in a relatively larger deadline than that under complete information. Accordingly, let $E(U_{CI,NonVE})$ denote the expected patient welfare across all possible values of cost type θ under the optimal deadline from Proposition 5. Then the expected patient welfare loss from asymmetric information can be measured by $E(U_{CI,NonVE}) - E(U_{AI,NonVE})$. Figure 6(a) graphs this loss as a function of enforceability s for the illustrative case in which $\alpha = 0.85$ and $w = 130$. As Figure 6(a) indicates, knowing the manufacturer's cost type becomes more valuable as s increases. Indeed, if $s = 0$, then even if the regulator knows the manufacturer's cost type and can correspondingly determine an appropriate deadline for its post-market study, such information has no value because the regulator cannot enforce its deadline regardless. Accordingly, obtaining information on the manufacturer's cost type is useful only to the extent that the regulator can enforce its deadline; and the greater is the enforceability, the more valuable is the information the regulator obtains.

In a similar vein, let $E(U_{AI,VE})$ be the expected patient welfare across all possible values of cost type θ under the optimal effort-dependent user fee from Proposition 6. Then, the expected patient welfare gain from verifying the manufacturer's effort can be measured by $E(U_{AI,VE}) - E(U_{AI,NonVE})$. Figure 6(b) graphs this gain as a function of enforceability s for the illustrative case in which $\alpha = 0.85$ and $w = 130$. As Figure 6(b) illustrates, enforceability does not affect $E(U_{AI,VE})$ if the regulator can verify the manufacturer's effort. However, recall, enforceability is crucial to in-

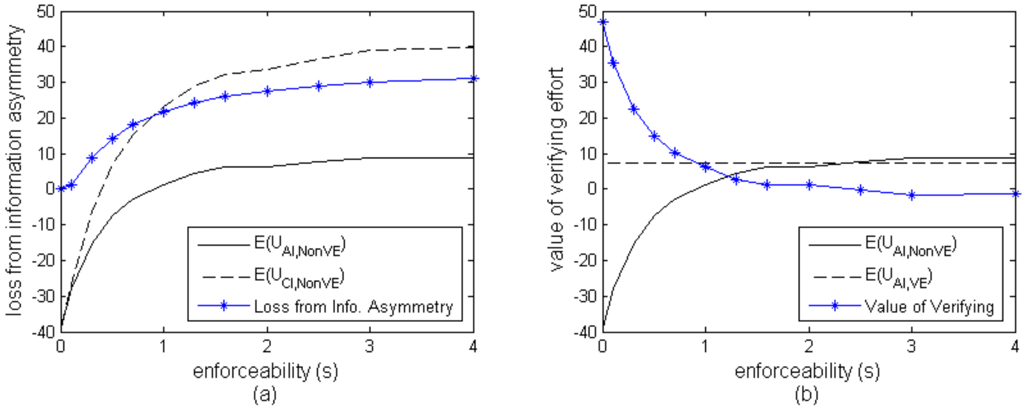

Figure 6: Welfare loss from asymmetric information and welfare gain from verifying effort ($w = 130$, $\alpha = 0.85$)

duce compliance if the manufacturer's effort is not verifiable. Accordingly, $E(U_{AI,NonVE})$ increases as s increases. As a result, the expected patient welfare gain from verifying the manufacturer's effort decreases with s. This implies that verifying effort is most valuable to the regulator when $s = 0$ because in such a case the regulator is unable to ensure compliance through a deadline-imposed sanction. As s increases, verifying the manufacturer's effort becomes less valuable because the enhanced enforceability can already induce the manufacturer's compliance with a deadline.

Figure 6 (a) and (b) together imply the substitutional effects between knowing the manufacturer's cost type and verifying the manufacturer's effort in managing a post-market study under AP. On the one hand, knowing the manufacturer's private cost type enables the regulator to set an appropriate deadline for the post-market study while compliance with the deadline partly depends on the enforceability of the ex-post sanction. On the other hand, verifying the manufacturer's effort renders the manufacturer's cost type irrelevant since the regulator's ex-ante monitoring prevents non-compliance. Therefore, the relative impact of ex-post sanction versus ex-ante monitoring depends on enforceability. Everything else being equal, under high enforceability, it is better for the regulator to implement the deadline-dependent user fee menu to set an appropriate deadline and, thus, to deter non-compliance through ex-post sanction; but under low enforceability, it is better for the regulator to implement the effort-dependent user fee to verify the manufacturer's effort on post-market study and, thus, to eliminate non-compliance through ex-ante monitoring.

8. Extensions

In this section, we further explore the implications of our results by providing a comparative analysis of three modeling extensions. Specifically, in Sections 8.1 and 8.2, we compare our optimal menu-based mechanism with an optimal single-deadline and a central planner's solution, respectively.

Then, in Section 8.3, we assess implications of private information regarding the success probability of the manufacturer's drug.

8.1 The Optimal Single Deadline

Our optimal mechanism indicates that the regulator should provide a menu of deadline-dependent user fees to the manufacturer. This would require FDA to change its current policy of a fixed user fee. Nevertheless, if such a policy change is prohibitive, our previous analysis can be modified to calculate a single deadline that induces compliance to the extent possible. In this subsection, we describe how to determine such a single deadline optimally, examine the value of the optimal menu compared to this single deadline, and investigate when the optimal menu would significantly outperform the single deadline as compared to when the single deadline can be a sufficient substitute for the optimal menu.

To calculate the optimal single deadline, the regulator would impose an additional constraint in Problem (1) to restrict $d(\theta)$ to be a constant. Proposition 7 describes the resulting solution.

Proposition 7 *If the regulator were restricted to impose a single deadline, then $d_1(\hat{\theta})$ is the optimal deadline to impose.*

Because of the compliance constraint in Problem (1), the optimal single deadline prescribed by Proposition 7 induces compliance from any manufacturer that would comply under the optimal menu-based mechanism. As a result, the regulator should impose the longest deadline from the optimal menu, which is $d_1(\hat{\theta})$. Correspondingly, the regulator would achieve the same compliance probability $\Psi(\hat{\theta})$ with its optimal single deadline as it would achieve with its optimal menu. However, with the single deadline, the regulator cannot induce the same level of effort from the manufacturer as it could with the optimal menu, thereby, resulting in, on average, a longer completion time for the post-market study. Thus, regardless of the manufacturer's cost type, patients would be forced to endure a higher risk of being exposed to a potentially ineffective drug and consequently suffer from less welfare.

To illustrate this, Table 3 compares results of the optimal single deadline with those of the optimal menu for the representative case of the ProAmatine example in which $\alpha = 0.85$. Note from Table 3 that the difference in completion time between the optimal single deadline and the optimal menu is less concerning if the regulator's weight on risk is relatively small (e.g., if $w = 20$). Hence, in such a case, the optimal single deadline would perform well as a substitute to the optimal menu. However, if the weight on risk is comparatively large, then the single deadline cannot be adjusted to induce the manufacturer to invest more effort, thereby resulting in a lower patient welfare as compared with the optimal menu. Thus, in such a case, the optimal single deadline would not be as suitable a substitute for the optimal menu.

Table 3: Comparison between the optimal menu-based mechanism and the optimal single deadline

	$w = 20$				$w = 130$			
	$s = 0$	$s = 0.5$	$s = 1$	$s = \infty$	$s = 0$	$s = 0.5$	$s = 1$	$s = \infty$
Optimal menu-based mechanism								
Mfg's Compliance Prob.	34.7%	65.8%	79.3%	99.0%	34.7%	65.8%	79.3%	99.0%
Mfg's Effort λ	37.22	68.90	81.23	97.09	37.22	75.66	88.38	105.65
Time-to-Complete (yrs)	3.93	3.85	3.92	4.10	3.93	3.58	3.68	3.86
Patient Welfare ($M)	57.92	62.94	66.38	72.21	-39.58	-7.48	1.19	10.03
Optimal Single Deadline								
Mfg's Compliance Prob.	34.7%	65.8%	79.3%	99.0%	34.7%	65.8%	79.3%	99.0%
Mfg's Effort λ	37.22	67.25	79.50	95.10	37.22	67.25	79.50	95.10
Time-to-Complete (yrs)	3.93	3.91	3.97	4.16	3.93	3.91	3.97	4.16
Patient Welfare ($M)	57.92	62.61	66.03	71.78	-39.58	-12.74	-4.41	2.87

8.2 The Centralized Solution

In this paper, we adopted a Stackelberg game framework to model the current decentralized regulator-manufacturer system for completing post-market studies. In this subsection, we explore implications if the regulator were a central planner that assumes the responsibility for conducting the post-market study itself. In principle, this provides an alternative solution to the regulator's compliance problem because it would altogether eliminate the need to control the manufacturer's effort either directly or indirectly. In essence, this alternative would mean that the regulator waives the manufacturer's obligation to conduct the post-market study and instead funds the study itself by charging the manufacturer a lump-sum payment.

In this centralized case, the regulator would be able to determine the effort invested in the post-market study directly, without concerns of asymmetric information or moral hazard. Accordingly, the regulator would determine the optimal effort for the post-market study by maximizing the combined utility of both patients and the manufacturer, i.e., $\max_\lambda U(\lambda) + \Pi(\lambda, \theta_R)$, where θ_R represents the regulator's efficiency in conducting the post-market study itself.

Let $\lambda^{cs}(\theta_R)$ denote the regulator's optimal effort for the post-market study given θ_R. Then, Proposition 8 describes the optimal effort $\lambda^{cs}(\theta_R)$.

