False Diagnosis and Overtreatment in Services

M. Fazıl Paç* Senthil Veeraraghavan†

May 2015

Abstract

In many services, consumers must rely on experts to identify the type of service they need. In such expert services, diagnosis is a crucial step in which the expert identifies the type of problem and also provides the corresponding treatment. The asymmetry of information between the expert and the consumer leads to inefficiencies in the form of over-treatment: The expert may have an incentive to not reveal the true diagnosis to the consumer in order to sell more expensive services. The expert may extract more revenue through the expensive treatment, but providing such a treatment also requires more service capacity and time. Hence, overtreatments impose longer delays and higher waiting costs for consumers. We find that congestion acts as a natural "fraud cost" and mitigates expert cheating and induces honesty, thus increasing social value. Experts with high capacity utilization are less prone to overtreat. To commit to an honest diagnosis, the expert has to charge high prices and limit demand for his services. Thus, low prices may serve as a signal for overtreatment.

Keywords: Queueing Games, Diagnosis, Overtreatment, Honesty, Service Specialization, Social Value.

1 Introduction

In a wide variety of services, consumers cannot self-diagnose their problems; hence, they cannot identify the type of service that will address their needs. As a result, they rely on knowledgeable experts for diagnosis, who are also often the providers of those service. These services are often referred to as *expert services*. Medical services, repair services, legal advice, and consulting services all fall into the broad category of expert services, in which a diagnosis precedes the actual provision of the service. In these settings, the experts are better informed about the potential problem types and the services required to treat those problems. Upon diagnosing and identifying the problems, the experts gain an informational advantage over the consumers.

*fazilpac@gmail.com, Google Inc. Mountain View, CA. 94043

†senthilv@wharton.upenn.edu. The Wharton School, University of Pennsylvania, Philadelphia, PA. 19104. Corresponding Author.

In recommending a service, an expert may not reveal the true diagnosis and may exploit his informational advantage in order to sell more services (that a consumer may not need). As a consequence, *even after consumption*, consumers are often unsure about the necessity of the recommended service, and its true benefit. Darby and Karni (1973) first introduced the term *credence good*, to refer to goods whose impact on consumers' utility is not completely revealed even after consumption.

Since the experts who diagnose are also often the providers of the associated service, there are incentives to provide false recommendations, leading to inefficiencies. The presence of aggressive diagnosis and overtreatment in services has been of significant interest to firms, consumers, policy makers and society as a whole.

In particular, in the health care industry, such incentives in diagnosis and treatments have received enormous recent interest in both academic literature and press. In the book *Overtreated*, Brownlee (2011) argues how several billion dollars of unnecessary tests and procedures are ordered in the United States due to perverse incentives. In a recent article, Gawande (2015) discusses how paying physicians for the quantity of care leads to pervasive overtreatment and unnecessary care. Welch and Passow (2014) estimate that 22% of all screen-detected invasive breast cancers are overdiagnosed. Sometimes, aggressive (rather, than false) diagnosis also leads to overtreatment. In a recent paper, Ong and Mandl (2015) estimate that such overdiagnoses of invasive breast cancer and ductal carcinoma in-situ (DCIS) cost the US health care system $1.2 billion annually. Based on 30 years of data, medical research (Bleyer and Welch 2012, Welch and Black 2010) argues that aggressive screening for mammography is associated with the harm of overtreatment caused by both false-positive mammograms and breast cancer overdiagnosis. In fact, the US Preventive Services Task Force has recommended against routine mammography screening for women ages 40–49 (USPSTF, 2009).

Such overtreatment concerns have been documented in other industries/instances of expert services. Gruber and Owings (1996) note a significant rise in cesarean delivery rates mainly due to high reimbursement pressures. Schneider (2012) estimates that unnecessary auto-repairs were recommended in 27% of all visits, using a field experiment. Also see Wennberg et al. (1982), Dranove (1988), Gruber and Owings (1996), and Delattre and Dormont (2003) for the evidence of physician-induced demand in various medical settings.

The extant academic research in operations literature has not addressed the impact of the capacity and service delays on expert diagnosis and overtreatment issues. Typically, when experts direct consumers toward more expensive services to extract higher revenues, the action comes at a cost to the society, and sometimes to the expert. Overtreatment creates a greater requirement of capacity and a tighter resource utilization. Often, the expensive services require more time which is a valuable resource for both experts and consumers. Further, longer service times result in long waiting times, which may deter some consumers from procuring the service from a service

provider. In this paper, we focus on an expert's pricing and diagnostic decisions in a market with with self-interested delay-sensitive consumers.

In services where waiting is required, the experts' diagnostic strategy and the aggregate procurement decisions of all consumers in the market may increase waiting for all other consumers. Therefore, even though overtreatment may increase revenues for the service provider, it also intensifies the congestion in the system, as the expert may provide consumers with longer services that the consumers do not need. We investigate how capacity and congestion concerns may mitigate expert cheating (service overprovision) and increase the efficiency of service provision.

We characterize the role of service capacity in welfare maximizing operations. In particular, we identify the market conditions for specialization and prioritization of problems. We show that, due to congestion considerations, it might be socially optimal to specialize in treating "minor" problems.

We show that honest price discrimination and socially efficient service can indeed occur and how they are driven by driven by workload considerations. For the expert to achieve efficient service provision, the demand for his services has to be large enough. An expert with excess capacity is more likely to cheat, as congestion costs are diminished under low utilization. As a result, service over-provision becomes less costly.

If the service value from the different treatment/service plans are disparate, it becomes harder for the service provider to signal honest diagnosis. In such cases, service over-provision becomes an unavoidable feature of the market.

2 Related Literature

Our work builds on the literature on (i) queuing games, specifically service value and time dependency literature, and (ii) the economics of credence goods in the context of queueing operations.

The research on credence goods originates from Darby and Karni (1973), who study the impact of market conditions on the equilibrium level of fraud in markets where consumers' utility has not been revealed even after consumption. Several papers (Wolinsky (1993), Pesendorfer and Wolinsky (2003)) note that overtreatments continue to occur in competitive settings. Fong (2005) finds that expert cheating arises as a substitute for price discrimination, and experts cheat on high valuation, high cost customers. Dulleck and Kerschbamer (2006) find that liability and verifiability are important institutional factors determining experts' behavior, while reputation and competition are important market factors.

Nevertheless, very little is known about the impact of resource constraints (service capacity, service delays) on experts' diagnostic decisions to overtreat. Emons (1997) in a deterministic model, finds that experts earn positive revenues through honesty only if the market demand exceeds market capacity. Emons (2001) shows that an expert can signal an honest diagnosis either by setting the

capacity exactly equal to the market demand or by charging a fixed admission price. However, in many operational settings in which the arrivals and service times are stochastic, and service capacities are fixed, these findings do not hold. We contribute to the literature by investigating the impact of service delays, and negative externalities on expert diagnosis behavior.

In the credence literature stream, committing false diagnosis is typically costless. For an exception, see Alger and Salanie (2006) where the fraud is costly but exogenous. In contrast, in our model, the fraud costs emerge endogeneously through waiting costs. Pitchik and Schotter (1987) find that experts become more honest as the price gap between major and minor repairs decreases. We find that congestion effects alter the finding. In congested stochastic environments, prices serve two different but (occasionally) contradictory purposes: they signal the credibility of expert's diagnosis, and they control congestion in the service system.We show that congestion tempers cheating behavior.

The research in Operations Management on credence services is nascent. The effect of workload on cheating/service inducement is discussed in Glazer and Hassin (1983). Debo et al (2008) studies how queue lengths to "service inducement" in a monopolistic market. The service provider can exploit the differences in the state of arrival, to *pad* additional services and extract higher revenues whenever the server is idle. We show how cheating can emerge even without state dependencies, and even with procedural verification, when the consumer is aware that a process was truly done.

Finally, our paper contributes to the recent renewed interest in queueing games (beginning with Naor 1969; see survey by Hassin and Haviv 2004). Specifically, we focus on the recent research in queues where the service value increases in service time. This service value-time relationship has been consistently observed in health care settings and diagnostic services: For instance, Alizamir et al (2013) and Wang et al (2009) focus on diagnostic services in congested system where the outcome quality improves with every additional diagnostic step (that adds to overall service time). We refer the reader to significant advances in different settings made in research by Alizamir et al 2013 (Diagnostic accuracy), Anand et al 2011 (Tradeoffs in static settings), Dai et al 2012 (Medical insurance pricing), Hasija et al 2009 (Work expansion), Lee et al 2012 (Gatekeeping and expert referrals), and Kostami and Rajgopalan 2014 (Tradeoffs in dynamic settings). This stream of literature has not examined cheating and overtreatment in services.

We find that such quality-service time dependency is critically inherent to expert services, and is critical to the understanding the drivers of overtreatment in expert services. For instance, we show how the social welfare planner may prioritize low-value minor services due to time dependencies of such services. To our knowledge, very little is known about the operational impact of service capacity and service value-time dependencies, on the diagnostic decisions, overtreatment and pricing in such expert services. Our paper attempts to fill this gap.

3 Model of Expert Service

Consumer Information. We consider an expert providing service to a market of self-interested, risk-neutral and utility maximizing consumers. Each consumer in the market has a problem that requires treatment. We assume that consumers' problems fall into two categories: major problems (type M problems), which occur with probability $\theta \in (0,1)$ in the population, and minor problems (type m), which occur with probability $\bar{\theta} \triangleq 1-\theta$. The presence of both types of problems is recognized by the same symptoms. As a result, the consumer *cannot* identify the type of problem himself.

Since consumers can observe the existence of *a* problem through exhibited symptoms, but cannot diagnose the underlying type of problem, they also cannot identify the proper service that will resolve their problem. Therefore, for both diagnosis and treatment, they rely on a knowledgeable service provider, whom we henceforth simply address as the *expert*. The resolution of a minor problem provides a consumer a value of v_L, while the resolution of a major problem provides a value of $v_H (> v_L)$.[1]

The consumers are ex ante homogeneous, i.e., they have the same prior on the incidence of the problem θ. Thus, the ex ante expected value from the treatment of their problem is $v \triangleq \theta v_H + (1-\theta) v_L$. However, upon diagnosis by the expert, they will have different valuations and updated beliefs based on the service recommendations that they receive from the expert. Specifically a consumer with a major problem diagnosis would have higher value for the problem to be fixed.

Consumer Arrival, Diagnosis and Service Recommendation. Consumers arrive to the server according to a Poisson process at an exogenously specified mean rate of Λ, which we refer to as the *potential demand* in market, or simply as *market size*. Upon the consumer's arrival, the expert performs an immediate diagnosis[2], and learns (i.e., identifies to himself) the problem type. While the expert's (internal) diagnosis is always accurate, his service recommendation, d, need not be. In other words, the expert does not necessarily reveal his true diagnosis to the consumer when he recommends a treatment. Thus, once the diagnosis is completed, the service provider has an informational advantage over the consumer.

For each consumer, following up on his diagnosis, the expert also offers service to treat the diagnosed problem. When waiting for service, each consumer incurs a waiting cost of c per unit of time spent in the system. The expert sets the price of the minor treatment at p_L and the price of the major treatment at p_H, which are both public information. The true diagnosis is not revealed.

[1] This setting is equivalent to a setting in which a consumer incurs a higher cost when major problem is left untreated compared to untreated minor problem.

[2] See related paper on gatekeeper diagnosis by Lee, Pinker and Shumsky (2012), where the diagnosis is also instantaneous. Minor problems (below a threshold complexity) are handled by the gatekeeper. More complex (and mistreated) problems are handled by the expert in that model. There is no incentive to overtreat in their model because mistreated and expert-retreated problems are costly.

The consumers, being rational, are aware of the expert's incentive for cheating.

There are economies of scope between the diagnosis and the service. For instance, providing the service might require the same equipment or facility used for the diagnosis. Therefore, as typically observed in practice, the same expert provides both the diagnosis *and* the service.

Service Value and Time Dependency. The expert recommends/provides one of the two types of services: a minor treatment L or a major treatment H. The service duration for each type of service is stochastic and exponentially distributed. For consumers with minor problems (m), the minor treatment of mean duration τ_L is sufficient to treat the problem. For consumers with major problems (M), a major treatment of mean duration τ_H is necessary to resolve the problem. The major treatment resolves both minor and major problems. Hence, its valuation is higher, i.e., $v_H > v_L$, but its duration is also longer, i.e., $\tau_H > \tau_L$.

We assume *diminishing marginal returns* in value accrued during service time, i.e., $v_H/\tau_H \leq v_L/\tau_L$. The diminishing marginal value assumption is consistent with quality-speed trade-off literature (see Anand et al 2011). Our conclusions continue to hold even when this assumption is relaxed. In fact, overtreatment becomes more likely if marginal value does not diminish in time.

We assume that the treatment times are exponentially distributed, but our conclusions continue to hold as long as the service times follow an independently and identically distributed general distribution.

Finally, the expert's services are *verifiable*; i.e., upon the completion of the service, consumers learn the type of service they received and can also verify whether their problems were treated (for example, it is easy to verify that a complicated additional tax document was completed and filed, or an MRI scan was done). Consumers pay the service fee only after their problems are treated.

After their problems are treated, the consumers may still not know *what type of problem they had* in the first place, and whether the provided service was indeed the appropriate service to treat their problem. In particular, a consumer who received a major treatment may have had a minor problem and there is no evidence to prove or disprove the need for the recommended procedure, after the fact. If the expert recommends minor treatments to fix major problems, the consumer problems remain unresolved, which forces the expert to lose resources (time) without accruing revenues.

Information Asymmetry and Honesty. On diagnosis, the expert gains information about the customer type, which the customer does not know himself. This information asymmetry arising upon the diagnosis gives the expert an opportunity to over-sell services by prescribing a major treatment to consumers with minor problems. Let $\alpha_i, i \in \{m, M\}$ be the probability that the expert recommends the major treatment H to a consumer with problem type i. Since consumers pay the service fee only on the resolution of their problem, the expert's services are *liable*. As minor

treatment does not treat major problems, its provision to consumers with major problems would not any yield any revenue. Hence, due to service liability, we have $\alpha_M = 1$. In other words, the expert does not gain from providing a minor service to a consumer with major problem. It does not resolve the consumer's problem, and no revenue is accrued.

Nevertheless, minor problem consumers may continue to be over-treated with probability α_m. We define an *honest* diagnostic strategy to the diagnostic strategy without any over-treatment, i.e., $\alpha_m = 0$. Henceforth, we refer to the proportion of consumers for whom major treatment is recommended, χ, as the expert's diagnostic decision variable. Clearly, $\chi = \theta + (1 - \theta)\alpha_m$. Also note that $\chi \geq \theta$ since $\alpha_M = 1$. Depending on the diagnosis strategy, consumers receiving the recommendation of major problem, may have different problem types.

Expert's Decisions. The expert chooses prices (p_L, p_H) for the treatments and adopts diagnostic strategy χ to maximize his revenues, based on consumer's utility maximizing queue joining decisions. We model the expert's revenue maximization problem as a two-stage game: First, the expert sets the prices (p_L, p_H) for his services and announces it to the market. Then, the expert and arriving consumers play the *diagnosis and queue joining game*. We first focus on the diagnosis and queueing subgame, and then the pricing game.

Consumer Decisions. On receiving the diagnosis ($d \in \{L, H\}$), every consumer decides whether to procure the prescribed treatment (i.e., join the service) or seek an alternate option (i.e., balk), based on the prices p_L and p_H, her belief about the service provider's diagnosis decision χ and the expected delay W. Note that a consumer only knows the service recommendation, d, and observes the posted prices p_L and p_H. However, the consumer can deduce the expert's diagnostic strategy and the resulting expected waiting time in equilibrium from the prices set by the provider and the market size.

Absent additional information, all consumers that are recommended the same treatment have the same updated beliefs. Consumers who get different recommendations will join with different probabilities. For sure, a consumer getting a major treatment recommendation, is concerned about the possibility of overtreatment, and hence, all else equal, less likely to join. Thus we focus on symmetric joining probabilities, *conditional* on the recommendation: $\beta_L \triangleq \beta(p_L, p_H|d = L)$ and $\beta_H \triangleq \beta(p_L, p_H|d = H)$.

Service Time Distribution. Consumers joining the queue after receiving a recommendation, will face different service times based on the recommendation they received. Therefore the ex ante expectation of the service rate of the system is endogenously determined by the expert's diagnostic strategy, χ, and consumers' service procurement decisions β_L and β_H. For a given strategy profile

(χ, β_L, β_H), the unconditional service time, τ, has the following hyper-exponential distribution:

$$\tau = \begin{cases} \text{Exponential with mean } \tau_L & \text{w.p. } \beta_L(1-\chi)/(\beta_L(1-\chi)+\beta_H\chi) \\ \text{Exponential with mean } \tau_H & \text{w.p. } \beta_H\chi/(\beta_L(1-\chi)+\beta_H\chi) \end{cases} \quad (1)$$

with mean and variance,

$$\begin{aligned} E[\tau] &= (\beta_L(1-\chi)\tau_L + \beta_H\chi\tau_H)/(\lambda_L+\lambda_H) \\ Var[\tau] &= \left((\chi\beta_H\tau_H - (1-\chi)\beta_L\tau_L)^2 + 2\chi(1-\chi)\beta_L\beta_H(\tau_L^2+\tau_H^2)\right)/(\beta_L(1-\chi)+\beta_H\chi)^2. \end{aligned}$$

Expected Waiting Time. The expert's diagnosis and pricing policy, and the consumer recommendation-dependent joining rates all affect the congestion in the system. The expected time in the system is the sum of expected time spent in the queue $W_q(\chi, \beta_L, \beta_H)$ and the time of service (τ_H or τ_L depending on the recommendation). The expected waiting times for consumers joining the service queue given the potential demand, Λ, and the strategy profile (χ, β_L, β_H), and diagnosis $d \in \{L, H\}$ are as follow:

$$W_L(\chi, \beta_L, \beta_H) = \tau_L + W_q(\chi, \beta_L, \beta_H) \quad (2)$$

$$W_H(\chi, \beta_L, \beta_H) = \tau_H + W_q(\chi, \beta_L, \beta_H) \quad (3)$$

$$\text{where } W_q(\chi, \beta_L, \beta_H) = \Lambda(\beta_L\tau_L^2(1-\chi) + \beta_H\tau_H^2\chi)/\left(1 - \Lambda(\beta_L\tau_L(1-\chi) + \beta_H\tau_H\chi)\right). \quad (4)$$

The above expressions apply if the stability condition $\Lambda(\beta_L\tau_L(1-\chi) + \beta_H\tau_H\chi) < 1$ holds. If the stability condition is not satisfied, the expected waiting time is infinite.[3]

3.1 Diagnosis and Queueing

The expert sets his diagnostic strategy, χ which is dependent on θ, Λ and prices (p_L, p_H). We suppress the dependencies on θ, Λ and price in the diagnostic strategy χ for the ease of exposition. Due to service liability, consumers with major problems (M) are recommended the major treatment, and therefore, $\chi \geq \theta$. On the other hand, consumers with minor problems (m) may be subject to overprovision. We focus on pure diagnostic strategies, i.e., $\chi \in \{\theta, 1\}$ for analytic tractability. Consumers do not observe the expert's diagnostic strategy χ, but deduce it through the prices (p_L, p_H) and Λ to make their queue joining decision.

Given the expert's diagnostic strategy χ, and the consumers symmetric queue joining strategy $\beta = (\beta_L, \beta_H)$, the effective demand for the minor treatment is $\lambda_L = \beta_L(1-\chi)\Lambda$, while the effective demand for the major treatment is $\lambda_H = \beta_H\chi\Lambda$.

[3]Equilibrium joining rates ensure stability, since the revenue falls as waiting times grow.

Consumers' Expected Payoff and Expert's Revenues. Given the prices (p_L, p_H) and the potential demand Λ, under the strategy profile (χ, β_L, β_H), the expected payoff of a consumer from joining the queue upon the recommendation d is:

$$V_d(\chi, \beta_L, \beta_H) = \begin{cases} v_L - p_L - cW_L(\chi, \beta_L, \beta_H) & \text{if } d = L \\ \\ v_H\theta/\chi + v_L(\chi - \theta)/\chi - p_H - cW_H(\chi, \beta_L, \beta_H) & \text{if } d = H, \end{cases} \tag{5}$$

while the expert's revenues can be written as:

$$R(p_L, p_H, \Lambda, \chi, \beta) = \Lambda\left(p_L(1-\chi)\beta_L + p_H\chi\beta_H\right). \tag{6}$$

An arriving customer will procure the recommended service (i.e., join the service queue) only if $V_d(\chi, \beta_L, \beta_H)$ is non-negative.

Equilibrium Conditions. Given the potential demand Λ, and the prices (p_L, p_H), under symmetric consumer strategies, the set of equilibria $\mathcal{E}(\Lambda, p_H, p_L) \subset \{\theta, 1\} \times [0,1] \times [0,1]$ of the diagnosis and queue joining game consists of strategy profiles $(\chi^e, \beta_L^e, \beta_H^e)$, which simultaneously satisfy the following set of conditions:

$$\beta_L^e = \max_{\{0 \le \beta_L \le 1\}} \left\{\beta_L | \beta_L V_d(\chi^e, \beta_L, \beta_H^e) \ge 0\right\}, \tag{7}$$

$$\beta_H^e = \max_{\{0 \le \beta_H \le 1\}} \left\{\beta_H | \beta_H V_d(\chi^e, \beta_L^e, \beta_H) \ge 0\right\}, \tag{8}$$

$$\chi^e = \arg\max_{\chi \in \{\theta, 1\}} \left\{\Lambda\left(p_L(1-\chi)\beta_L^e + p_H\chi\beta_H^e\right)\right\}. \tag{9}$$

Equations (7) and (8) guarantee that under the strategy profile $(\chi^e, \beta_L^e, \beta_H^e)$, a self-interested consumer will join the service queue with positive probability upon receiving recommendation $d \in \{L, H\}$, as long as there is non-negative value in doing so. The diagnosis and queue joining game may have multiple symmetric equilibria satisfying the above conditions, for a given set of prices (p_H, p_L).

However, as we will show in Section 5, under optimal prices, the equilibrium of the diagnosis and queue joining game is unique in symmetric consumer strategies (i.e., the set $\mathcal{E}(\Lambda, p_H, p_L)$ consists of a single strategy profile) in all but one scenario. In that particular scenario we focus on the consumer equilibrium, that maximizes the expert's revenues.

Given the prices (p_L, p_H) and Λ, the equilibrium demands for the minor and major treatments is $\lambda_L^e(p_H, p_L) = (1-\chi^e)\beta_L^e\Lambda$, and $\lambda_H^e(p_H, p_L) = \chi^e\beta_H^e\Lambda$, respectively. The expert's incentives and resultant consumer skepticism clearly lead to destruction of social value. To measure the extent of value reduction due to expert's equilibrium pricing and service strategy, we first discuss the socially

optimal (first-best) provision of the service in Section 4.

4 Socially Optimal Provision of the Service

We consider the first-best provision of the expert service by a social planner with complete information on each problem type. The social planner's objective is to maximize the value generated by the capacitated service system net of congestion costs, by suitably controlling the admission to the queue for each treatment.

Remark: The social planner will always provide true diagnosis ($\chi = \theta$). Overtreatment decreases social welfare, as it increases consumer waiting costs without increasing the value of the service provided.

Now, we explore the admission control problem. Admitting a consumer with a major problem into the system generates a value v_H for the customer, while increasing the expected workload at the server by τ_H units of time. Similarly, admitting a minor problem consumer into the system, generates value v_L while increasing the expected workload by τ_L. Thus, every admission increases workload causing the expected wait time to increase for all admitted consumers, since they all share the same service queue.

Let the admission rates for minor and major problems be λ_L and λ_H respectively. Thus, ex ante admission probability for a minor problem customer is therefore $\lambda_L/(\Lambda\bar{\theta})$ and the admission probability for a major problem customer is $\lambda_H/(\Lambda\theta)$. Under socially optimal admission policy, we have $\chi = \theta$. Using equation (4), we can write the expected waiting time in queue as $W_q\left(\theta, \frac{\lambda_L}{(1-\theta)\Lambda}, \frac{\lambda_H}{\theta\Lambda}\right)$. For expositional ease, we drop other parametric dependencies and write $W_q\left(\theta, \frac{\lambda_L}{(1-\theta)\Lambda}, \frac{\lambda_H}{\theta\Lambda}\right)$ as $W_q(\lambda_L, \lambda_H)$.