Proposition 8 *The central planner's optimal effort $\lambda^{cs}(\theta_R)$ solves the first-order condition $(\frac{3}{2}(\alpha\pi_R - \pi_A) + w(1-\alpha))G(T, \lambda, n+1)\frac{n}{\lambda^2} = \theta_R k \lambda^{k-1}$.*

Since the regulator, as a central planner, determines directly the effort to invest in completing the post-market study, we benchmark the centralized solution with the verifiable effort case from Section 6.2 to evaluate the potential benefit if the regulator could conduct the post-market study itself. To ensure a fair comparison, we assume that the regulator is equally efficient as the manufacturer in conducting the study, i.e., $\theta_R = \theta$. Table 4 summarizes the comparison for the representative case of ProAmatine example in which $\alpha = 0.85$.

Table 4: Comparison between the centralized solution and the decentralized solution for the verifiable effort case

	Centralized		Decentralized with Verifiable Effort	
	$w = 130$	$w = 20$	$w = 130$	$w = 20$
Post-market study Effort λ	124.58	102.31	105.77	86.78
Time-to-Complete (yrs)	3.38	4.11	4.06	4.95
Patient Welfare (\$M)	24.57	73.79	7.38	66.43

Table 4 suggests that, if the regulator were to conduct the post-market study itself, it would invest higher effort than that of the manufacturer in the decentralized case. As a result, under centralization, the post-market study would be completed sooner and hence, patients would benefit significantly from the early access to a drug with its true clinical benefits verified. This is especially beneficial for cases in which the regulator's weight on risk is relatively high.

8.3 Asymmetric Information on Success Probability

In our analysis of the regulator's compliance problem, we assumed that a drug's success probability is common knowledge between the regulator and the manufacturer. We made this assumption because the manufacturer is required by FDA to disclose all available clinical trial information when it applies for AP approval, thus assuring that the regulator is equally as informed as the manufacturer on a given drug's success probability. It nevertheless is conceivable that, in some cases, the manufacturer could hold private information on the drug's potential for success. Thus, in this subsection, we extend our model by considering the asymmetric information case in which the manufacturer knows α, but the regulator is limited to characterizing α by a subjective probability distribution. Intuitively, this suggests that the regulator's compliance problem is amplified because the manufacturer would be further deterred to comply with its post market study if it knew its drug had a lower success probability than the regulator inferred.

In terms of the complication introduced because of the regulator's limited knowledge on the success probability of the manufacturer's drug, if information regarding both α and θ is asymmetric, then the regulator's corresponding mechanism design problem would boil down to having to establish a two-dimensional deadline-dependent user fee menu $(d(\theta, \alpha), F(\theta, \alpha))$ that induces the manufacturer to implicitly reveal both α and θ. However, characterizing such a menu is, in general, an intractable problem (Kostamis and Duenyas 2011)[5]. Thus, to develop insight on how the two-

[5] The specific technical challenge in solving for the optimal two-dimensional revelation mechanism in our context is two-fold: First, there exists no exogenous ordering of the two-dimensional type (θ, α) that, in turn, depends on the deadline menu $d(\theta, \alpha)$. Second, by stipulation, we have only one independent contract instrument (because the user fee depends on the deadline) available to force the manufacturer to reveal two distinct components of privately

dimensional menu should depend on α, we investigate here the special case in which information regarding α is asymmetric, but information regarding θ is not. For this special case variant of the regulator's compliance problem, analogous to Problem (1), the regulator would design a deadline-dependent user fee menu contingent on α, $(d(\alpha), F(\alpha))$, to induce the manufacturer to reveal α by solving the following:

$$\begin{aligned}
\max_{d(\alpha),F(\alpha)} \int_{\underline{\alpha}}^{\bar{\alpha}} & [U(d(\alpha)) + F(\alpha)]\psi(\alpha)d\alpha \\
\text{s.t. } & \Pi_C(d(\alpha),\alpha) - F(\alpha) \geq \Pi_C(d(\tilde{\alpha}),\alpha) - F(\tilde{\alpha}), \forall \tilde{\alpha} \in [\underline{\alpha}, \bar{\alpha}] \\
& \Pi_C(d(\alpha),\alpha) - F(\alpha) \geq \Pi_0, \\
& \Pi_C(d(\alpha),\alpha) \geq \Pi_N(d(\alpha)), \forall \theta \leq \hat{\theta}.
\end{aligned}$$

where $\psi(\alpha)$ denotes the subjective distribution characterizing the regulator's belief of α. Proposition 9 next establishes a key structural property of the regulator's optimal menu for this problem.

Proposition 9 *Let $(d^{AI}(\alpha), F^{AI}(\alpha))$ denote the regulator's optimal deadline-dependent user fee menu as a function of α. Then, both $d^{AI}(\alpha)$ and $F^{AI}(\alpha)$ are decreasing in α.*

Intuitively, this optimal menu prescribes a longer deadline as an incentive to offset the manufacturer's deterrence to comply when its drug has a lower success probability. But, in exchange, the regulator then can charge a correspondingly higher user fee to subsidize the implicit cost of that incentive.

Now, with the characterization of the one-dimensional mechanism for asymmetric α from Proposition 9 and its analog for asymmetric θ from Proposition 4, we conjecture that, for the two-dimensional information asymmetry case in which both α and θ are privately known by the manufacturer, the regulator's optimal deadline-dependent user fee menu $(d^{AI}(\theta,\alpha), F^{AI}(\theta,\alpha))$ should increase with θ but decrease with α such that shorter deadlines are associated with high cost efficient manufacturers that develop drugs with high probabilities of success while longer deadlines are associated with low cost efficient manufacturers that develop drugs with low probabilities of success. However, we defer to future research for precise indexing of deadlines for specific manufacturer-drug profiles.

9. Conclusion

FDA instituted the accelerated-approval pathway (AP) in 1992 for drugs targeted at serious diseases without alternative treatments to expedite access to new drugs. Essentially, AP allows promising

held information (namely, θ and α), thereby rendering it impossible to perfectly differentiate different manufacturer types.

drugs to enter the market based on limited evidence of efficacy, thereby permitting clinical trials required to verify true clinical benefits to be conducted as post-market studies. However, many required post-market studies are not completed as promised. Moreover, FDA must endure an onerous process to withdraw an unproven drug from the market when a post-market study is uncompleted. Consequently, FDA faces substantial risk of an ineffective drug remaining on the market indefinitely when a manufacturer does not comply with its requirement to complete its post-market study, thereby compromising the original purpose of AP.

Our study thus aims to explore a potentially implementable and internally consistent solution to FDA's non-compliance problem through a comprehensive analysis of the myriad complicating factors and tradeoffs. Toward that end, we develop a deadline-dependent user fee menu to ensure compliance. In establishing this menu, we optimize the regulator's tradeoff between providing public access to potentially effective drugs and mitigating public health risks from ineffective drugs. In so doing, we address the regulator's challenge of having to impose a post-market study deadline without being able to observe the manufacturer's private cost information or its level of effort. From a practical standpoint, by tying the user fee already in place to fund a new drug application to the post-market study deadline, we leverage an existing FDA mechanism into an incentive for the manufacturer to complete its post-market study, thus addressing FDA's associated asymmetric information and moral hazard challenges. If, in contrast, the current format of the fixed user fee cannot be altered into an deadline-dependent fee, then our analysis can be modified to calculate an optimal single deadline that can be imposed as a simple substitute to the menu, but to limited effect: The suitability of this substitute dissipates if the regulator's weight on risk is comparatively high.

While, in principle, our optimal deadline-dependent user fee menu enables the regulator to impose a deadline that better motivates the manufacturer to complete its post-market study, our analysis also shows that the effectiveness of the menu in inducing compliance depends not only on the drug's success probability, but also on what we call the enforceability of sanction, s, a measure indicating how difficult it is to withdraw a drug from the market if a given post-market study is not completed. For any given success probability α, as long as a drug's enforceability is sufficiently high, in particular, if $s \geq \hat{\theta}^{-1}(\bar{\theta}; \alpha)$, our deadline-dependent user fee menu is guaranteed to induce manufacturer compliance. And, the higher is the drug's success probability, the lower is the enforceability needed for the menu to guarantee compliance. More generally, the higher is either a drug's enforceability or its success probability, the higher is the probability that the menu will induce the manufacturer to comply with its required post-market study; and the lower is either, the lower is the probability that the menu will induce the manufacturer to comply. In the extreme, this means that, with everything else equal, if a drug's enforceability is especially low, in particular, if $s \leq \hat{\theta}^{-1}(\underline{\theta}; \alpha)$, then no deadline would induce manufacturer compliance.

Otherwise, effort monitoring very well might be the only way for the regulator to ensure compliance under AP. While monitoring the manufacturer effort currently is not standard practice and potentially requires significant cost and planning to implement, FDA nevertheless is beginning to explore the viability of the option (GAO, 2009). Hence, to respond to such a case, the regulator should explore how to design unambiguous milestones and to monitor the manufacturer's effort efficiently to ensure compliance. To the extent that such milestones are infeasible or impractical, a last resort would be for the regulator to consider monitoring effort directly by managing the post-market study itself rather than leaving it to the manufacturer to manage. While such an extreme measure would no doubt have serious resource implications for the regulator, it nevertheless would assure a timely completion of the study.

While we have assumed in this paper that enforceability is exogenously given, our results indicate that the regulator's benefits increase with the enforceability of a given drug. Indeed, with the implementation of the deadline-dependent user fee menu, not only would the probability of manufacturer compliance increase as s increases, but also would the patient welfare increase as a result. In fact, the welfare loss from not being able to withdraw from the market an unproven drug is substantial, especially when the risk of exposing patients to a potentially ineffective drug is prominent. Therefore, if enforceability is especially low for a given drug, then as an alternative to investing in effort monitoring, the regulator also should consider investing resources to increase the drug's enforceability as, for example, GAO (2009) has suggested. Toward that end, the regulator could embark on a campaign to educate the public about the potential risks embedded in AP as well as the potential negative consequences associated with unproven drugs left on the market indefinitely.

Finally, while we have focused on the non-compliance problem that arises once AP is granted for a given drug, our analysis does provide important insights to the regulator on granting AP approval in the first place. If the enforceability of a given drug is expected to be low, then the regulator may require a higher success probability for granting AP approval to achieve a desired level of compliance. However, if either a given drug's success probability or enforceability is too low, if effort monitoring is too costly, and if enforceability cannot be increased, then the regulator probably would be better served requiring the manufacturer to adhere to the regular approval pathway rather than granting AP. In short, moving forward, the regulator should incorporate its likelihood of inducing post-market study compliance into its AP decision making process.

References

Butler D (2008) Translational research: crossing the valley of death. *Nature News* 453(7197):840–842.

Chick SE, Hasija S, Nasiry J (2016) Information elicitation and influenza vaccine production. *Oper. Res.* 65(1):75–96.

Cohen MA, Eliasberg J, Ho TH (1996) New product development: The performance and time-to-market tradeoff. *Management Sci.* 42(2):173–186.

Collins F (2012) How can we do better? accessed Sept. 21, 2017, `https://irp.nih.gov/catalyst/v20i3/collins-how-can-we-do-better` .

Crocker KJ, Letizia P (2014) Optimal policies for recovering the value of consumer returns. *Production and Operations Management* 23(10):1667–1680.

Cutting Edge Information (2004) Accelerating clinical trials – budgets, patients recruitment and productivity. accessed Sept. 21, 2017, `https://firstclinical.com/journal/2005/0503_Accelerating.pdf` .

Dagher R, Johnson J, Williams G, Keegan P, Pazdur R (2004) Accelerated approval of oncology products: a decade of experience. *Journal of the National Cancer Institute* 96(20):1500–1509.

Dai T, Cho SH, Zhang F (2016) Contracting for on-time delivery in the us influenza vaccine supply chain. *Manufacturing & Service Oper. Management* 18(3):332–346.

DiMasi JA, Feldman L, Seckler A, Wilson A (2010) Trends in risks associated with new drug development: success rates for investigational drugs. *Clinical Pharmacology & Therapeutics* 87(3):272–277.

DiMasi JA, Grabowski HG, Hansen RW (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics* 47:20–33.

FDA (2015) The beginnings: Laboratory and animal studies. accessed Sept. 21, 2017, `https://www.fda.gov/Drugs/ResourcesForYou/Consumers/ucm143475.htm` .

FDA (2017) Prescription Drug User Fee Act (PDUFA). accessed Sept. 21, 2017, `https://www.fda.gov/ForIndustry/UserFees/PrescriptionDrugUserFee/default.htm` .

Fleming TR (2005) Surrogate endpoints and FDA's accelerated approval process. *Health Affairs* 24(1):67–78.

Fudenberg D, Tirole J (1991) *Game Theory* (The MIT Press).

GAO (2009) Government Accountability Office: FDA needs to enhance its oversight of drugs approved on the basis of surrogate endpoints. accessed at Sept. 21, 2017, `http://www.gao.gov/new.items/d09866.pdf` .

Gawande K, Bohara AK (2005) Agency problems in law enforcement: Theory and application to the us coast guard. *Management Sci.* 51(11):1593–1609.

Gellad WF, Kesselheim AS (2017) Accelerated approval and expensive drugs -?a challenging combination. *New England Journal of Medicine* 376(21):2001–2004.

Gray WB, Deily ME (1996) Compliance and enforcement: Air pollution regulation in the US steel industry. *Journal of Environmental Economics and Management* 31(1):96–111.

Guide for Industry (2014) Expedited programs for serious conditions drugs and biologics. accessed Sept. 21, 2017, `https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm358301.pdf` .

Iyer AV, Schwarz LB, Zenios SA (2005) A principal-agent model for product specification and production. *Management Sci.* 51(1):106–119.

Johnson JR, Ning YM, Farrell A, Justice R, Keegan P, Pazdur R (2011) Accelerated approval of oncology products: the Food and Drug Administration experience. *Journal of the National Cancer Institute* 103(8):636–644.

Kolstad CD, Ulen TS, Johnson GV (1990) Ex post liability for harm vs. ex ante safety regulation: substitutes or complements? *The American Economic Review* 80(4):888–901.

Kostamis D, Duenyas I (2011) Purchasing under asymmetric demand and cost information: When is more private information better? *Oper. Res.* 59(4):914–928.

Kraft T, Zheng Y, Erhun F (2013) The NGO's dilemma: How to influence firms to replace a potentially hazardous substance. *Manufacturing & Service Oper. Management* 15(4):649–669.

Mayo Clinic (2010) Letter to the proposal to withdraw marketing approval. accessed Sept. 21, 2017, `https://www.regulations.gov/document?D=FDA-2007-N-0475-0032` .

Moore TJ, Furberg CD (2014) Development times, clinical testing, postmarket follow-up, and safety risks for the new drugs approved by the us food and drug administration: the class of 2008. *JAMA Internal Medicine* 174(1):90–95.

Olson MK (2004) Are novel drugs more risky for patients than less novel drugs? *Journal of Health Economics* 23(6):1135–1158.

Shabtay D, Steiner G (2007) Optimal due date assignment and resource allocation to minimize the weighted number of tardy jobs on a single machine. *Manufacturing & Service Oper. Management* 9(3):332–350.

Shavell S (1984) A model of the optimal use of liability and safety regulation. *The Rand Journal of Economics* 15(2):271–280.

Willyard C (2014) FDA's post-approval studies continue to suffer delays and setbacks. *Nature Medicine* 20:1224–1225.

Wood AJ (2006) A proposal for radical changes in the drug - approval process. *New England Journal of Medicine* 355:618–623.

Proof of Lemma 1

Define $\Phi(t) = E_\tau(\phi(t,\tau))$ as below,

$$\Phi(t) = \begin{cases} t\pi_A + (T-t)\alpha\pi_R & \text{if } t \le d, \\ \displaystyle\int_d^t \tau\pi_A s e^{-s(\tau-d)}d\tau + (t\pi_A + (T-t)\alpha\pi_R)e^{-s(t-d)} & \text{if } d < t \le T, \\ \displaystyle\int_d^T \tau\pi_A s e^{-s(\tau-d)}d\tau + T\pi_A e^{-s(T-d)} & \text{if } t > T. \end{cases}$$

It follows directly that $\Phi(t)$ is continuous and mostly differentiable, and $\Phi_t^{'}(t)$ is,

$$\Phi_t^{'}(t) = \begin{cases} -(\alpha\pi_R - \pi_A) & \text{if } t < d, \\ -e^{-s(t-d)}\big(\alpha\pi_R - \pi_A + (T-t)\alpha\pi_R s\big) & \text{if } d < t < T, \\ 0 & \text{if } t > T. \end{cases}$$

Thus, through integration by parts, we have $\Pi(\lambda,d,\theta) = \Phi(T) - \int_0^T G(t,\lambda,n)\Phi_t^{'}(t)dt - C(\lambda,\theta)$. Based on the property of Possion process, $G(t,\lambda,n) = \sum_{j=n}^{\infty} e^{-\lambda t}\frac{(\lambda t)^j}{j!}$. Hence, $G_\lambda^{'}(t,\lambda,n) = \frac{n}{\lambda^2}g(t,\lambda,n+1)$. Therefore, we have $\frac{\partial\Pi}{\partial\lambda} = n\lambda^{k-1}\big(\int_0^T \frac{g(t,\lambda,n+1)}{\lambda^{k+1}}(-\Phi_t^{'})dt - \frac{k\theta}{n}\big)$.

We define $M(\lambda) = \int_0^T \frac{g(t,\lambda,n+1)}{\lambda^{k+1}}(-\Phi_t^{'})dt - \frac{k\theta}{n}$, which has the same sign as $\frac{\partial\Pi}{\partial\lambda}$. Since $-\Phi_t^{'} \ge 0$ and bounded, we have $\lim_{\lambda\to\infty} M(\lambda) = -\frac{k\theta}{n} < 0$. Additionally, we have $\lim_{\lambda\to 0} M(\lambda) = \lim_{\lambda\to 0}\int_0^T \frac{\lambda^{n-k}t^n}{e^{\lambda t}n!}(-\Phi_t^{'})dt - \frac{k\theta}{n} = -\frac{k\theta}{n}$. Therefore, $\lim_{\lambda\to 0}\Pi^{'}(\lambda) < 0$ and thus $\lambda = 0$ is one local optimal of $\Pi(\lambda)$. Moreover, we have $\lim_{\lambda\to\infty}\Pi^{'}(\lambda) < 0$ and thus $\bar{\lambda}$ must solve the first-order condition $\frac{\partial\Pi}{\partial\lambda} = 0$ to be another local optimal.