A consumer with a minor problem derives a net value of $v_L - c\tau_L - cW_q(\lambda_L, \lambda_H)$, and the consumer with a major problem derives a net value of $v_H - c\tau_H - cW_q(\lambda_L, \lambda_H)$. Let $SW(\lambda_L, \lambda_H)$ denote social welfare when the admission rates for minor and major treatments are λ_L and λ_H respectively. Then,

$$SW(\lambda_L, \lambda_H) \triangleq \lambda_L(v_L - c\tau_L) + \lambda_H(v_H - c\tau_H) - c(\lambda_L + \lambda_H)W_q(\lambda_L, \lambda_H). \tag{10}$$

The social planner's objective would be to choose admission to maximize welfare, i.e.,

$$\max_{\{0\leq\lambda_H\leq\theta\Lambda, 0\leq\lambda_L\leq\bar{\theta}\Lambda\}} SW(\lambda_L, \lambda_H).$$

We will demonstrate that the socially optimal actions are highly dependent on various parameter scenarios, despite our model parsimony. The objective function in (10) is not quasi-concave. In what follows, we will develop on the key drivers of the optimal social welfare maximizing policy

can be characterized in the population. As a first step, we analyze social welfare, when the expert fixes the admission rate for one of the treatments in the following Section 4.1.

4.1 Type-Dependent Welfare Decisions

In Lemma 1, we characterize the admission control policy for a specific treatment (minor or major) by holding the other admission decision exogenous. The first part of Lemma 1 illustrates the socially optimal admission rate for consumers with minor problems, $\lambda_L^*(\lambda_H)$, for a fixed admission rate of major problems. In the second part of Lemma 1, the socially optimal admission rate for consumers with major problems, $\lambda_H^*(\lambda_L)$, for a fixed minor problems admission is presented. All proofs and technical derivations are presented in the Appendix.

Lemma 1. *1. Given λ_H, the social welfare $SW(\lambda_L, \lambda_H)$ is maximized by:*

$$\lambda_L^*(\lambda_H) = \begin{cases} \frac{1-\lambda_H\tau_H}{\tau_L} - \sqrt{\frac{c\gamma_L(\lambda_H)}{v_L\tau_L}} & \text{if } 0 \le \lambda_H \le \widehat{\lambda}_H \\ \\ 0 & \text{if } \lambda_H > \widehat{\lambda}_H \end{cases} \tag{11}$$

where $\gamma_L(\lambda_H) = 1+(1/\tau_L)(1-\lambda_H\tau_H)\lambda_H(\tau_H-\tau_L)^2$, and $\widehat{\lambda}_H$ is such that $\partial SW(0,\widehat{\lambda}_H)/\partial\lambda_L = 0$.
$\lambda_L^*(\lambda_H)$ is decreasing in λ_H for $\lambda_H \le \widehat{\lambda}_H$.

2. Given λ_L, the social welfare $SW(\lambda_L, \lambda_H)$ is maximized by:

$$\lambda_H^*(\lambda_L) \;=\; \begin{cases} \frac{1-\lambda_L\tau_L}{\tau_H} - \sqrt{\frac{c\gamma_H(\lambda_L)}{v_H\tau_H}} & \text{if } 0 \le \lambda_L \le \widehat{\lambda}_L \\ \\ 0 & \text{if } \lambda_L > \widehat{\lambda}_L \end{cases} \tag{12}$$

where $\gamma_H(\lambda_L) = 1 + \frac{(1-\lambda_L\tau_L)\lambda_L(\tau_H-\tau_L)^2}{\tau_H}$, and $\widehat{\lambda}_L$ is such that $\frac{\partial SW(\widehat{\lambda}_L,0)}{\partial\lambda_H} = 0$.
$\lambda_H^*(\lambda_L)$ is decreasing in λ_L for $\lambda_L \le \widehat{\lambda}_L$.

Lemma 1(1) illustrates the optimal admission rate for consumers with minor problems, $\lambda_L^*(\lambda_H)$, when the admission rate for consumers with major problems is fixed at λ_H. Since $\lambda_L^*(\lambda_H)$ is decreasing in λ_H, fewer minor problems are admitted as the admission rate for major problems increases. When major problem consumers are admitted at a higher rate, both the utilization of the system and overall waiting costs are higher. As a result, fewer minor problem consumers can be admitted.

In particular, if the fixed admission rate for consumers with major problems is higher than $\widehat{\lambda}_H$, the incremental waiting cost for the population caused by the admission of one more minor problem consumer, exceeds the additional value of treating the minor problem v_L. Hence, it is not

socially beneficial to admit one more consumer with a minor problem. Above a threshold of major problem admissions, it is socially optimal that no more minor problem consumers are admitted: $\lambda_L^*(\lambda_H) = 0$ for $\lambda_H \geq \widehat{\lambda}_H$. When the major problem admission rate is below $\widehat{\lambda}_H$, minor problem consumers can be accommodated.

The results of Lemma 1(part 2) mirrors that of Lemma 1(part 1). Fix λ_L. The socially optimal admission rate of major problem consumers $\lambda_H^*(\lambda_L)$ is decreasing in λ_L, and after the certain threshold admission rate of λ_L, it is (conditionally) optimal to not admit any major problem consumers. This is a direct outcome of the value-time dependency. Even though major treatments bring higher value, they are insufficient to make up for the additional waiting times created due to the longer service time.

To wit, after sufficient consumers of a type are admitted, it is socially optimal to deny admission to the other type. Social benefits accrue by trading off one type of problems against the other. In particular, it is socially optimal for monopolist service providers with high utilization (i.e., high admission compared to capacity) to engage in *specialization*. We further characterize the social welfare function in Lemma 2.

Lemma 2. *1. $SW(\lambda_L^*(\lambda_H), \lambda_H)$ is convex in λ_H for $\lambda_H \leq \widehat{\lambda}_H$ and concave in λ_H for $\lambda_H > \widehat{\lambda}_H$.*

2. $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is convex in λ_L for $\lambda_L \leq \widehat{\lambda}_L$ and concave in λ_L for $\lambda_L > \widehat{\lambda}_L$.

Lemma 2(1) shows that the social welfare is convex in λ_H for major problem admission rate is sufficiently low ($\lambda_H \leq \widehat{\lambda}_H$), as long as the minor problem customers are admitted optimally (as prescribed by Lemma 1). If the major admission rate is above the threshold, i.e., $\lambda_H > \widehat{\lambda}_H$, it is optimal to deny service to minor problem consumers. In this case, the social welfare is concave in λ_H. A corresponding result for welfare as a function of minor problem admission rate is provided in Lemma 2(2).

4.2 Priorities and Service Specialization

Service Specialization. Following Lemma 2, we observe that in large market sizes (markets in which the arrival rates are comparable to capacity or exceed it), the social planner would prefer to serve only one of the problem types. Recall that service value has *diminishing returns* as service times increase, i.e., $v_H > v_L$ but $v_H/\tau_H < v_L/\tau_L$. Hence depending on the relative *value accrued per unit time* of the treatments, one of the treatments is preferred. Thus, social welfare is maximized by specialization. Let the admission policy be written as (minor admission rate, major admission rate). Then two possible modes of specialization exist.

(a) Admitting some minor-problem customers and denying all major problem customers. The optimal admission policy is $(\lambda_L^*(0), 0))$, or

(b) Admitting some major problem customers and denying all minor problem customers. The optimal admission policy is $(0, \lambda_H^*(0))$.

Admission Priorities. In smaller markets (i.e., arrival rates are small compared to service capacity), there is enough capacity to admit more than one type of customer problems. The social planner prioritizes by

(a) Admitting all minor problems customers compared to major problem customers, or

(b) Admitting all major problem customers over minor problem arrivals.

In Lemma 3, we explore the conditions that impose the preference order in service prioritization. In order to understand the priority rules better, particularly to understand why minor problems may be prioritized, recall that major treatments take a longer duration. There is some social value lost as the service time increases, as a result of diminishing marginal returns and the service times imposing negative waiting cost externalities on other customers. For notational convenience, let us define this welfare difference as Δ, where it can be shown that

$$\Delta \stackrel{\text{def}}{=\!=} \frac{c(v_H - \sqrt{v_H c \tau_H})(\tau_H - \tau_L)^2}{\tau_L \tau_H \sqrt{c v_H \tau_H}} > 0.$$

The term Δ could be thought of as the additional social value (rate) achieved, when the value of service increases from v_L to v_H, even as the expected service time increases from τ_L to τ_H. When

$$\frac{v_L}{\tau_L} = \frac{v_H}{\tau_H} + \Delta,$$

the social welfare planner would find no benefit in admitting one type of consumers over the other.[4] In other words, when the above relationship holds, treatments of minor and major problems provide the same social value.

Hence, depending on how v_L/τ_L compares with $\Delta + v_H/\tau_H$, we have two cases:
(a) $\Delta \leq \frac{v_L}{\tau_L} - \frac{v_H}{\tau_H}$. The marginal value of major treatments over time is *strongly* diminishing. Longer major treatments do not create sufficient additional value to overcome negative externalities.
(b) $\Delta > \frac{v_L}{\tau_L} - \frac{v_H}{\tau_H}$. The marginal value of major treatments over time is *weakly* diminishing. Longer major treatments bring in additional value sufficient to overcome waiting cost externalities.
Hence, as such, in Scenario (a) social planner exclusively offers minor treatments or prioritize them over major treatments. In Scenario (b), he prioritizes or specializes in major treatments. Before we formalize the social welfare action, we prove a technical result in Lemma 3.

[4]Recall that due to diminishing marginal value of service time, we have $v_H/\tau_H < v_L/\tau_L$. This assumption can be relaxed without affecting our conclusions. For instance, it can be intuited from the conditions in §4, that overtreatment fraud is more likely when marginal value of more treatment is not diminishing in time.

Lemma 3. *When* $\frac{v_H}{\tau_H} \leq \frac{v_L}{\tau_L} < \Delta + \frac{v_H}{\tau_H}$ *then:*

1. *$SW(\lambda_L, \lambda_H)$ has a unique saddle point $(\underline{\lambda_L}, \underline{\lambda_H}) \in [0, \widehat{\lambda}_L] \times [0, \widehat{\lambda}_H]$, such that $\underline{\lambda_L} = \lambda_L^*(\underline{\lambda_H})$ and $\underline{\lambda_H} = \lambda_H^*(\underline{\lambda_L})$.*

2. *For $\bar{\theta}\Lambda \geq \underline{\lambda_L}$ define $\widetilde{\Lambda}_H(\bar{\theta}\Lambda) \in [\underline{\lambda_H}, \infty)$ as,*

$$f\widetilde{\Lambda}_H(\bar{\theta}\Lambda) \quad = \quad \begin{cases} \lambda_H \text{ s.t. } SW(\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda)) = SW(\lambda_L^*(\lambda_H), \lambda_H) & \text{if } \underline{\lambda_L} \leq \bar{\theta}\Lambda \leq \widetilde{\lambda}_L \\ \infty & \text{if } \bar{\theta}\Lambda > \widetilde{\lambda}_L \end{cases}$$

where $\widetilde{\lambda}_L > \underline{\lambda_L}$ is unique solution to $SW(\widetilde{\lambda}_L, \lambda_H^*(\widetilde{\lambda}_L)) = SW(0, \lambda_H^*(0))$.

From Lemma 3, if the treatment value is weakly diminishing, i.e., $\frac{v_H}{\tau_H} > \frac{v_L}{\tau_L} - \Delta$, in small markets with few consumers with minor problems, and sufficiently large arrivals with major problems,major treatments are prioritized. Equipped with the proven results, we are now ready to describe the socially optimal service policy in Proposition 1.

4.3 Welfare Maximizing Admission Policy

In Proposition 1, we segment our finding following the preceding discussion into two scenarios: (a), *strongly* diminishing marginal returns on the value in time and in Scenario (b), *weakly* diminishing marginal returns. The socially optimal admission rates are marked by superscript S.