Next, we prove when $\alpha\pi_R \ge \pi_A$, $\bar{\lambda}$ must be the larger root of the two roots of $\frac{\partial\Pi}{\partial\lambda} = 0$. We prove this by showing that $M(\lambda)$ first increases and then decreases with λ. Taking the derivative of $M(\lambda)$ against λ gives $M_\lambda^{'} = \frac{1}{\lambda^{k+2}}\int_0^T (n-k-\lambda t)g(t,\lambda,n+1)(-\Phi_t^{'})dt$. We define $N(\lambda) = \int_0^T (n-k-\lambda t)g(t,\lambda,n+1)(-\Phi_t^{'})dt$, which has the same sign as $M_\lambda^{'}$. If $\lambda \le \frac{n-k}{T}$, then $N(\lambda) > 0$. Hence, to prove $M(\lambda)$ first increases and then decreases with λ, it suffices to show that $N(\lambda)$ has exactly one non-zero root.

We first prove the existence of non-zero roots for $N(\lambda)$. Suppose there exists no λ such that $N(\lambda) = 0$. Since $N(\lambda) > 0$ for $\lambda \le \frac{n-k}{T}$, then $N(\lambda) > 0$ for all λ. Thus, $M(\lambda)$ always increases with λ, which contradicts that $\lim_{\lambda\to 0} M(\lambda) = \lim_{\lambda\to\infty} M(\lambda) = -\frac{k\theta}{n}$. Hence, there must exist at least one λ_0 such that $N(\lambda_0) = 0$.

Next, we show the uniqueness of λ_0 under two cases 1) $n-k < \lambda_0 d$; 2) $n-k \ge \lambda_0 d$.

Case 1: $n-k < \lambda_0 d$. Suppose there exist two roots λ_0^1 and λ_0^2, and $\lambda_0^2 > \lambda_0^1 > \frac{n-k}{d}$. Through change of variable by setting $x = \lambda t$, we can rewrite $N(\lambda)$ as $N(\lambda) = \int_0^{n-k}(n-k-x)g(x,1,n+1)(\alpha\pi_R - \pi_A)dx + \int_{n-k}^{T\lambda}(n-k-x)g(x,1,n+1)(-\Phi_t^{'}(x/\lambda))dx$.

Thus, we have

$$N(\lambda_0^2)-N(\lambda_0^1) = \int_{n-k}^{\lambda_0^1 T}(n-k-x)g(t,1,n+1)(-\Phi_t'(x/\lambda_0^2)+\Phi_t'(x/\lambda_0^1))dx + \int_{\lambda_0^1 T}^{\lambda_0^2 T}(n-k-x)g(t,1,n+1)(-\Phi_t'(x/\lambda_0^2))dx$$

Because $\Phi_t''(t)\geq 0$, we have $\Phi_t'(x/\lambda_0^1)\geq \Phi_t'(x/\lambda_0^2)$. Additionally, we have $n-k<\lambda_0^1 d<\lambda_0^1 T$. Together, we have $N(\lambda_0^2)-N(\lambda_0^1)<0$, which contradicts that $N(\lambda_0^2)=N(\lambda_0^1)=0$. Therefore, $N(\lambda)$ has at most one root on interval $(\frac{n-k}{d},\infty)$.

Case 2: $n-k\geq\lambda_0 d$. Suppose $N(\lambda_0)=0$, it suffices to show that $N'(\lambda_0)<0$. We evaluate $N(\lambda)$ at λ_0 as $N(\lambda_0)=\frac{1}{n!}\int_0^{\lambda_0 d}(n-k-x)e^{-x}x^n(\alpha\pi_R-\pi_A)dx+\frac{1}{n!}\int_{\lambda_0 d}^{\lambda_0 T}(n-k-x)e^{-x}x^n(-\Phi_t'(\frac{x}{\lambda_0}))dx$. If $\alpha\pi_R\geq\pi_A$, then $\int_0^{\lambda_0 d}(n-k-x)e^{-x}x^n(\alpha\pi_R-\pi_A)dx\geq 0$. Since $N(\lambda_0)=0$, then $\int_{\lambda_0 d}^{\lambda_0 T}(n-k-x)e^{-x}x^n(-\Phi_t'(\frac{x}{\lambda_0}))dx\leq 0$.

We take the derivative of $N(\lambda)$ against λ and evaluate at λ_0 as follow,

$$N'(\lambda_0) = \frac{1}{n!}\Big(-d(n-k-\lambda_0 d)e^{-\lambda_0 d}(\lambda_0 d)^n(T-d)\alpha\pi_R s + T(n-k-\lambda_0 T)e^{-\lambda_0 T}(\lambda_0 T)^n(-\Phi_t'(T))+\int_{\lambda_0 d}^{\lambda_0 T}(n-k-x)e^{-x}x^n(\Phi_t''(\frac{x}{\lambda_0})\frac{x}{\lambda_0^2})dx\Big).$$

It follows directly that the first two terms of $N'(\lambda_0)$ are both negative. Hence, we now sign the third term of $N'(\lambda_0)$. When $t\in(d,T]$, we have $\Phi_t''=(-\Phi_t')(s+\frac{\alpha\pi_R s}{\alpha\pi_R-\pi_A+(T-t)\alpha\pi_R s})$. For shorthand, we define $Q(t)=(s+\frac{\alpha\pi_R s}{\alpha\pi_R-\pi_A+(T-t)\alpha\pi_R s})\frac{t}{\lambda_0}$. $Q(t)$ is positive and increases with t for $t\geq 0$. We evaluate the third term of $N'(\lambda_0)$ as below,

$$\begin{aligned}
&\int_{\lambda_0 d}^{\lambda_0 T}(n-k-x)e^{-x}x^n(\Phi_t''(\frac{x}{\lambda_0})\frac{x}{\lambda_0^2})dx\\
=\ &\int_{\lambda_0 d}^{n-k}(n-k-x)e^{-x}x^n(-\Phi_t'(\frac{x}{\lambda_0}))Q(\frac{x}{\lambda_0})dx+\int_{n-k}^{\lambda_0 T}(n-k-x)e^{-x}x^n(-\Phi_t'(\frac{x}{\lambda_0}))Q(\frac{x}{\lambda_0})dx\\
<\ &\int_{\lambda_0 d}^{n-k}(n-k-x)e^{-x}x^n(-\Phi_t'(\frac{x}{\lambda_0}))Q(\frac{n-k}{\lambda_0})dx+\int_{n-k}^{\lambda_0 T}(n-k-x)e^{-x}x^n(-\Phi_t'(\frac{x}{\lambda_0}))Q(\frac{n-k}{\lambda_0})dx\\
=\ &Q(\frac{n-k}{\lambda_0})\int_{\lambda_0 d}^{\lambda_0 T}(n-k-x)e^{-x}x^n(-\Phi_t'(\frac{x}{\lambda_0}))dx\leq 0.
\end{aligned}$$

Together, we proved that $N'(\lambda_0)<0$. Hence, $N(\lambda)$ has at most one root on $(0,\frac{n-k}{d}]$.

We next prove that $N(\lambda)$ cannot have another root on $(\frac{n-k}{d},\infty)$ if $N(\lambda)$ has already one root on $(0,\frac{n-k}{d}]$. To prove this, we first show that $\lim_{\lambda\to\infty}N(\lambda)<0$. When $\lambda>\frac{n-k}{d}$, we have

$$\begin{aligned}
N(\lambda) &= \int_0^d(n-k-\lambda t)g(t,\lambda,n+1)(\alpha\pi_R-\pi_A)dt+\int_d^T(n-k-\lambda t)g(t,\lambda,n+1)(-\Phi_t')dt\\
&< \int_0^d(n-k-\lambda t)g(t,\lambda,n+1)(\alpha\pi_R-\pi_A)dt=(\alpha\pi_R-\pi_A)\int_0^{\lambda d}(n-k-x)e^{-x}\frac{x^n}{n!}dx\\
&= -(\alpha\pi_R-\pi_A)\Big(\frac{n-k-x}{n!}x^n e^{-x}\Big|_0^{\lambda d}-\int_0^{\lambda d}\big[(n-k)\frac{e^{-x}x^{n-1}}{(n-1)!}-(n+1)\frac{e^{-x}x^n}{n!}\big]dx\Big)
\end{aligned}$$

Thus, we have

$$\begin{aligned}\lim_{\lambda\to\infty} N(\lambda) &< -(\alpha\pi_R-\pi_A)\lim_{\lambda\to\infty}\left(\frac{n-k-x}{n!}x^n e^{-x}\bigg|_0^{\lambda d}-(n-k)\int_0^{\lambda d}\frac{e^{-x}x^{n-1}}{(n-1)!}dx\right.\\ &\left.+(n+1)\int_0^{\lambda d}\frac{e^{-x}x^n}{n!}dx\right)=-(\alpha\pi_R-\pi_A)(k+1)<0\end{aligned}$$

Hence, suppose $N(\lambda)$ has another root on $(\frac{n-k}{d},\infty)$. Based on the proof of Case 1, $N(\lambda)$ has at most one root on $(\frac{n-k}{d},\infty)$. Therefore, $\lim_{\lambda\to\infty}N(\lambda)>0$, which contradicts that $\lim_{\lambda\to\infty}N(\lambda)<0$. Therefore, $N(\lambda)$ has no another root on $(\frac{n-k}{d},\infty)$. This completes the proof.

Proof of Proposition 1

Let $f(\tau)$ and $F(\tau)$ denote the pdf and cdf of τ, respectively. Define $R(\lambda,d)$ as manufacturer's profit under immediate-sanction case with deadline d. Hence, $R(\lambda,d)=\int_0^d(t\pi_A+(T-t)\alpha\pi_R)g(t,\lambda,n)dt+d\pi_A(1-G(d,\lambda,n))-C(\lambda,\theta)$. Let $\tilde{d}=\min(d+\tau,T)$. We then can rewrite $\Pi(\lambda,d,\theta)$ as,

$$\begin{aligned}\Pi(\lambda,d,\theta) &= \int_0^\infty R(\lambda,\tilde{d})f(\tau)d\tau=R(\lambda,T)-\int_d^T F(t-d)R_d^{'}(\lambda,t)dt\\ &= R(\lambda,T)-\int_d^T F(t-d)\big(\pi_A(1-G(t,\lambda,n))+(T-t)\alpha\pi_R g(t,\lambda,n)\big)dt\end{aligned}$$

According to Lemma 1, we have $\Pi_C(d,\theta)=\Pi(\bar{\lambda},d,\theta)$ and $\Pi_N(d)=d\pi_A+\frac{1}{s}(1-e^{-s(T-d)})\pi_A$. Define $J(d)=\Pi_C(d,\theta)-\Pi_N(d)$. Manufacturer will only comply if $J(d)\geq 0$. We next prove that $J(d)$ is quasi-concave for $d\in[0,T]$ by showing that $J^{'}(d)$ has at most one root on $[0,T]$.

According to Envelope Theorem, we have $J^{'}(d)=\frac{\partial\Pi}{\partial d}(\bar{\lambda},d,\theta)-\frac{\partial\Pi_N}{\partial d}=\int_d^T se^{-s(t-d)}\big(-\pi_A G(t,\bar{\lambda},n)+(T-t)\alpha\pi_R g(t,\bar{\lambda},n)\big)dt$. Define $K(t)=-\pi_A G(t,\bar{\lambda},n)+(T-t)\alpha\pi_R g(t,\bar{\lambda},n)$. Hence, $J^{'}(d)=\int_d^T se^{-s(t-d)}K(t)dt$ and $J^{''}(d)=sJ^{'}(d)-sK(d)$. Suppose d_0 such that $J^{'}(d_0)=0$. Thus, we have $J^{''}(d_0)=-sK(d_0)$. Next, we show by contradiction that $K(d_0)>0$.

Taking derivative of $K(t)$ against t gives, $K^{'}(t)=\big(-(\pi_A+\alpha\pi_R)t+\alpha\pi_R(T-t)(n-1-\bar{\lambda}t)\big)g(t,\bar{\lambda},n)/t$. Note that $K^{'}(t)$ has the same sign as the quadratic function of t, i.e., $-(\pi_A+\alpha\pi_R)t+(T-t)\alpha\pi_R(n-1-\bar{\lambda}t)$. Since and $K^{'}(0)>0$ and $K^{'}(T)<0$, then $K^{'}(t)$ has exactly one root. Since $K(0)>0$ and $K(T)<0$, thus, $K(t)$ must first increase and decrease and cross x-axis exactly once. Suppose $K(d_0)\leq 0$, we must have $K(t)\leq 0$ for all $t\in[d_0,T]$. Therefore, we have $\int_{d_0}^T se^{-s(t-d)}K(t)dt<0$, which contradicts that $J^{'}(d_0)=0$. Therefore, we must have $K(d_0)>0$. Hence, $J^{''}(d_0)=-sK(d_0)<0$. This implies that $J^{'}(d)$ has at most one root d_0, and $J^{'}(d)>(<)$ 0 if $d<(>)\ d_0$. Thus, we proved that $J(d)$ is quasi-concave.

Next, we characterize the cost type threshold $\hat{\theta}$. It follows directly that $\Pi_C(d,\theta)-\Pi_N(d)>0$ if $\theta=0$, and $\lim_{\theta\to\infty}\Pi_C(d,\theta)-\Pi_N(d)\leq 0$. In addition, based on Envelope Theorem, $\frac{d(\Pi_C(d,\theta)-\Pi_N(d))}{d\theta}=\frac{\partial\Pi}{\partial\theta}(\bar{\lambda},d,\theta)=-\bar{\lambda}^2<0$. Hence, given d, there must exist one and only one $\theta_0(d)$ such that

$\Pi_C(d,\theta_0(d)) = \Pi_N(d)$, and if $\theta >(<)\ \theta_0(d)$ then $\Pi_C(d,\theta) <(>)\ \Pi_N(d)$. Let $\hat{\theta} = \max\left\{\theta_0(d)\middle| d \in [0,T]\right\}$. Hence, if $\theta > \hat{\theta}$, we have $\theta > \theta_0(d)$ for all $d \in [0,T]$. Therefore, $\Pi_C(d,\theta) < \Pi_N(d)$ for all d. If $\theta \le \hat{\theta}$, there exists d such that $\theta \le \theta_0(d)$ and thus we have $\Pi_C(d,\theta) \ge \Pi_N(d)$. Since $J(d)$ is quasi-concave, the level set, defined as $\left\{d\middle|J(d) \ge 0,\ 0 \le d \le T\right\}$, is thus convex. Hence, there exist d_1 and d_2 such that $J(d) \ge 0$ if $d \in [d_1, d_2]$, where d_1 and d_2 are defined by $J(d) = 0$.

Based on the derivative of implicit function, we have $\frac{\partial d_i}{\partial \theta} = -\frac{\partial J/\partial \theta}{\partial J/\partial d}, \forall i = 1,2$. Since $J(d)$ is quasi-concave, $\partial J/\partial d \ge 0$ at d_1 and $\partial J/\partial d \le 0$ at d_2. In addition, $\frac{\partial J}{\partial \theta} = -\bar{\lambda}^2 < 0$. Hence, $\frac{\partial d_1}{\partial \theta} \ge 0$ and $\frac{\partial d_2}{\partial \theta} \le 0$. This completes the proof.

Proof of Proposition 2

We prove by contradiction that $\hat{\theta}(s,\alpha)$ increases with s. Given s_1 and the corresponding $\hat{\theta}_1$, we have $\max_d\left(\Pi_C(d,\hat{\theta}_1;s_1) - \Pi_N(d;s_1)\right) = 0$. Thus, it suffices to show that for any $s_2 > s_1$, we have $\max_d\left(\Pi_C(d,\hat{\theta}_1;s_2) - \Pi_N(d;s_2)\right) > 0$.

Let $d^1 = \arg\max_d\left(\Pi_C(d,\hat{\theta}_1;s_1) - \Pi_N(d;s_1)\right)$ and $\bar{\lambda}^1 = \arg\max_\lambda \Pi(\lambda, d^1, \hat{\theta}_1; s_1)$. Hence, by definition, $\Pi(\bar{\lambda}^1, d, \hat{\theta}_1; s_1) - \Pi_N(d;s_1) \le \Pi(\bar{\lambda}^1, d^1, \hat{\theta}_1; s_1) - \Pi_N(d^1;s_1) = 0$. As defined in the proof of Proposition 1, we have $\Pi(\bar{\lambda}^1, d, \hat{\theta}_1; s) - \Pi_N(d;s) = R(\bar{\lambda}^1, T) - \pi_A T - \int_d^T F(t-d)K(t)dt$, and $K(0) > 0$ and $K(T) < 0$, and $K(t)$ crosses x-axis exactly once. Denote t^0 such that $K(t^0) = 0$. Hence, we must have $K(t) >(<)\ 0$ if $t <(>)\ t^0$. We next prove by contradiction that $d^1 \le t^0$. Suppose $d^1 > t^0$, then $K(t) < 0$ for all $t \in [d^1, T]$. Hence, for any $t^0 < d < d^1$, we have

$$\begin{aligned}
\Pi(\bar{\lambda}^1, d, \hat{\theta}_1; s_1) - \Pi_N(d;s_1) &= R(\bar{\lambda}^1, T) - \pi_A T - \int_d^{d_1} F(t-d)K(t)dt - \int_d^T F(t-d)K(t)dt \\
&> R(\bar{\lambda}^1, T) - \pi_A T - \int_{d^1}^T F(t-d^1)K(t)dt = \Pi(\bar{\lambda}^1, d^1, \hat{\theta}_1; s_1) - \Pi_N(d^1;s_1)
\end{aligned}$$

which contradicts that $d^1 = \arg\max_d\left(\Pi_C(d,\hat{\theta}_1;s_1) - \Pi_N(d;s_1)\right)$. Thus, we proved $d^1 \le t^0$.

For $s_2 > s_1$, we construct $d^2 = t_0 - (t_0 - d^1)s_1/s_2$. It follows directly that $t_0 > d^2 > d^1$. It is easy to verify that $F(t-d^1;s_1) - F(t-d^2;s_2)$ and $K(t)$ have the same sign. Since $\Pi(\bar{\lambda}^1, d^1, \hat{\theta}_1; s_1) - \Pi_N(d^1;s_1) = R(\bar{\lambda}^1, T) - \pi_A T - \int_{d^1}^T F(t-d^1;s_1)K(t)dt = 0$, thus,

$$\begin{aligned}
\Pi(\bar{\lambda}^1, d^2, \hat{\theta}_1; s_2) - \Pi_N(d^2;s_2) &= R(\bar{\lambda}^1, T) - \pi_A T - \int_{d^2}^T F(t-d^2;s_2)K(t)dt \\
&= \int_{d^2}^T \left(F(t-d^1;s_1) - F(t-d^2;s_2)\right)K(t)dt + \int_{d^1}^{d^2} F(t-d^1;s_1)K(t)dt.
\end{aligned}$$

Because $t_0 > d^2$, thus, we have $K(t) > 0$ for $t \le d^2$. Hence, we have $\Pi(\bar{\lambda}^1, d^2, \hat{\theta}_1; s_2) - \Pi_N(d^2;s_2) > 0$. By definition, we have $\max_d\left(\Pi_C(d,\hat{\theta}_1;s_2) - \Pi_N(d;s_2)\right) > \Pi(\bar{\lambda}^1, d^2, \hat{\theta}_1; s_2) - \Pi_N(d^2;s_2)$. Thus, $\max_d\left(\Pi_C(d,\hat{\theta}_1;s_2) - \Pi_N(d;s_2)\right) > 0$.