Proposition 1. *(a) If $\Delta \leq \frac{v_L}{\tau_L} - \frac{v_H}{\tau_H}$, the socially optimal service policy is:*

$$(\lambda_L^S, \lambda_H^S) \quad = \quad \begin{cases} \big(\min\{\bar{\theta}\Lambda, \lambda_L^*(0)\}, 0\big) & \text{if } \bar{\theta}\Lambda > \widehat{\lambda}_L \\ \big(\bar{\theta}\Lambda, \min\{\theta\Lambda, \lambda_H^*(\bar{\theta}\Lambda)\}\big) & \text{if } \bar{\theta}\Lambda \leq \widehat{\lambda}_L \end{cases}$$

(b) If $\frac{v_L}{\tau_L} - \frac{v_H}{\tau_H} < \Delta$, the socially optimal service policy is:

$$(\lambda_L^S, \lambda_H^S) \quad = \quad \begin{cases} \big(\min\{\bar{\theta}\Lambda, \lambda_L^*(\theta\Lambda)\},\ \min\{\lambda_H^*(0), \theta\Lambda\}\big) & \text{if } \bar{\theta}\Lambda < \underline{\lambda_L} \\ (\lambda_L^*(\theta\Lambda),\ \min\{\lambda_H^*(0), \theta\Lambda\}) & \text{if } \bar{\theta}\Lambda \geq \underline{\lambda_L} \text{ and } \theta\Lambda > \widetilde{\Lambda}_H(\bar{\theta}\Lambda) \\ \big(\min\{\bar{\theta}\Lambda, \lambda_L^*(0)\},\ \min\{\lambda_H^*(\bar{\theta}\Lambda), \theta\Lambda\}\big) & \text{if } \bar{\theta}\Lambda \geq \underline{\lambda_L} \text{ and } \theta\Lambda \leq \widetilde{\Lambda}_H(\bar{\theta}\Lambda) \end{cases}$$

Proposition 1(a) addresses strongly diminishing marginal returns of treatment time. In this case, prioritizing major treatments leads to welfare-loss, either due to longer service time, or weak additional value provided for the time taken. Therefore, it is socially optimal to prioritize minor

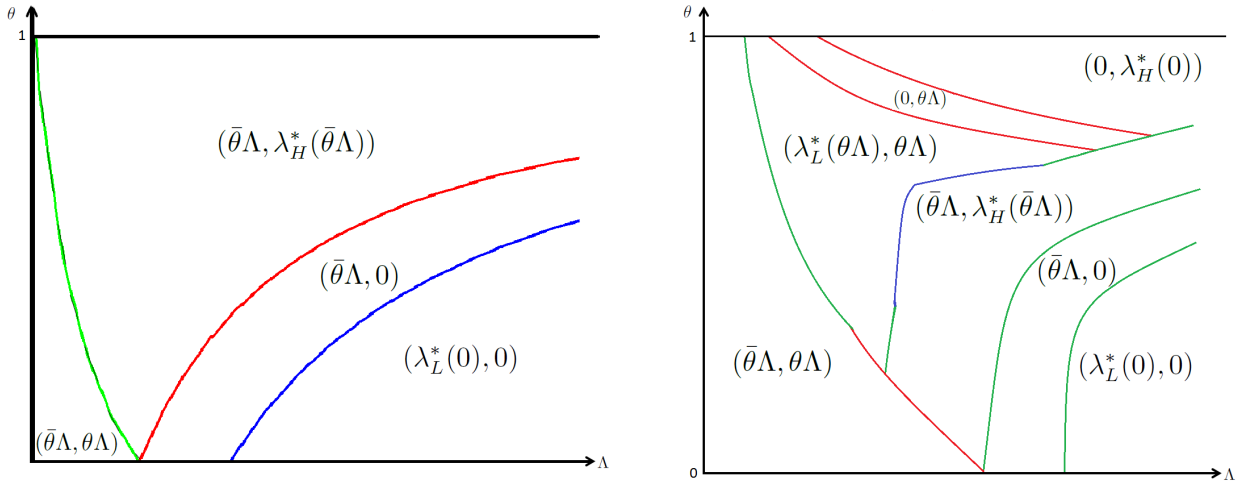

Figure 1: Socially optimal admission rates (minor types λ_L^S, major types λ_H^S). (a) Strongly diminishing returns (left panel), (b) Weakly diminishing returns (right panel). x-axis: Market size (Λ), y-axis: Incidence probability of major problems θ.

problems regardless of the market size (Λ). If the market is small or when the minor problem incidence is low, i.e., $\Lambda\bar{\theta} < \widehat{\lambda}_L$ or simply, $\Lambda < \widehat{\lambda}_L/(1-\theta)$, the service is *rationed* for major problem customers. Major problem admission is capped just to allow for treatment of *all* minor problem arrivals. When the market size is large, $\Lambda > \widehat{\lambda}_L/(1-\theta)$, utilization levels are high, hence all major problem types are denied admission, and social welfare is maximized by service specialization.

Left panel of Figure 1 characterizes the space of social-welfare maximizing policies for the strongly diminishing returns case. It can be seen that, when the market is sufficiently large (south-east corner), no major admissions are made. As can be seen in the Figure 1(a), it is welfare improving to admit major problem treatments when (i) the market size is sufficiently small – the service provider has excess service capacity (left area in the figure), or (ii) the major problem incidence is high – there are too few minor problems in the market compared to the available capacity. In both cases, any admission of major treatment cases is driven by capacity utilization considerations. The admission rate for major treatments, is decreasing in arrival rates and increasing in θ.

Proposition 1(b) addresses weakly diminishing marginal returns of treatment time. Major treatments are prioritized since they increase welfare – they bring in sufficiently high additional service value that compensates for the additional delays created due to longer major treatment time. The socially optimal admission control is an outcome of the interplay between major problem incidence and market size, as before, but also based on value per capita (v_H vs. v_L), as can be seen

in the right panel of Figure 1.

As before, capacity considerations drive similar welfare actions (as in Scenario (a)). When the market is small and the available capacity is in excess, i.e., under low utilization, the negative impact of capacity constraints and service delay is low. All types are admitted into service. (See left side of the figure in the right panel). For sufficiently large markets, $\Lambda > \widehat{\lambda}_L/(1-\theta)$, it is optimal to admit only the minor problems types as they generate higher value per unit of service time. However, when prevalence of major problems is significant, i.e., $(1-\theta)\Lambda < \underline{\lambda_L}$, i.e, $\theta > 1 - \underline{\lambda_L}/\Lambda$, consumers with major problems are prioritized over those with minor problems.

Finally, when the incidence of major problem is sufficiently low in large markets – how low depends on the market size – there are not enough major problems to treat. Hence, a social planner would just prioritize or specialize in serving minor problem customers.

To conclude, the social welfare maximizing planner either specializes in serving a class of customer problems, or prioritizes the class of problems. Whether the social planner prioritizes or specializes in minor or major problem types depends on the service value vs. time, specifically how quickly the marginal service value diminishes with additional time.

5 Revenue Maximization: Equilibrium Prices and Diagnosis

We now focus on the expert's pricing decision in equilibrium. The expert decides on the prices of the minor and major treatments in order to maximize his revenues. Information asymmetry plays a key role in the outcomes. In order to characterize the asymmetry created by incomplete information, we first discuss the complete information case. We follow the same information structure as in the welfare case except that consumers can choose to join or balk.

Equilibrium under Complete Information. If consumers have complete information about their own problem types, the equilibrium demand will be identical to the socially optimal admission rates $(\lambda_L^S, \lambda_H^S)$ given in Proposition 1. The expert's diagnosis is honest. Indeed, the need for diagnosis is obviated by complete information. Note that it would not be rational for a customer who knows her own type to wait for an incorrect treatment: Major types do not benefit from minor treatment, and minor types do not incur longer waits for unnecessary major treatment at a higher price. Under complete information, the expert can price the services to achieve the maximum social welfare, and fully extract the surplus from consumers joining the unobservable service queue. Thus, the first best (social welfare maximum) is achieved if there is no information asymmetry. This observation is consistent with existing literature on unobservable queues (Hassin and Haviv 2004) in which the server extracts all consumer welfare.

Efficiency: We address an equilibrium with first best outcomes, as an "efficient" equilibrium since no social welfare is lot. Now, we examine the equilibria under information asymmetry.

Information Asymmetry. When there is information asymmetry between consumers and the expert, the diagnostic strategies play a bigger role. The prices serve several purposes: Prices determine the extent of consumer demand that is covered by the firm; they control congestion in the system and they inform the market on the expert's diagnostic strategy.

The expert's objective function can be written as:

$$\max_{\{p_L \geq 0, p_H \geq 0\}} \left\{ p_L \lambda_L^e(p_H, p_L) + p_H \lambda_H^e(p_H, p_L) \right\}, \tag{13}$$

For brevity, we will demonstrate our results for strongly diminishing returns for service value. Recall that, to improve welfare, consumers with minor problems are prioritized; This is hence the most stringent existent condition for the experts incentives to overtreat. We will show the overtreatment equilibrium nevertheless emerges.

To begin the equilibrium analysis, we first shed light on prices under two extreme cases: *honesty* – the case when all decisions are accurate and there is no overtreatment, and *complete overprovision* – in which all minor problems are recommended overtreatment.

$$\bar{p}_L(\lambda_L, \lambda_H) = v_L - cW_L\left(\theta, \frac{\lambda_L}{(1-\theta)\Lambda}, \frac{\lambda_H}{\theta\Lambda}\right), \tag{14}$$

$$\bar{p}_H(\lambda_L, \lambda_H) = v_H - cW_H\left(\theta, \frac{\lambda_L}{(1-\theta)\Lambda}, \frac{\lambda_H}{\theta\Lambda}\right), \tag{15}$$

Hence $\bar{p}_L$ and $\bar{p}_H$ are the prices for minor (L) and major (H) treatments respectively under *honest* diagnosis. Next, we characterize the price for major treatment when the expert false-diagnoses, recommending major treatment to all arriving customers regardless of their type.

$$\bar{p}_C(\lambda) = v_L + \theta(v_H - v_L) - cW_H\left(1, 0, \frac{\lambda}{\Lambda}\right). \tag{16}$$

It is evident that $\bar{p}_L$, $\bar{p}_H$ and $\bar{p}_C$ all extract the entire consumer welfare. Thus, the ex-ante expected utility net of waiting costs from consuming the offered service is zero.

In order to characterize the equilibrium, we require some technical results that are characterized in Lemmas 4 and 5. In Lemma 4, we focus on the equilibrium demand on major types under honest diagnostic strategy, which helps us to compare the results with the social welfare (which is honest), and thus develop market conditions amenable for honest (and efficient) equilibrium. To begin with, we define $\bar{\lambda}_H(\Lambda)$ as the *maximum* rate of major problem consumers that can be treated through an honest diagnostic policy, if the expert serves all the minor problem types. $\bar{\lambda}_H$ captures the market coverage limit on major treatments under an honest diagnostic policy. This helps us pin down the

conditions conducive for overtreatment in diagnosis.

Lemma 4. *1. For $\Lambda \leq \frac{\lambda_L^*(0)}{(1-\theta)}$, the maximum major type demand $\bar{\lambda}_H(\Lambda)$ satisfies:*

$$\frac{\bar{p}_H((1-\theta)\Lambda, \bar{\lambda}_H(\Lambda))\bar{\lambda}_H(\Lambda)}{\theta\Lambda} = \bar{p}_L((1-\theta)\Lambda, \bar{\lambda}_H(\Lambda)). \tag{17}$$

2. $\bar{\lambda}_H(\Lambda)$ is increasing in θ.

3. $\exists\ \overline{\Lambda}_s$ such that the efficient (first best) provision cannot be achieved for $\Lambda < \overline{\Lambda}_s$.

Lemma 4(1) provides an implicit characterization of the market coverage of major problem types under an honest diagnostic policy. For a given Λ, if all the consumers with minor problems join the service queue, $\bar{\lambda}_H(\Lambda)$ denotes the maximal major problem consumers that can join under *honest* diagnosis. This upper-bound on admission under honest diagnosis helps us determine the market size threshold when the expert benefits from providing false recommendations. As the incidence of major problem θ increases, the upper bound on major problem treatments increases.

Lemma 4(3) derives the market size threshold below which the expert's incentive to misdiagnose creates large consumer costs and diminished social welfare. Thus, when the market demand is sufficiently low, the expert has capacity to provide overtreatment and accrue excess revenues through strategic diagnosis.

Although such false diagnosis and overtreatments lead to improved revenues for the service providers, they create two costs for the consumers, which reduces welfare. First, some consumers pay more for services they do not necessarily need. Second, the increased service times increase overall delays for all consumers. All consumers now expect to wait longer for a *less* honest service. As a result, there is consumer welfare loss, which may exceed the service providers' revenue gains, resulting in social welfare loss. Hence, the equilibrium is inefficient (the first-best is not achieved) in small markets.

Providing false recommendations dissuades customers from joining the service queue because of two factors: the reduced expected utility due to anticipated strategic diagnosis, and longer waits due to overtreatment. This leads to some consumers balking and hence causes reduced equilibrium demand for the service provider. Such demand loss due to congestion, acts as a *deterrent* to the expert from providing false recommendations and overtreating minor problem consumers.