We next prove that $\hat{\theta}(s,\alpha)$ increases with α. By definition, we have $\hat{\theta} = \max\{\theta_0(d)\big| d \in [0,T]\}$. Based on the definition of θ_0, we have $\frac{\partial \theta_0}{\partial \alpha} = -\frac{\partial(\Pi_C(d,\theta)-\Pi_N(d))/\partial\alpha}{\partial(\Pi_C(d,\theta)-\Pi_N(d))/\partial\theta}$. Since $\frac{\partial(\Pi_C(d,\theta)-\Pi_N(d))}{\partial\alpha} > 0$ and $\frac{\partial(\Pi_C(d,\theta)-\Pi_N(d))}{\partial\theta} < 0$, thus $\frac{\partial\theta_0}{\partial\alpha} > 0$. Therefore, $\hat{\theta}(s,\alpha)$ increases with α. This completes the proof.

Proof of Proposition 3

Based on the proof of Proposition 1, we have $\partial\Pi/\partial\lambda = R'_\lambda(\lambda,T) - \int_d^T F(t-d)R''_{d,\lambda}(\lambda,t)dt$, where $R''_{d,\lambda}(\lambda,t) = \frac{g(t,\lambda,n+1)n}{\lambda^2 t}\big((n-\lambda t)(T-t)\alpha\pi_R - t\pi_A\big)$. Taking the first and second derivative of $\partial\Pi/\partial\lambda$ against d gives $\frac{\partial^2\Pi}{\partial\lambda\partial d} = \int_d^T se^{-s(t-d)}R''_{d,\lambda}(\lambda,t)dt$ and $\frac{\partial^3\Pi}{\partial\lambda\partial d^2} = -sR''_{d,\lambda}(\lambda,d) + s\frac{\partial^2\Pi}{\partial\lambda\partial d}$.

Denote d_λ such that $\frac{\partial^2\Pi}{\partial\lambda\partial d} = 0$, we prove by contraction that $\frac{\partial^3\Pi}{\partial\lambda\partial d^2} = -sR''_{d,\lambda}(\lambda,d_\lambda) < 0$. Suppose $\frac{\partial^3\Pi}{\partial\lambda\partial d^2} \geq 0$, it implies $R''_{d,\lambda}(\lambda,d_\lambda) \leq 0$. Hence, $(n-\lambda d_\lambda)(T-d_\lambda)\alpha\pi_R - d_\lambda\pi_A \leq 0$. Thus, we have $(n-\lambda t)(T-t)\alpha\pi_R - t\pi_A \leq 0$ for all $t \in [d_\lambda, T]$. Therefore, $R''_{d,\lambda}(\lambda,t) \leq 0$ for all $t \in [d_\lambda,T]$. Hence, $\frac{\partial^2\Pi}{\partial\lambda\partial d} < 0$, which contradicts that $\frac{\partial^2\Pi}{\partial\lambda\partial d} = 0$. Hence, if $d \geq d_\lambda$, we have $\frac{\partial^2\Pi}{\partial\lambda\partial d} \leq 0$; and if $d < d_\lambda$, we have $\frac{\partial^2\Pi}{\partial\lambda\partial d} > 0$. Based on the derivative of implicit function, $\frac{\partial\bar{\lambda}}{\partial d} = -\frac{\partial^2\Pi}{\partial\lambda\partial d}/\frac{\partial^2\Pi}{\partial\lambda^2}$. By definition, $\frac{\partial^2\Pi}{\partial\lambda^2} < 0$ at $\bar{\lambda}$. Hence, $\frac{\partial\bar{\lambda}}{\partial d} < 0$ for $d \geq d_\lambda$; and $\frac{\partial\bar{\lambda}}{\partial d} > 0$ for $d < d_\lambda$.

Next, we prove that $d_\lambda(\theta)$ increases with θ. Based on derivative of implicit function, we have $\frac{\partial d_\lambda(\theta)}{\partial\theta} = -\frac{\frac{\partial^3\Pi}{\partial\lambda\partial d\partial\theta}}{\frac{\partial^3\Pi}{\partial\lambda\partial d^2}} = -\frac{\frac{\partial^3\Pi}{\partial\lambda^2\partial d}\frac{\partial\bar{\lambda}}{\partial\theta}}{\frac{\partial^3\Pi}{\partial\lambda\partial d^2}}$. Because we proved that $\frac{\partial^3\Pi}{\partial\lambda\partial d^2}\big|_{\bar{\lambda},d_\lambda(\theta)} < 0$ and $\frac{\partial\bar{\lambda}}{\partial\theta} < 0$, it suffices to prove that $\frac{\partial^3\Pi}{\partial\lambda^2\partial d}\big|_{\bar{\lambda},d_\lambda(\theta)} < 0$. Based on the proof of Lemma 1, we have $\frac{\partial^2\Pi}{\partial\lambda\partial d} = -\frac{d^2}{n-1}g(d,\lambda,n-1)(T-d)\alpha\pi_R s + s\int_d^T \frac{t^2}{n-1}g(t,\lambda,n-1)(e^{-s(t-d)}(\alpha\pi_R - \pi_A + (T-t)\alpha\pi_R s)dt$. Taking derivative of $\frac{\partial^2\Pi}{\partial\lambda\partial d}$ against λ gives $\frac{\partial^3\Pi}{\partial\lambda^2\partial d} = -\frac{d^2}{n-1}\frac{n-1-\lambda d}{\lambda}g(d,\lambda,n-1)(T-d)\alpha\pi_R s + s\int_d^T \frac{t^2}{n-1}\frac{n-1-\lambda t}{\lambda}g(t,\lambda,n-1)(e^{-s(t-d)}(\alpha\pi_R - \pi_A + (T-t)\alpha\pi_R s)dt$. Substituting $\frac{\partial^2\Pi}{\partial\lambda\partial d}\big|_{\bar{\lambda},d_\lambda(\theta)} = 0$ into $\frac{\partial^3\Pi}{\partial\lambda^2\partial d}$ gives $\frac{\partial^3\Pi}{\partial\lambda^2\partial d} = s\int_d^T(d-t)\frac{t^2}{n-1}g(t,\lambda,n-1)(e^{-s(t-d)}(\alpha\pi_R - \pi_A + (T-t)\alpha\pi_R s)dt < 0$. Hence, $\frac{\partial d_\lambda(\theta)}{\partial\theta} > 0$. This completes the proof.

Proof of Proposition 4

Given $q_i(p) = a_i - b_i p$, $i = A, R$, and $\pi_i = \max_q q_i(p)p$, we can solve $p_i^* = \frac{a_i}{2b_i}$, $q_i^* = \frac{a_i}{2}$ and $\pi_i = \frac{a_i^2}{4b_i}$. Meanwhile, $\mu_i = \int_0^{q_i^*}(p_i(q) - p_i^*) = \frac{a_i^2}{8b_i} = \frac{\pi_i}{2}$. Thus, patient welfare is $\nu(t,\tau;d) = \frac{1}{2}\phi(t,\tau) - w(1-\alpha)\min(\tilde{d},t)$. Hence, if $\lambda = \bar{\lambda}$, expected patient welfare is $U(d) = E_{t,\tau}\nu(t,\tau;d) = \frac{1}{2}\big(\Pi_C(d,\theta) + C(\bar{\lambda},\theta)\big) - w(1-\alpha)E_{t,\tau}(\min(\tilde{d},t))$.

We next establish the single-crossing condition. The marginal rate of substitution $\frac{dF}{dd} = \frac{d\Pi_C}{dd} = \frac{\partial\Pi}{\partial d}(\bar{\lambda},d,\theta)$. And $\frac{\partial}{\partial\theta}(\frac{dF}{dd}) = \frac{\partial^2\Pi}{\partial d\partial\lambda}\frac{\partial\bar{\lambda}}{\partial\theta}$. Based on Proposition 3, when $d \geq d_\lambda(\theta)$, we have $\frac{\partial^2\Pi}{\partial d\partial\lambda} \leq 0$ and thus $\frac{\partial}{\partial\theta}(\frac{dF}{dd}) \geq 0$. Hence, the single-crossing condition holds if $d \geq d_\lambda(\theta)$. Depending on θ, we discuss two cases.

Case 1. We solve the optimal $(d^{AI}(\theta), F^{AI}(\theta))$ for $\theta \in [\underline{\theta}, \hat{\theta}]$. Based on Fudenberg and Tirole (1991), for a given $d(\theta)$, the corresponding $F(\theta)$ is $F(\theta) = \Pi_C(d(\theta),\theta) - \Pi_0 - \int_\theta^{\hat{\theta}}\bar{\lambda}^k\big(d(\tilde{\theta});\tilde{\theta},s\big)d\tilde{\theta}$.