Nevertheless, the incentive to overtreat is not eliminated. Let us examine the maximum revenue the expert can earn through overtreatments following consistent false diagnosis to *all* arriving minor problems. Given some Λ, the revenues are given by:

$$\Pi_c(\Lambda) = \mathbb{I}(\Lambda > \lambda_c^*)\left(\sqrt{\frac{\theta v_H + (1-\theta)v_L}{\tau_H}} - \sqrt{c}\right)^2 + \mathbb{I}(\Lambda \leq \lambda_c^*)\Lambda\left((\theta v_H + (1-\theta)v_L) - cW_H(1,0,1)\right),$$

where $\mathbb{I}(\cdot)$ is the indicator function. $\lambda_c^* = \frac{1}{\tau_H} - \sqrt{\frac{c}{(\theta v_H + \bar{\theta} v_L)\tau_H}}$ is the effective arrival rate maximizing the expert's revenues, when major treatment is recommended to every customer. Using the above result, we provide conditions when it is optimal for the expert to cheat/overtreat customers with minor problems in Lemma 5.

Lemma 5. *When $\theta > \frac{c(\tau_H - \tau_L)}{v_H - v_L}$ and $\forall \lambda < \bar{\Lambda}_c$, false recommendations accrue higher revenues for the expert than honest recommendations. $\bar{\Lambda}_c$ is the unique solution to $\Pi_c(\Lambda_c) = (1-\theta)\Lambda_c \bar{p}_L(\bar{\theta}\Lambda_c, \bar{\lambda}_H(\Lambda_c)) + \bar{\lambda}_H(\Lambda_c)\bar{p}_H(\bar{\theta}\Lambda_c, \bar{\lambda}_H(\Lambda_c))$.*

Lemma 5 if the market size is smaller than a threshold *and* if there is sufficient incidence of major problems, overtreatment becomes more profitable than honest diagnostic strategy. The first condition is a direct effect of capacities. If the utilization is high, the service provider would maximize revenues without having to misdiagnose. On the other hand, when the arrival rates are small, deliberate overtreatment is a mechanism is to employ the excess service capacity that would otherwise remain unused.

However, a significant prevalence of major problems in the population, is essential for major treatment diagnosis to be believable for consumers. If the known incidence rate is low, consumers are less likely to trust major recommendations and would balk from the queue, i.e., "falseness" in recommendations is not sufficiently credible. Using the Lemmas 4 and 5, we are ready to characterize the optimal prices and the equilibrium outcomes in Proposition 2.

Proposition 2. *The equilibrium arrival rates $(\lambda_L^e, \lambda_H^e)$ and optimal prices (p_L^*, p_H^*) are as follow:*

1. ***Minor Treatment (Specialization) Equilibrium:*** *When $\Lambda \geq \frac{\widehat{\lambda}_L}{1-\theta}$, the expert provides an honest diagnosis, $\chi^* = \theta$, and only minor problem consumers join.*

$$\begin{aligned}(\lambda_L^e, \lambda_H^e) &= \begin{cases} (\lambda_L^*(0), 0) & \text{if } \Lambda > \lambda_L^*(0)/\bar{\theta} \\ (\bar{\theta}\Lambda, 0) & \text{if } \lambda_L^*(0)/\bar{\theta} \geq \Lambda \geq \widehat{\lambda}/\bar{\theta} \end{cases} \\ (p_L^*, p_H^*) &= (\bar{p}_L(\lambda_L^e, 0), -) \text{ if } \Lambda > \lambda_L^*(0)/\bar{\theta}.\end{aligned}$$

2. ***Honest Price Discrimination:*** *When $\bar{\Lambda}_c \leq \Lambda < \frac{\widehat{\lambda}_L}{1-\theta}$, the expert provides an honest diagnosis, $\chi^* = \theta$. All minor problem consumers and some major problem consumers join.*

$$\begin{aligned}(\lambda_L^e, \lambda_H^e) &= \begin{cases} (\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda)) & \text{if } \widehat{\lambda}/\bar{\theta} > \Lambda \geq \bar{\Lambda}_s \\ (\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda)) & \text{if } \bar{\Lambda}_s \geq \Lambda \geq \bar{\Lambda}_c \end{cases} \\ (p_L^*, p_H^*) &= (\bar{p}_L(\lambda_L^e, \lambda_H^e),\ \bar{p}_H(\lambda_L^e, \lambda_H^e)) \quad \text{if } \widehat{\lambda}/\bar{\theta} > \Lambda \geq \bar{\Lambda}_c\end{aligned}$$

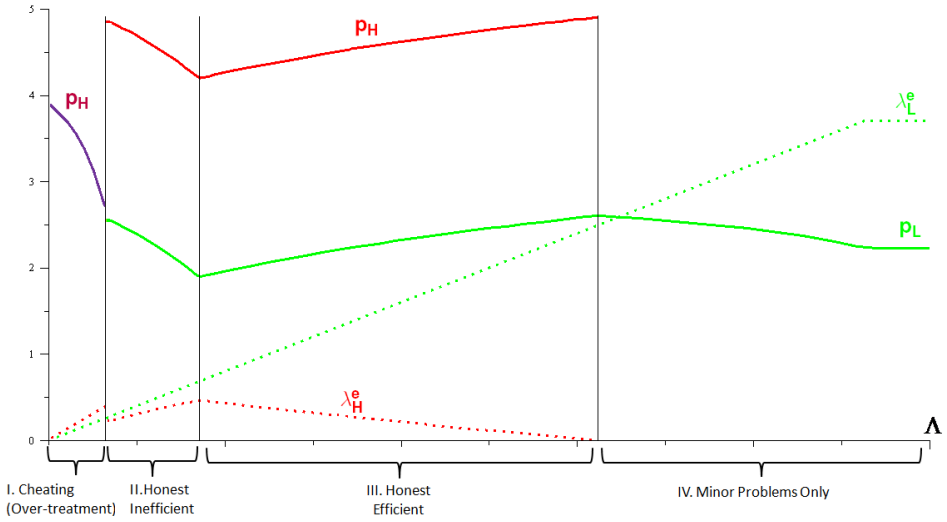

Figure 2: Optimal prices and the equilibrium demand as a function of the potential demand Λ. Solid lines: optimal prices, dotted lines: equilibrium demand.

3. ***Overtreatment Equilibrium:*** *When $\Lambda < \bar{\Lambda}_c$, the expert provides false recommendations to all minor problem consumers $\chi^* = 1$. If $\theta > \widehat{\theta}$, the market coverage is partial when Λ is sufficiently high.*

$$(\lambda_L^e, \lambda_H^e) = \begin{cases} (0, \lambda_c^*) & \text{if } \bar{\Lambda}_c > \Lambda \geq \lambda_c^* \\ (0, \Lambda) & \text{if } \min\{\lambda_c^*, \bar{\Lambda}_c\} \geq \Lambda \geq 0 \end{cases}$$

$$(p_L^*, p_H^*) = (\bar{p}_C(\lambda_H^e) - \delta, \bar{p}_C(\lambda_H^e)) \quad \text{if } \bar{\Lambda}_c > \Lambda \geq 0$$

for any $\delta \in (0, \bar{p}_C)$.

Proposition 2 presents the optimal prices and equilibrium demands for different levels of market size. Despite the information asymmetry between the expert and the consumers, the optimal prices achieve the first-best provision of the service as long as arrival rate is sufficiently high, i.e., $\Lambda \geq \bar{\Lambda}_s$. Figure 2 illustrates the optimal prices, (p_L^*, p_H^*), and the equilibrium demand, $(\lambda_L^e, \lambda_H^e)$ as functions of demand, Λ.

5.1 Minor Treatment (Service Specialization) Equilibrium

When the market size is sufficiently large $\Lambda \geq \widehat{\theta\lambda}_L$, the expert specializes in providing only the minor treatment while making *honest* recommendations. He chooses to price out all consumers with major problems (by setting an arbitrarily high price for those services).[5] This equilibrium is captured in region IV in the Figure 2. Even though only minor problem consumers are attended

[5]Major treatments are prioritized when the service value is weakly diminishing in time, as shown before.

to, only some of them can join, since congestion costs can be make net value negative.

The optimal price for the minor treatment, p_L^*, is decreasing in Λ (as evident in Figure 4) for $\Lambda < \lambda_L(0)/\bar{\theta}$. As the market grows (i.e., Λ increases), the expert cuts price to admit more minor consumers into the system. However, it is not optimal to reduce prices further and admit more than $\lambda_L^*(0)$ consumers into the system. In this case, an increased incidence of major problem, i.e., high θ, results in a lower demand for the expert as it necessarily implies fewer minor problems in the market ($(1-\theta)\Lambda$ is lower). The expert responds to the decreasing demand by increasing the price p_L^*.

5.2 Honest Equilibrium

When the market size is sufficiently large but not too large (i.e., when $\bar{\Lambda}_c < \Lambda \leq \frac{\widehat{\lambda}_L}{\bar{\theta}}$), the expert provides honest recommendations, $\chi^* = \theta$, (i.e., all minor problem consumers are recommended only the minor treatment), but also has enough excess capacity to provide major treatment for major problem consumers. Depending on the market size, honesty could be efficient.

Honest Efficient Equilibrium. If market size is such that $\bar{\Lambda}_s \leq \Lambda \leq \widehat{\lambda}_L/\bar{\theta}$, (see Region III in Figure 2) the optimal prices (p_L^*, p_H^*), achieve the first-best service provision. Note as the potential demand or market size Λ increases, both prices increase. *Hence, the number of joining customers increases even as the price increases.* Typically, service providers need to cut the price to admit more consumers into the system. The observed result is due to preference of types.

Note that the expert prefers consumers with minor problems to those with major problems. As Λ increases, the equilibrium demand for the major treatment decreases, while the equilibrium demand for the minor treatment increases. The expert sets prices to preferentially admit minor treatments. This increases the ratio of minor treatment consumers to major treatment consumers joining the queue, λ_L^e/λ_H^e. Hence the average equilibrium service time of the service provider decreases. This leads to lower waiting times, despite the higher number of consumers $(\lambda_L^e + \lambda_H^e)$ joining the service queue in equilibrium. Finally, the reduced cost of waiting allows the expert to charge higher prices to extract the additional welfare.

Honest Inefficient Equilibrium The provision of the service is honest but not efficient (i.e. the first best is not attained) for $\Lambda_c \leq \bar{\Lambda} \leq \bar{\Lambda}_s$. (See corresponding Region II in Figure 2). In fact, there may be underprovision of the major treatment due to information asymmetry. The expert would be better off by serving more consumers with major problems. But consumers believe that major treatments are often overtreatments. Hence, the expert cannot commit to an honest diagnosis. Note that as major problem incidence θ increases, the expert's revenues decrease. As the potential demand for the minor treatment ($(1-\theta)\Lambda$) decreases with increasing θ, the expert admits more consumers with major problems into the system (λ_H^e is increasing), which results in a

slower service rate in equilibrium and longer waiting times. As a result, both prices (p_L^*, p_H^*) and revenues fall.

Interestingly, as the value of treating major problems, v_H, increases, the equilibrium demand for the major treatment, λ_H^e, decreases. This is mainly due to the increasing likelihood of overprovision (expert cheating) with the increasing value gap between the minor and major treatments. Despite the decreasing equilibrium demand, the expert is able to achieve higher revenues by increasing prices (p_L^*, p_H^*). When the value of treating minor problems, v_L, increases, the equilibrium demand for the major treatment, λ_H^e, and the prices, (p_L^*, p_H^*), increase, yielding higher revenues to the expert.

5.3 Overtreatment Equilibrium

In small markets ($\Lambda < \bar{\Lambda}_c$), the expert recommends major treatment to all arriving consumers, $\chi^* = 1$. As a result, even though consumers do not know their type, they know that with probability $\bar{\theta}$ they could be overtreated. Hence, consumers have a lower willingness to pay for the major treatment recommended. In sufficiently small markets with with low incidence of major problems ($\theta \leq \widehat{\theta}$ and $\Lambda \leq \bar{\Lambda}_c$), there exists an overtreatment equilibrium in which all consumers knowingly join. The expert provides all consumers with the major treatment, i.e., $\lambda_H^e = \Lambda$.

Expert false diagnosis becomes more likely as v_H and θ increase. The optimal price for major treatment (which is often an overtreatment) increases with the incidence of major problem (or the value of resolving those major problem). Since, consumers anticipate that the expert will cheat. It makes them reluctant to join due to skepticism about the true value of service, leading to some consumers balking (likely to seek external advice).

Note that under the overtreatment equilibrium, the minor treatment price can take any value lower than $\bar{p}_C$. The lower minor treatment can essentially be considered "bait-and-switch" prices that are offered, but never materialized. Hence low prices for minor treatments do not provide any further confidence in service provider's honesty.