According to Proposition 1, the compliance constraint is equivalent to $d_1(\theta) \le d(\theta) \le d_2(\theta)$. Since $d(\theta)$ must be non-decreasing, we thus can obtain a more binding constraint $d_1(\theta) \le d(\theta) \le d_2(\hat{\theta})$. Additionally, we have $d \ge d_\lambda(\theta)$ to satisfy the single-crossing condition. Substituting $F(\theta)$ into regulator's objective and changing the order of integral gives

$$\begin{aligned} &\max_{d(\theta)} \int_{\underline{\theta}}^{\hat{\theta}} [U(d) + \Pi_C(d,\theta) - \Pi_0 - \bar{\lambda}^k\big(d(\theta);\theta,s\big)\frac{\Psi(\theta)}{\psi(\theta)}]\psi(\theta)d\theta \\ &\text{s.t. } d(\theta) \ge d_\lambda(\theta), \\ &\qquad d_1(\theta) \le d(\theta) \le d_2(\hat{\theta}). \end{aligned}$$

According to Envelope Theorem, we have $U_d' = \frac{1}{2}\frac{\partial \Pi}{\partial d}\big(\bar{\lambda}, d, \theta\big) + \frac{C_\lambda'}{2}\frac{\partial \bar{\lambda}}{\partial d} - w(1-\alpha)\frac{\partial E_{t,\tau}(\min(\tilde{d},t))}{\partial d}$. Suppose the constraints are not biding, then the optimal $d^{AI} = d^*(\theta)$, where $d^*(\theta)$ solves the following first-order condition $\frac{3}{2}\frac{\partial \Pi}{\partial d}\big(\bar{\lambda}, d, \theta\big) + \frac{C_\lambda'}{2}\frac{\partial \bar{\lambda}}{\partial d} - w(1-\alpha)\frac{\partial E_{t,\tau}(\min(\tilde{d},t))}{\partial d} - k\lambda^{k-1}\frac{\partial \bar{\lambda}}{\partial d}\frac{\Psi(\theta)}{\psi(\theta)} = 0$. Taking into the constraints into consideration, the optimal deadline is $d^{AI}(\theta) = \min\big(d_2(\hat{\theta}), \max\big(d^*(\theta), d_1(\theta), d_\lambda(\theta)\big)\big)$.

Case 2: We next solve the optimal $(\bar{d}^{AI}, \bar{F}^{AI})$ for $\theta > \hat{\theta}$. We first study the incentive compatible constraint associated with $(\bar{d}^{AI}, \bar{F}^{AI})$. Given the menu from Case 1, $(\bar{d}^{AI}, \bar{F}^{AI})$ must ensure that (i) manufacturer with $\theta > \hat{\theta}$ with prefer $(\bar{d}^{AI}, \bar{F}^{AI})$ over $(d^{AI}(\theta), F^{AI}(\theta))$; and (ii) manufacturer with $\theta < \hat{\theta}$ will prefer $(d^{AI}(\theta), F^{AI}(\theta))$ over $(\bar{d}^{AI}, \bar{F}^{AI})$.

We first prove that manufacturer with $\theta > \hat{\theta}$ will prefer $(d^{AI}(\hat{\theta}), F^{AI}(\hat{\theta}))$ over $(d^{AI}(\theta), F^{AI}(\theta))$ for all $\theta < \hat{\theta}$. To prove that, it suffices to show that $\Pi_N(d^{AI}(\hat{\theta})) - F^{AI}(\hat{\theta}) \ge \Pi_N(d^{AI}(\theta)) - F^{AI}(\theta)$ for all $\theta < \hat{\theta}$. According to the proof of Case 1, we have $\Pi_N(d^{AI}(\hat{\theta})) = \Pi_C(d^{AI}(\hat{\theta}), \hat{\theta})$. Additionally, we have $\Pi_C(d^{AI}(\hat{\theta}), \hat{\theta}) - F^{AI}(\hat{\theta}) = \Pi_0$. Hence, we have $\Pi_N(d^{AI}(\hat{\theta})) - F^{AI}(\hat{\theta}) = \Pi_C(d^{AI}(\hat{\theta}), \hat{\theta}) - F^{AI}(\hat{\theta}) = \Pi_0$. For $\theta < \hat{\theta}$, we have $\Pi_N(d^{AI}(\theta)) - F^{AI}(\theta) = \Pi_N(d^{AI}(\theta)) - \Pi_C(d^{AI}(\theta), \hat{\theta}) + \Pi_C(d^{AI}(\theta), \hat{\theta}) - F^{AI}(\theta)$. By the incentive compatible constraint, we have $\Pi_C(d^{AI}(\theta), \hat{\theta}) - F^{AI}(\theta) \le \Pi_C(d^{AI}(\hat{\theta}), \hat{\theta}) - F^{AI}(\hat{\theta}) = \Pi_0$. By the compliance constraint, we have $\Pi_N(d^{AI}(\theta)) - \Pi_C(d^{AI}(\theta), \hat{\theta}) < 0$. Thus, $\Pi_N(d^{AI}(\theta)) - F^{AI}(\theta) < \Pi_0$. Therefore, we have shown that $\Pi_N(d^{AI}(\hat{\theta})) - F^{AI}(\hat{\theta}) > \Pi_N(d^{AI}(\theta)) - F^{AI}(\theta)$ for all $\theta < \hat{\theta}$. Hence, to be incentive compatible, we must have $\Pi_N(\bar{d}^{AI}) - \bar{F}^{AI} = \Pi_0$ such that a manufacturer with $\theta > \hat{\theta}$ is actually indifferent between $(\bar{d}^{AI}, \bar{F}^{AI})$ and $(d^{AI}(\hat{\theta}), F^{AI}(\hat{\theta}))$.

We next prove that, given $\Pi_N(\bar{d}^{AI}) - \bar{F}^{AI} = \Pi_0$, manufacturer with $\theta < \hat{\theta}$ will prefer $(d^{AI}(\theta), F^{AI}(\theta))$ over $(\bar{d}^{AI}, \bar{F}^{AI})$. We essentially have to show that $\Pi_C(d^{AI}(\theta), \theta) - F^{AI}(\theta) \ge \Pi_C(\bar{d}^{AI}, \theta) - \bar{F}^{AI}$ for all $\theta < \hat{\theta}$. By the incentive compatible constraint, we have $\Pi_C(d^{AI}(\theta), \theta) - F^{AI}(\theta) \ge \Pi_C(d^{AI}(\hat{\theta}), \theta) - F^{AI}(\hat{\theta})$. Therefore, it suffices to prove $\Pi_C(d^{AI}(\hat{\theta}), \theta) - F^{AI}(\hat{\theta}) \ge \Pi_C(\bar{d}^{AI}, \theta) - \bar{F}^{AI}$. Since $F^{AI}(\hat{\theta}) = \Pi_C(d^{AI}(\hat{\theta}), \hat{\theta}) - \Pi_0$ and $\bar{F}^{AI} = \Pi_N(\bar{d}^{AI}) - \Pi_0$ and $\Pi_C(d^{AI}(\hat{\theta}), \hat{\theta}) = \Pi_N(d^{AI}(\hat{\theta}))$, thus, it suffices to prove that $\Pi_C(d^{AI}(\hat{\theta}), \theta) - \Pi_N(d^{AI}(\hat{\theta})) \ge \Pi_C(\bar{d}^{AI}, \theta) - \Pi_N(\bar{d}^{AI})$.

As defined in the proof of Proposition 1, $J(d;\theta) = \Pi_C(d,\theta) - \Pi_N(d)$. Hence, we essentially must prove $J(d^{AI}(\hat{\theta});\theta) \ge J(\bar{d}^{AI};\theta)$. The proof of Proposition 1 suggests that there exists d_0 such that $J^{'}(d_0) = 0$ and $J(d)$ increases with d if $d > d_0$. By definition, we have $d_0(\theta) = d^{AI}(\hat{\theta})$ when $\theta = \hat{\theta}$. We next prove that $d_0(\theta) < d^{AI}(\hat{\theta})$ when $\theta < \hat{\theta}$. Based on the derivative of implicit function, we have $\frac{\partial d_0}{\partial \theta} = -\frac{\partial^2 J/\partial d \partial \theta}{\partial^2 J/\partial d^2}$. Since $\frac{\partial^2 J}{\partial d \partial \theta} = \frac{\partial^2 J}{\partial \theta \partial d} = -k\bar{\lambda}^{k-1}\frac{\partial \bar{\lambda}}{\partial d} > 0$ and $\frac{\partial^2 J}{\partial d^2} < 0$, we have $\frac{\partial d_0}{\partial \theta} > 0$. Therefore, $d_0(\theta) < d^{AI}(\hat{\theta}) \leq \bar{d}^{AI}$ for $\theta < \hat{\theta}$. Thus, we have $J(d^{AI}(\hat{\theta}); \theta) \geq J(\bar{d}; \theta)$. Together, we proved that $(\bar{d}^{AI}, \bar{F}^{AI})$ specified by $\Pi_N(\bar{d}^{AI}) - \bar{F}^{AI} = \Pi_0$ is incentive compatible.

Hence, for $\theta > \hat{\theta}$, regulator solves the following problem to determine $(\bar{d}^{AI}, \bar{F}^{AI})$:

$$
\begin{aligned}
\max \;\; & U(\bar{d}^{AI}, \bar{F}^{AI}) = \frac{1}{2}\Pi_N(\bar{d}^{AI}) - w(1-\alpha)(\bar{d}^{AI} + 1/s) + \bar{F}^{AI} \\
\text{s.t.} \;\; & \bar{F}^{AI} = \Pi_N(\bar{d}^{AI}) - \Pi_0 \\
& \bar{d}^{AI} \geq d^{AI}(\hat{\theta}),
\end{aligned}
$$

Thus, we have $\bar{d}^{AI} = \max(d^{AI}(\hat{\theta}), \frac{1}{s}\ln(1 - \frac{2w(1-\alpha)}{3\pi_A}) + T)$. This completes the proof.