As consumers become more sensitive to delays, overtreatment becomes less likely in equilibrium: As the cost of waiting c increases, the expert can achieve honest price discrimination while serving fewer consumers ($\bar{\Lambda}_c$ decreases) in equilibrium; hence, the potential for honesty increases. If consumers are more impatient to delays, overtreatment becomes less profitable. Thus, congestion dampens expert cheating behavior.

6 Conclusion

The information asymmetry between service providers and consumers is an obstacle preventing honest and efficient provision in expert service markets. Experts often have incentives to exploit

their informational advantage over consumers by recommending unnecessary services and treatments to consumers. Despite the growth of service economy, the overtreatment problem remains unexplored in service operations. In our paper, we consider an expert with serving consumers who arrive with problems that they cannot self-diagnose. Upon diagnosis, consumers update their beliefs on offered diagnosis treatment and corresponding price, and decide whether to go through the service.

We show that when service capacity is limited and consumers are sensitive to delay, exploiting the informational advantage through overtreatment is costly for an expert. Unnecessary treatments take longer and use up capacity, which intensifies the congestion in the system, leading to longer wait times and smaller demand. Hence, congestion costs act a deterrent to expert overtreatment.

As a result of the incentives, committing to an honest diagnostic policy becomes costly for the expert, especially in small markets. In fact, as major treatments become more valuable, there is increased skepticism (due to overprovision) that deters more consumers from joining an honest service. In these cases, prices reveal information on likely over-treatment. To make his actions credible, an honest expert has to charge sufficiently high prices for both minor and major treatments, and limit the demand for the more profitable major treatment. As the market-size decreases, signaling an honest diagnosis becomes costlier, since having excess service capacity makes overtreatment easier. As a result, overtreatment is unavoidable in small markets.

The social planner would recommend appropriate treatments. Nevertheless, it may be prudent for the social planner under certain conditions, if the expert *specializes or prioritizes* in providing low-value minor treatments (or major treatments). Minor treatment specialization occurs when major treatments which provide more value per service take up too much resource and time, diminishing the availability of the service for minor treatments, which are typically more common. We recommend that in such cases, capping prices (or reimbursements), would lead to investment in low-capital, less expensive minor interventions.

Despite capturing many facets of overtreatment in expert services, our model is a stylized model that is subject to limitations. For instance, some of the treatment fraud can be ameliorated (but not eliminated, because overtreatment incentives do not disappear) by seeking second opinions. In the empirical data from the health care industry, significant overtreatments have been observed, despite the presence of second opinion options.

If consumer access-to-service is a significant concern, as in health care settings, the social welfare maximization does not necessarily improve access for all customer problems. The lack of access is further exacerbated when the arrival rate (i.e., capacity utilization) is high, and the incidence of complex problems become predominant. Overtreatments are also costly in several other ways. The social after-effects of overtreatment are deeply felt in other medical side-effects, long-term care expenses and suicide rates (Schairer et al 2006). Finally, the pricing issues in insurance-related aspects of the health-care industry are complex. We hope that our model is a step towards more

theoretical or empirical investigations into unraveling the complexity.

References

Afèche, P. 2013. Incentive-Compatible Revenue Management in Queueing Systems: Optimal Strategic Delay. *Manufacturing & Service Operations Management*. 15(3), 423–443.

Alger, I., F. Salanie. 2006. A theory of fraud and over-treatment in expert markets. *Journal of Economic Management Strategy*. **15**(4), 853–881.

Alizamir, S., F. de Vericourt, P. Sun. 2013. Diagnostic Accuracy Under Congestion. *Management Science*. **59**(1), 157–171.

Anand, K., M.F. Paç, S. Veeraraghavan. 2011. Quality-Speed Conundrum: Tradeoffs in Customer Intensive Services. *Management Science*. 57(1). 40–56.

Bleyer, A., Welch H. G. 2012. Effect of three decades of screening mammography on breast-cancer incidence. *New Engl J. Medicine*. 367(21). pp. 1998–2005.

Dai, T., M. Akan, S. Tayur. 2012. Imaging Room and Beyond: The Underlying Economics Behind Physicians' Test-Ordering Behavior in Outpatient Services. Johns Hopkins Working Paper.

Darby, M. R., E. Karni. 1973. Free competition and the optimal amount of fraud. *Journal of Law and Economics*, **16**, pp. 67–88.

de Vericourt, F., Y. Zhou. 2005. Managing Response Time in a Call-Routing Problem with Service Failure. *Operations Research*. **53**(6), pp. 968–981.

Debo, L. G., L.B. Toktay, L.N. Van Wassenhove. 2009. Queuing for Expert Services. *Management Science*. **54**(8), 1497–1512.

Delattre, E., B. Dormont. 2003. Fixed Fees and Physician Induced Demand: A Panel Study on French Physicians. *Health Economics*. **12**, 741–754.

Dranove, D. 1988. Demand Inducement and the Physician Patient Relationship. *Economic Inquiry*. **26**(2), 281–298.

Dulleck, U., R. Kerschbamer. 2006. On doctors, mechanics and computer specialists: The economics of credence goods. *Journal of Economics Literature*. **44**(1), 5–42.

Edelson, N.M., D.K. Hildebrand. 1975. Congestion Tolls for Poisson Queuing Process. *Econometrica*, **43**(1), 81–92.

Emons, W. 1997. Credence Goods and Fraudulent Experts. *RAND Journal of Economics*, **28**, 107–119.

Emons, W. 2001. Credence Goods Monopolists. *International Journal of Industrial Organizations*, **19**, 375–389.

Fong, Y., 2005. When do experts cheat and whom do they target? *RAND Journal of Economics*, **36**(1), 113–130.

Gawande, A., 2015. Overkill: America's Epidemic of Unnecessary Care. *The New Yorker*, Issue: May 11, 2015.

Glazer, A., R. Hassin. 1983. The Economics of Cheating in the Taxi Market. *Transportation Research*. **17A**(1), 25–31.

Gruber, J., M. Owings. 1996. Physician Financial Incentives and Cesarean Delivery. *The RAND Journal of Economics*. **27**(1), 99–123.

Ha, A. Y. 2001. Optimal Pricing That Coordinates Queues with Customer-Chosen Service Requirements. *Management Science*, **47**(7), 915–930.

Hasija, S., E. Pinker, R.A. Shumsky. 2009. Work Expands to Fill the Time Available: Capacity Estimation and Staffing under Parkinson's Law. *Manufacturing and Service Operations Management*, **12**(1), 1–18.

Hassin, R., M. Haviv. 2003. To Queue or not to Queue: Equilibrium behavior in queuing systems. Kluwer Academic Publishers, Norwell, MA.

Kostami, V., S. Rajagopalan. 2014. Speed Quality Tradeoffs in a Dynamic Model. *Manufacturing and Service Operations Management*, **16**(1), pp. 104–118.

Lee, H., E. Pinker, R. Shumsky. 2012. Outsourcing a Two-Level Service Process. *Management Science*, **58**(8). pp. 1569–1584.

Naor, P. 1969. The Regulation of Queue Sizes by Levying Tolls. *Econometrica*, **37**(1), 15–24.

Ong, M., Mandl, K. D. 2015. Overdiagnoses Estimated At $4 Billion A Year National Expenditure For False-Positive Mammograms And Breast Cancer. Health Affairs, 34(4) pp. 576–583.

Pesendorfer, W., A. Wolinsky. 2003. Second opinions and price competition: Inefficiency in the market for expert advice. *Review of Economic Studies*. **70**(2), 417–437.

Pitchik, C., A. Schotter. 1987. Honesty in a Model of Strategic Information Transmission. *American Economic Review*. **77**, 1032–1036.

Schairer C, Brown LM, Chen BE, Howard R, Lynch CF, Hall P. 2006. Suicide after breast cancer: an international population-based study of 723,810 women. J Natl Cancer Inst. 98(19):1416–1419.

Schneider, H. 2012. Agency Problems and Reputation in Expert Services: Evidence From Auto Repair. *The Journal of Industrial Economics.* 60(3), pp. 406–433.

USPSTF. 2009. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 151(10): 716–26.

Wang, X., L. Debo, A. Scheller-Wolf, S. Smith. 2010. Design and Analysis of Diagnostic Service Centers. *Management Science*, 56(11). 1873–1890.

Wennberg, J. E., B. A. Barnes, M. Zubkoff. 1982. Professional Uncertainty and the Problem of Supplier-Induced Demand. *Social Sciences and Medicine*. **16**, 811–824.

Welch, H.G., Black, W.C. 2010. Overdiagnosis in cancer. *J Natl Cancer Inst.* 102(9), pp. 605-613.

Welch, H.G., Passow H.J. 2014. Quantifying the benefits and harms of screening mammography. *JAMA Intern Med.* 174(3), pp. 448–454.

Wolinsky, A. 1993. Competition in a market for informed experts' services. *RAND Journal of Economics*. **24**, 380–398.

Appendix: Proofs

Proof of Lemma 1: Plugging in $\chi = \theta$, $\beta_L = \frac{\lambda_L}{(1-\theta)\Lambda}$ and $\beta_H = \frac{\lambda_H}{\theta\Lambda}$ into equation (4), we can write the social welfare function as:

$$SW(\lambda_L, \lambda_H) = \lambda_L(v_L - c\tau_L) + \lambda_H(v_H - \tau_H) - c\frac{(\lambda_L + \lambda_H)(\lambda_L\tau_L^2 + \lambda_H\tau_H^2)}{(1 - \lambda_L\tau_L - \lambda_H\tau_H)}. \tag{18}$$

1. For a given λ_H, $SW(\lambda_L, \lambda_H)$ is concave in λ_L as the second order condition $\frac{\partial^2 SW(\lambda_L,\lambda_H)}{\partial\lambda_L^2} = -\frac{2c\tau_L^2}{1-\lambda_L\tau_L-\lambda_H\tau_H} - \frac{2c\tau_L((\lambda_L\tau_L^2+\lambda_H\tau_H^2)+\tau_L^2(\lambda_L+\lambda_H))}{(1-\lambda_L\tau_L-\lambda_H\tau_H)^2} - \frac{2c\tau_L^2(\lambda_L+\lambda_H)(\lambda_L\tau_L^2+\lambda_H\tau_H^2)}{(1-\lambda_L\tau_L-\lambda_H\tau_H)^3} < 0$, for $(\lambda_L\tau_L + \lambda_H\tau_H) < 1$ which is for stability.

   To determine the optimal admission rate, $\lambda_L^*(\lambda_H)$, we first find that for $1 > (\lambda_L\tau_L + \lambda_H\tau_H)$, the first order condition,
   $0 = \frac{\partial SW(\lambda_L,\lambda_H)}{\partial\lambda_L} = v_L - c\tau_L - c\frac{\lambda_L\tau_L^2+\lambda_H\tau_H^2+\tau_L^2(\lambda_L+\lambda_H)}{1-\lambda_L\tau_L-\lambda_H\tau_H} - c\tau_L\frac{(\lambda_L+\lambda_H)(\lambda_L\tau_L^2+\lambda_H\tau_H^2)}{(1-\lambda_L\tau_L-\lambda_H\tau_H)^2}$, is uniquely satisfied at $\lambda_L^* = \frac{1-\lambda_H\tau_H}{\tau_L} - \sqrt{\frac{c\gamma_L(\lambda_H)}{v_L\tau_L}}$.

   The social welfare function $SW(\lambda_L, \lambda_H)$ is sub-modular in (λ_L, λ_H) for $1 > (\lambda_L\tau_L + \lambda_H\tau_H)$, i.e.,

$$0 < \frac{\partial^2 SW(\lambda_L, \lambda_H)}{\partial\lambda_L\partial\lambda_H} = -\frac{c(\tau_L^2 + \tau_H^2)}{1 - \lambda_L\tau_L - \lambda_H\tau_H} - \frac{c(\tau_H + \tau_L)(\lambda_L\tau_L^2 + \lambda_H\tau_H^2) + c(\tau_L\tau_H)(\lambda_L + \lambda_H)(\tau_L + \tau_H)}{(1 - \lambda_L\tau_L - \lambda_H\tau_H)^2}$$
$$- \frac{2c\tau_L\tau_H(\lambda_L + \lambda_H)(\lambda_L\tau_L^2 + \lambda_H\tau_H^2)}{(1 - \lambda_L\tau_L - \lambda_H\tau_H)^3}.$$

   Therefore, λ_L^* is decreasing in λ_H.