Proof of Proposition 5

If regulator knew manufacturer's cost type θ, it then solves the following problem,

$$
\begin{aligned}
\max_{d} \;\; U(d) = \;& \frac{1}{2}\big(\Pi_C(d,\theta) + C(\bar{\lambda},\theta)\big) - w(1-\alpha)E_{t,\tau}(\min(\tilde{d}, t)) \\
\text{s.t.} \;\; & d_1(\theta) \leq d(\theta) \leq d_2(\theta).
\end{aligned}
$$

Taking derivative of $U(d)$ against d gives $U^{'}_d = \frac{1}{2}\frac{\partial \Pi}{\partial d}(\bar{\lambda}, d, \theta) + \frac{C^{'}_{\lambda}}{2}\frac{\partial \bar{\lambda}}{\partial d} - w(1-\alpha)\frac{\partial E_{t,\tau}(\min(\tilde{d},t))}{\partial d}$. Let $d^0(\theta)$ solve the above first-older condition. Given the compliance constraint, the optimal $d^{CI}(\theta)$ is then $d^{CI}(\theta) = \min\big(d_2(\theta), \max(d^0(\theta), d_1(\theta))\big)$. This completes the proof.

Proof of Corollary 1

According to the proof of Proposition 4, by definition, $d^*(\theta)$ solves $U^{'}_d + \frac{\partial \Pi}{\partial d}\big(\bar{\lambda}, d, \theta\big) - k\bar{\lambda}^{k-1}\frac{\partial \bar{\lambda}}{\partial d}\frac{\Psi(\theta)}{\psi(\theta)} = 0$. And according to the proof of Proposition 5, $d^0(\theta)$ solves $U^{'}_d = 0$. Since $\frac{\partial \Pi}{\partial d}(\lambda, d, \theta) = \int_0^\infty \Phi^{'}_d(t) g(t, \lambda, n) dt$ and $\Phi^{'}_d(t)$ is

$$
\Phi^{'}_d(t) = \begin{cases} 0 & \text{if } t \leq d, \\ (1 - e^{-s(t-d)})\pi_A + (T-t)s\alpha\pi_R e^{-s(t-d)} & \text{if } d < t \leq T, \\ (1 - e^{-s(T-d)})\pi_A & \text{if } t > T. \end{cases}
$$

It follows directly that $\Phi^{'}_d(t) > 0$ and, thus, $\frac{\partial \Pi}{\partial d}(\bar{\lambda}, d, \theta) > 0$. If $d^*(\theta) \geq d_\lambda(\theta)$, according to Proposition 3, we have $\frac{\partial \bar{\lambda}}{\partial d} < 0$. Together, we have $U^{'}_d + \frac{\partial \Pi}{\partial d}\big(\bar{\lambda}, d, \theta\big) - k\bar{\lambda}^{k-1}\frac{\partial \bar{\lambda}}{\partial d}\frac{\Psi(\theta)}{\psi(\theta)} > U^{'}_d$. Therefore, we have $d^*(\theta) > d^0(\theta)$. This completes the proof.

Proof of Proposition 6

For a given λ, $\Pi(\lambda,\theta)=\int_0^T \big(t\pi_A+(T-t)\alpha\pi_R\big)g(t,\lambda,n)dt+\int_T^\infty T\pi_A g(t,\lambda,n)dt-C(\lambda,\theta)$. Regulator's expected patient welfare is then $U(\lambda)=\frac{1}{2}\big(\Pi(\lambda,\theta)+C(\lambda,\theta)\big)-w(1-\alpha)\big(T-\int_0^T G(t,\lambda,n)dt\big)$.

We then verify the single-crossing condition. The marginal rate of substitution is $\frac{dF}{d\lambda}=\frac{\partial\Pi}{\partial\lambda}$. Therefore, we have $\frac{\partial}{\partial\theta}(\frac{dF}{d\lambda})=\frac{\partial^2\Pi}{\partial\lambda\partial\theta}=-k\lambda^{k-1}$. Thus, the $\frac{dF}{d\lambda}$ is monotone in θ and the single-crossing condition holds. Hence, based on Fudenberg and Tirole (1991), for a given $\lambda(\theta)$, the corresponding $F(\theta)$ is $F(\theta)=\Pi(\lambda(\theta),\theta)-\Pi_0-\int_\theta^{\bar\theta}\lambda^k(\tilde\theta)d\tilde\theta$. Thus, the participation constraint will not be biding. Substituting $F(\theta)$ into regulator's objective and changing the order of integral, we have

$$\begin{aligned}&\max_{d(\theta)}\int_{\underline{\theta}}^{\bar\theta}[U(\lambda(\theta))+\Pi(\lambda(\theta),\theta)-\Pi_0-\lambda^k(\theta)\frac{\Psi(\theta)}{\psi(\theta)}]\psi(\theta)d\theta\\&\text{s.t. }\Pi(\lambda(\theta))\geq\Pi_0\end{aligned}$$

It follows directly that $\frac{\partial U}{\partial\lambda}=\frac{1}{2}(\frac{\partial\Pi}{\partial\lambda}+\theta k\lambda^{k-1})+\frac{n}{\lambda^2}w(1-\alpha)G(T,\lambda,n+1)$ and $\frac{\partial\Pi}{\partial\lambda}=\frac{n}{\lambda^2}G(T,\lambda,n+1)(\alpha\pi_R-\pi_A)-\theta k\lambda^{k-1}$. The optimal $\lambda^{VE}(\theta)$ solves the following first-order condition $\big(\frac{3}{2}(\alpha\pi_R-\pi_A)+w(1-\alpha)\big)\frac{n}{\lambda^2}G(T,\lambda,n+1)-k\lambda^{k-1}(\theta+\frac{\Psi(\theta)}{\psi(\theta)})=0$. This completes the proof.

Proof of Proposition 8

The central planner solves $\lambda^{cs}(\theta_R)$ to maximize $U(\lambda)+\Pi(\lambda,\theta_R)$. Based on the proof of Proposition 6, we have $\frac{\partial U}{\partial\lambda}=\frac{1}{2}(\frac{\partial\Pi}{\partial\lambda}+\theta_R k\lambda^{k-1})+\frac{n}{\lambda^2}w(1-\alpha)G(T,\lambda,n+1)$, and $\frac{\partial\Pi}{\partial\lambda}=\frac{n}{\lambda^2}G(T,\lambda,n+1)(\alpha\pi_R-\pi_A)-\theta_R k\lambda^{k-1}$. Thus, $\lambda^{cs}(\theta_R)$ solves the following first-order condition $\big(\frac{3}{2}(\alpha\pi_R-\pi_A)+w(1-\alpha)\big)\frac{n}{\lambda^2}G(T,\lambda,n+1)=\theta_R k\lambda^{k-1}$. This completes the proof.

Proof of Proposition 9

For the case in which information regarding α is asymmetric, but information regarding θ is not, the regulator would design a deadline-dependent user fee $(d(\alpha),F(\alpha))$ as follow,

$$\begin{aligned}\max_{d(\alpha),F(\alpha)}&\int_{\underline{\alpha}}^{\bar\alpha}[U(d(\alpha),\alpha)+F(\alpha)]\psi(\alpha)d\alpha\\\text{s.t. }&\Pi_C(d(\alpha),\alpha)-F(\alpha)\geq\Pi_C(d(\tilde\alpha),\alpha)-F(\tilde\alpha),\forall\tilde\alpha\in[\underline{\alpha},\bar\alpha]\\&\Pi_C(d(\alpha),\alpha)-F(\alpha)\geq\Pi_0,\\&\Pi_C(d(\alpha),\alpha)\geq\Pi_N(d(\alpha)),\forall\theta\leq\hat\theta.\end{aligned}$$

where $\psi(\alpha)$ is the distribution of α that describes the regulator's belief about the drug's success probability. The first constraint is the incentive compatibility constraint. The second constraint ensures the manufacturer's participation and the third ensures that the deadline creates incentive for compliance.

We next establish property of the optimal menu of deadline and user fee by analyzing the manufacturer's iso-profit curve, which is defined by $\Pi_C(d,\alpha) - F = \Pi_0$. Thus, the marginal rate of substitution between deadline and user fee is

$$\frac{dF}{dd} = \frac{d\Pi_C}{dd} = \frac{\partial \Pi}{\partial d}(\bar{\lambda}, d, \alpha) = \int_d^T f(t-d)\big(\pi_A(1 - G(t, \bar{\lambda}, n)) + (T-t)\alpha\pi_R g(t, \bar{\lambda}, n)\big)dt.$$

Thus, we have

$$\frac{\partial}{\partial \alpha}(\frac{dF}{dd}) = \frac{\partial^2 \Pi}{\partial d \partial \lambda}\frac{\partial \bar{\lambda}}{\partial \alpha} + \int_d^T f(t-d)(T-t)\pi_R g(t, \bar{\lambda}, n)dt$$

Based on Proposition 3, we have showed that $\frac{\partial^2 \Pi}{\partial d \partial \lambda} \leq 0$ and $\frac{\partial \bar{\lambda}}{\partial \alpha} > 0$. Hence, we must have $\frac{\partial}{\partial \alpha}(\frac{dF}{dd}) < 0$ when evaluating $d = T$. Therefore, the single-crossing condition, which requires $\frac{dF}{dd}$ is monotone in α, enforces that $\frac{\partial}{\partial \alpha}(\frac{dF}{dd}) < 0$. Thus, according to Fudenberg and Tirole (1991), under the optimal deadline menu, $d(\alpha)$ should decrease with α.