   There exists a $\widehat{\lambda}_H \in (0, 1/\tau_H)$ such that $\lambda_L^*(\lambda_H) = 0$ for $\lambda_H \geq \widehat{\lambda}_H$, since $\frac{\partial SW(0,0)}{\partial\lambda_L} = v_L - c\tau_L > 0$, $\frac{\partial SW(0,1/\tau_H)}{\partial\lambda_L} = -\infty$ and $\frac{\partial SW(\lambda_L,\lambda_H)}{\partial\lambda_L}$ is decreasing in λ_H due to submodularity.

Note that $SW(\lambda_L, \lambda_H)$, is symmetric in λ_L and λ_H. Therefore, the proof of Lemma 1.2 is similar to that of Lemma 1.1.

Proof of Lemma 2:

1. Plugging $\lambda_L^*(\lambda_H)$ into equation (18) and differentiating, we get

$$\frac{d^2SW(\lambda_L^*(\lambda_H),\lambda_H)}{d\lambda_H^2} = \begin{cases} \frac{\sqrt{cv_L}(\tau_H^2-\tau_L^2)^2}{2\tau_L(\lambda_H(1-\lambda_H\tau_H)(\tau_H-\tau_L)^2+\tau_L)^{3/2}} & \text{if } \lambda_H \le \widehat{\lambda}_H \\ \frac{-2c\tau_H^2}{(1-\lambda_H\tau_H)^3} & \text{if } \lambda_H > \widehat{\lambda}_H \end{cases} \tag{19}$$

Note that $1 > \lambda_H\tau_H$ since the system utilization must be below one, and, $\tau_H > \tau_L$ since the major treatments take longer than the minor treatments on average. Therefore, the second order derivative is ≥ 0 for $\lambda_H \le \widehat{\lambda}_H$ and < 0 for $\lambda_H > \widehat{\lambda}_H$, which proves the desired result.

2. Similarly, plugging $\lambda_H^*(\lambda_L)$ into equation (18) and differentiating, we get

$$\frac{d^2SW(\lambda_L,\lambda_H^*(\lambda_L))}{d\lambda_L^2} = \begin{cases} \frac{\sqrt{cv_H}(\tau_H^2-\tau_L^2)^2}{2\tau_H(\lambda_L(1-\lambda_L\tau_L)(\tau_H-\tau_L)^2+\tau_H)^{3/2}} & \text{if } \lambda_L \le \widehat{\lambda}_L \\ \frac{-2c\tau_L^2}{(1-\lambda_L\tau_L)^3} & \text{if } \lambda_L > \widehat{\lambda}_L \end{cases} \tag{20}$$

proving the desired result when $\tau_H > \tau_L$ and $\lambda_L\tau_L < 1$.

Proof of Lemma 3:

1. To prove the result, we first show that social welfare $SW(\lambda_L,\lambda_H^*(\lambda_L))$ is maximized at $\lambda_L^*(0) = \frac{1}{\tau_L} - \sqrt{\frac{c}{v_L\tau_L}}$ for $\frac{v_H}{\tau_H} < \frac{v_L}{\tau_L}$:

   - $SW(0,\lambda_H^*(0)) = \left(\sqrt{\frac{v_H}{\tau_H}} - \sqrt{c}\right)^2 < \left(\sqrt{\frac{v_L}{\tau_L}} - \sqrt{c}\right)^2 = SW(\lambda_L^*(0),0)$.
   - $\widehat{\lambda}_L < \lambda_L^*(0)$ since,
     (i) $\frac{\partial SW(\lambda_L^*(0),0)}{\partial\lambda_H} = -\frac{c(\tau_H-\tau_L)^2(\sqrt{v_Lc\tau_L}-c\tau_L)+c\tau_L(v_L\tau_H-v_H\tau_L)}{c\tau_L^2} < 0 \left(\frac{v_L}{\tau_L} \ge \frac{v_H}{\tau_H}\right)$ and $v_L > c\tau_L$,
     (ii) $SW(\lambda_L,\lambda_H)$ is concave in λ_H for a given λ_L (Lemma 1), and
     (iii) $SW(\lambda_L,\lambda_H)$ is submodular in (λ_L,λ_H) (Lemma 1).

   From Lemma 1 we have that $SW(\lambda_L,\lambda_H^*(\lambda_L)$ is convex in λ_L for $\lambda_L < \widehat{\lambda}_L$ and concave in λ_L for $\lambda_L \ge \widehat{\lambda}_L$. Therefore, $SW(\lambda_L,\lambda_H^*(\lambda_L))$ can be maximized at either the corner point $\lambda_L = 0$ or at the local maximum $\lambda_L = \lambda_L^*(0)$. Since $SW(0,\lambda_H^*(0)) < SW(\lambda_L^*(0),0)$, $SW(\lambda_L,\lambda_H)$ is maximized at $(\lambda_L^*(0),0)$.

   To prove the existence of a unique saddle point it suffices to show that $SW(\lambda_L,\lambda_H^*(\lambda_L))$ has an interior global minimum (at $\lambda_L = \underline{\lambda_L}$). $(\underline{\lambda_L},\underline{\lambda_H})$ is the unique saddle point for $SW(\lambda_L,\lambda_H)$ since;

   • The first order condition is satisfied for $SW(\lambda_L,\lambda_H)$ at $(\underline{\lambda_L},\underline{\lambda_H})$:

$$\frac{dSW(\lambda_L,\lambda_H^*(\lambda_L))}{d\lambda_L} = \frac{\partial SW(\lambda_L,\lambda_H^*(\lambda_L))}{\partial\lambda_L} + \frac{\partial SW(\lambda_L,\lambda_H^*(\lambda_L))}{\partial\lambda_H}\frac{\partial\lambda_H^*(\lambda_L)}{\partial\lambda_L} = 0$$

$$\Rightarrow \frac{\partial SW(\lambda_L, \lambda_H^*(\lambda_L))}{\partial \lambda_L} = 0.$$

• $\underline{\lambda_L}$ is a local maximizer for $SW(\lambda_L, \lambda_H)$ for $\underline{\lambda_H}$ since $SW(\lambda_L, \lambda_H)$ is concave in λ_L for a given λ_H,

• $\underline{\lambda_H}$ is a local maximizer for $SW(\lambda_L, \lambda_H)$ for a given $\underline{\lambda_H}$ since $SW(\lambda_L, \lambda_H)$ is concave in λ_H for a given λ_L.

Therefore, if $SW(\lambda_L, \lambda_H^*(\lambda_L))$ has a local minimizer in λ_L at $\underline{\lambda_L}$, then $(\underline{\lambda_L}, \lambda_H^*(\underline{\lambda_L}))$ is a saddle point for $SW(\lambda_L, \lambda_H)$.

Since $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is convex in λ_L for $\lambda_L < \widehat{\lambda}_L$ and concave in λ_L for $\lambda_L > \widehat{\lambda}_L$, $SW(\lambda_L, \lambda_H^*(\lambda_L))$ has an interior global minimum only if $\frac{dSW(\lambda_L, \lambda_H^*(\lambda_L))}{d\lambda_L}$ is negative at $\lambda_L = 0$.

$$\left.\frac{dSW(\lambda_L, \lambda_H^*(\lambda_L))}{d\lambda_L}\right|_{\lambda_L=0} = \frac{\sqrt{v_H c \tau_H}(v_L \tau_H - v_H \tau_L) - c(\tau_H - \tau_L)^2(v_H - \sqrt{v_H c \tau_H})}{\tau_H \sqrt{v_H c \tau_H}} < 0$$
$$\Leftrightarrow \quad \frac{v_L}{\tau_L} < \frac{v_H}{\tau_H} + \frac{c(v_H \sqrt{v_H c \tau_H})(\tau_H - \tau_L)^2}{\tau_L \tau_H \sqrt{c v_H \tau_H}},$$

which proves the desired result.

2. If $SW(\lambda_L, \lambda_H^*(\lambda_L))$ has an interior minimizer in λ_L, then there exists a unique $\widetilde{\lambda}_L > \underline{\lambda_L}$ such that $SW(\widetilde{\lambda}_L, \lambda_H^*(\widetilde{\lambda}_L)) = SW(0, \lambda_H^*(0))$ since, (i) $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is convex and decreasing in λ_L at $\lambda_L = 0$, and (ii) The global maximum is $SW(\lambda_L^*(0), 0)$ at $\lambda_L^*(0) > \underline{\lambda_L}$.

Lemma A1. *If* $\frac{v_H}{\tau_H} \leq \frac{v_L}{\tau_L} < \Delta$ *then:*
a. $\widetilde{\Lambda}_H(\bar{\theta}\Lambda)$ *is increasing in* Λ *and decreasing in* θ *for* $\bar{\theta}\Lambda \leq \widetilde{\lambda}_L$.
b. *For* $\bar{\theta} \in [\underline{\lambda_L}, \widetilde{\lambda}_L]$, $SW(\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda)) \leq SW(\lambda_L^*(\theta\Lambda), \theta\Lambda)$ *if and only if* $\widetilde{\Lambda}_H(\bar{\theta}\Lambda) \leq \theta\Lambda \leq \lambda_H^*(0)$.

Proof of Lemma A1:
a. For $\lambda_L \in [\underline{\lambda_L}, \widetilde{\lambda}_L]$, $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is convex and increasing in λ_L. Therefore, $SW(\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda))$ is increasing in Λ and decreasing in θ. $SW(\lambda_L^*(\lambda_H), \lambda_H)$ is increasing in λ_H for $\lambda_H \in [\underline{\lambda_H}, \lambda_H^*(0)]$. Therefore, $\widetilde{\Lambda}_H(\bar{\theta}\Lambda)$ is increasing in Λ and $\bar{\theta}$.
b. For $\lambda_L \in [\underline{\lambda_L}, \widetilde{\lambda}_L]$, $SW(\lambda_L, \lambda_H^*(\lambda_L)) \leq SW(0, \lambda_H^*(0))$. $SW(\lambda_L^*(\lambda_H), \lambda_H)$ is increasing in λ_H for $\lambda_H \in [\widetilde{\Lambda}_H(\bar{\theta}\Lambda), \lambda_H^*(0)]$ which proves the desired result.

Proof of Proposition 1:

The social planner's objective function is given by:

$$\max_{\{0 \leq \lambda_H \leq \theta\Lambda, 0 \leq \lambda_L \leq \bar{\theta}\Lambda\}} \{SW(\lambda_L, \lambda_H)\}. \tag{21}$$

1. When $\Delta < \frac{v_L}{\tau_L} - \frac{v_H}{\tau_H}$, from Lemmas 1 and 3, we know:
   $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is convex and increasing in λ_L for $\lambda_L \in [0, \widehat{\lambda}_L)$ and concave in λ_L for $\lambda_L \geq \widehat{\lambda}_L$, and attains its maximum at $\lambda_L^*(0)$. As a result, $SW(\lambda_L, \lambda_H^*(\lambda))$, is increasing in $\lambda_L \in [0, \lambda_L^*(0)]$ and decreasing in $\lambda_L > \lambda_L^*(0)$ when $\frac{v_L}{\tau_L} \geq \Delta$ (from Lemma 3).

   Case 1: $\bar{\theta}\mathbf{\Lambda} > \widehat{\lambda}_\mathbf{L}$:
   From Lemma 1, we know that $\lambda_H^*(\lambda_L) = 0$ for $\lambda_L > \widehat{\lambda}_L$ and $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is increasing in λ_L for $\lambda_L \in [0, \lambda_L^*(0)]$. Therefore, the objective function in (21) is maximized by $(\lambda_L^S, \lambda_H^S) = (\bar{\theta}\Lambda, 0)$ for $\widehat{\lambda}_L \leq \bar{\theta}\Lambda < \lambda_L^*(0)$.

   If the potential demand for minor problems, $\bar{\theta}\Lambda \geq \lambda_L^*(0)$, then social welfare is maximized at the interior point $(\lambda_L^S, \lambda_H^S) = (\lambda_L^*(0), 0)$. Thus, $(\lambda_L^S, \lambda_H^S) = (\min(\bar{\theta}\Lambda, \lambda_L^*(0)), 0)$.

   Case 2: $\bar{\theta}\mathbf{\Lambda} \leq \widehat{\lambda}_\mathbf{L}$:
   Again Lemma 1, we know that $\lambda_H^*(\lambda_L) > 0$ and decreasing in λ_L for $\lambda_L < \widehat{\lambda}_L$. $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is increasing in λ_L for $\lambda_L \in [0, \lambda_L^*(0)]$. Therefore, the objective function is maximized by $(\lambda_L^S, \lambda_H^S) = (\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda))$ for $\bar{\theta}\Lambda \leq \widehat{\lambda}_L$ and $\theta\Lambda \geq \lambda_H^*(\bar{\theta}\Lambda)$.

   If $\theta\Lambda < \lambda_H^*(\bar{\theta}\Lambda)$, then the optimal policy is $(\lambda_L^S, \lambda_H^S) = (\bar{\theta}\Lambda, \theta\Lambda)$, since $SW(\lambda_L, \lambda_H)$ is concave in λ_H for a given λ_L and increasing in λ_H for $\lambda_H \leq \lambda_H^*(\lambda_L)$. Thus, $(\lambda_L^S, \lambda_H^S) = (\bar{\theta}\Lambda, \min\{\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda)\})$.

   Thus, combining cases 1 and 2,

$$(\lambda_L^S, \lambda_H^S) \;=\; \begin{cases} \big(\min\{\bar{\theta}\Lambda, \lambda_L^*(0)\}, 0\big) & \text{if } \bar{\theta}\Lambda > \widehat{\lambda}_L \\ \\ \big(\bar{\theta}\Lambda, \min\{\theta\Lambda, \lambda_H^*(\bar{\theta}\Lambda)\}\big) & \text{if } \bar{\theta}\Lambda \leq \widehat{\lambda}_L \end{cases}$$

2. For $\frac{v_H}{\tau_H} \leq \frac{v_L}{\tau_L} < \Delta$, from Lemma 3 we know that $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is convex and decreasing for $\lambda_L \in [0, \underline{\lambda_L}]$, and is convex and increasing in λ_L for $\lambda_L \in [\underline{\lambda_L}, \widehat{\lambda}_L]$ and concave for $\lambda_L \geq \widehat{\lambda}_L$. From Lemmas 1 and 3 we know that $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is minimized at $\underline{\lambda_L}$ and maximized at $\lambda_L^*(0)$, since $SW(\lambda_L^*(0), 0) = \left(\sqrt{\frac{v_L}{\tau_L}} - \sqrt{c}\right)^2 \geq SW(0, \lambda_H^*(0)) = \left(\sqrt{\frac{v_H}{\tau_H}} - \sqrt{c}\right)^2$ when $\frac{v_L}{\tau_L} \geq \frac{v_H}{\tau_H}$.

   Case 1: $\bar{\theta}\mathbf{\Lambda} < \underline{\lambda_\mathbf{L}}$:
   $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is convex and decreasing in λ_L for $\lambda_L \in [0, \underline{\lambda_L})$ for $\frac{v_L}{\tau_L} < \Delta$ from Lemma 3. This implies that $(0, \lambda_H^*(0))$ maximizes $SW(\lambda_L, \lambda_H)$ for $\lambda_L \leq \underline{\lambda_L}$ and $\lambda_H \geq \lambda_H^*(0)$.

   Let us define $\mathcal{F}$ as the pair of points $(\lambda_L, \lambda_H^*(\lambda))$ for $\lambda_L \in [0, \infty)$. We will check the local neighborhood to find the optimal admission policy.
   $\frac{dSW(\lambda_L, \lambda_H^*(\lambda_L))}{d\lambda_L} < 0$ for $\lambda_L \in [0, \underline{\lambda_L})$, implies that $\frac{\partial SW(\lambda_L, \lambda_H^*(\lambda_L))}{\partial \lambda_L} < 0$ for any λ_L on $\mathcal{F}$, since: $\frac{\partial SW(\lambda_L, \lambda_H^*(\lambda_L))}{\partial \lambda_L} = \frac{dSW(\lambda_L, \lambda_H^*(\lambda_L))}{d\lambda_L}$ as $\frac{\partial SW(\lambda_L, \lambda_H^*(\lambda_L))}{\partial \lambda_H} = 0$ on $\mathcal{F}$.

$\frac{\partial SW(\lambda_L,\lambda_H^*(\lambda_L))}{\partial \lambda_L} < 0$ implies that $\lambda_L^*(\lambda_H^*(\lambda_L)) < \lambda_L$ for $\lambda_L \in [0, \underline{\lambda_L})$. As a result, for any point (λ_L, λ_H) above the frontier $\mathcal{F}$, i.e., $\lambda_H > \lambda_H^*(\lambda_L)$, the social welfare $SW(\lambda_L, \lambda_H)$ is decreasing in both λ_L and λ_H, since: (i) $SW(\lambda_L, \lambda_H)$ is univariate concave in both λ_L and λ_H, (ii) $\lambda_L^*(\lambda_H)$ lies below the frontier as $\lambda_L^*(\lambda_H^*(\lambda_L)) < \lambda_L$ and $\lambda_L^*(\lambda_H)$ is decreasing in λ_H and (iii) $\lambda_H^*(\lambda_L)$ is on the frontier. Therefore, the socially optimal admission policy does not lie above the frontier $\mathcal{F}$ for $\lambda_L \in [0, \underline{\lambda_L})$.

For $\lambda_L \leq \underline{\lambda_L}$, on any point $(\lambda_L, \lambda_H^*(\lambda_L))$ on the frontier $\mathcal{F}$, $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is decreasing in λ_L, which implies that $(0, \lambda_H^*(0))$ maximizes $SW(\lambda_L, \lambda_H)$ for $\lambda_L \in [0, \underline{\lambda_L})$. This implies that the socially optimal admission policy $(\lambda_L^S, \lambda_H^S)$ is equal to $(0, \lambda_H^*(0))$ for $\bar{\theta}\Lambda \in [0, \underline{\lambda_L})$ and $\theta\Lambda \geq \lambda_H^*(0)$.

For $\lambda_L \leq \underline{\lambda_L}$, on any point (λ_L, λ_H) below the frontier $\mathcal{F}$, i.e., $\lambda_H < \lambda_H^*(\lambda_L)$, $SW(\lambda_L, \lambda_H)$ is increasing in λ_H since $SW(\lambda_L, \lambda_H)$ is concave in λ_H and the maximizer, $\lambda_H^*(\lambda_L)$, is on the frontier. Also, recall that $\lambda_L^*(\lambda_H^*(\lambda_L)) < \lambda_L$. Therefore, for any point below the frontier there are two potential cases:

(i) $SW(\lambda_L, \lambda_H)$ is increasing in λ_H and decreasing in λ_L for $\lambda_L \geq \lambda_L^*(\lambda_H)$, and,

(ii) $SW(\lambda_L, \lambda_H)$ is increasing in both λ_L and λ_H for $\lambda_L \in [0, \lambda_L^*(\lambda_H))$.

As a result, if $\theta\Lambda < \lambda_H^*(0)$ and $\bar{\theta}\Lambda \in [\lambda_L^*(\theta\Lambda), \underline{\lambda_L}]$, then the socially optimal admission policy, $(\lambda_L^S, \lambda_H^S)$, is $(\lambda_L^*(\theta\Lambda), \theta\Lambda)$. If $\theta\Lambda < \lambda_H^*(0)$ and $\bar{\theta}\Lambda < \lambda_L^*(\theta\Lambda)$, then the socially optimal admission policy, $(\lambda_L^S, \lambda_H^S)$, is $(\bar{\theta}\Lambda, \theta\Lambda)$.

The socially optimal admission policy for case 1 can be written as:

$$(\lambda_L^S, \lambda_H^S) = \begin{cases} (0, \lambda_H^*(0)) & \text{if } \bar{\theta}\Lambda \in [0, \underline{\lambda_L}) \\ & \text{and } \theta\Lambda \geq \lambda_H^*(0) \\ (\lambda_L^*(\theta\Lambda), \theta\Lambda) & \text{if } \bar{\theta}\Lambda \in [\lambda_L^*(\theta\Lambda), \underline{\lambda_L}) \\ & \text{and } \theta\Lambda \in [0, \lambda_H^*(0)) \\ (\bar{\theta}\Lambda, \theta\Lambda) & \text{if } \bar{\theta}\Lambda \in [0, \lambda_L^*(\theta\Lambda)) \\ & \text{and } \theta\Lambda \in [0, \lambda_H^*(0)) \end{cases} \tag{22}$$

Case 2: $\bar{\theta}\Lambda \geq \underline{\lambda_L}$ and $\theta\Lambda > \widetilde{\Lambda}_H(\bar{\theta}\Lambda)$:

Recall that, for $\frac{v_L}{\tau_L} < \Delta$, $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is convex and decreasing in λ_L for $\lambda_L \leq \underline{\lambda_L}$ and increasing in λ_L for $\lambda_L \in [\underline{\lambda_L}, \lambda_L^*(0))$ from Lemmas 1 and 3.

In Lemma 3.2, we define $\widetilde{\lambda}_L$ as the smallest λ_L such that:

$$SW(0, \lambda_H^*(0)) = SW(\widetilde{\lambda}_L, \lambda_H^*(\widetilde{\lambda}_L)).$$

Note that, $\widetilde{\lambda}_L$ has to be within the interval $(\underline{\lambda_L}, \lambda_L^*(0))$, since

$$SW(\underline{\lambda_L}, \lambda_H^*(\lambda_L)) < SW(0, \lambda_H^*(0)) < SW(\lambda_L^*(0), 0),$$

and $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is increasing in λ_L for $\lambda_L \in [\underline{\lambda_L}, \lambda_L^*(0))$ (from Lemmas 1 and 3).

Using the above fact, and the definition of $\widetilde{\Lambda}_H(\bar{\theta}\Lambda)$, we find that Case 2 can occur only if $\bar{\theta}\Lambda \leq \widetilde{\lambda}_L$, since $\widetilde{\Lambda}_H(\bar{\theta}\Lambda) = \infty$ for $\bar{\theta}\Lambda > \widetilde{\lambda}_L$.

We analyze Case 2 under two subcases.

Case 2.a. $\bar{\theta}\mathbf{\Lambda} \in [\underline{\lambda_L}, \widetilde{\lambda}_L]$ and $\theta\mathbf{\Lambda} \geq \lambda_H^*(0)$:
For any $\lambda_L \in [\underline{\lambda_L}, \widetilde{\lambda}_L]$, $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is increasing in λ_L and $SW(\lambda_L, \lambda_H^*(\lambda_L)) \leq SW(0, \lambda_H^*(0))$. Therefore, given $\bar{\theta}\Lambda \in [\underline{\lambda_L}, \widetilde{\lambda}_L]$ and $\theta\Lambda \geq \lambda_H^*(0)$, the socially optimal policy, $(\lambda_L^S, \lambda_H^S)$, is equal to $(0, \lambda_H^*(0))$.

Case 2.b. $\bar{\theta}\mathbf{\Lambda} \in [\underline{\lambda_L}, \widetilde{\lambda}_L]$ and $\widetilde{\mathbf{\Lambda}}_\mathbf{H}(\bar{\theta}\mathbf{\Lambda}) < \theta\mathbf{\Lambda} \leq \lambda_\mathbf{H}^*(\mathbf{0})$:
For this case we need to compare the two local maxima for $SW(\lambda_L, \lambda_H)$, given $\lambda_L \leq \bar{\theta}\Lambda$ and $\lambda_H \leq \theta\Lambda$. From Lemma A1(b) we know that $SW(\lambda_L^*(\theta\Lambda), \theta\Lambda) \geq SW(\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda))$ for this case. Therefore the socially optimal policy is $(\lambda_L^*(\theta\Lambda), \theta\Lambda)$.

We can write the socially optimal admission policy for Case 2 as:

$$(\lambda_L^S, \lambda_H^S) = \begin{cases} (0, \lambda_H^*(0)) & \text{if } \bar{\theta}\Lambda \in [\underline{\lambda_L}, \widetilde{\lambda}_L] \\ & \text{and } \theta\Lambda \geq \lambda_H^*(0) \\ (\lambda_L^*(\theta\Lambda), \theta\Lambda) & \text{if } \bar{\theta}\Lambda \in [\underline{\lambda_L}, \widetilde{\lambda}_L] \\ & \text{and } \theta\Lambda \in [\widetilde{\Lambda}_H(\bar{\theta}\Lambda), \lambda_H^*(0)) \end{cases} \tag{23}$$

Case 3: $\bar{\theta}\mathbf{\Lambda} \geq \underline{\lambda_\mathbf{L}}$ and $\theta\mathbf{\Lambda} \leq \widetilde{\mathbf{\Lambda}}_\mathbf{H}(\bar{\theta}\mathbf{\Lambda})$:

From Lemmas 1 and 3, we know that $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is convex and decreasing in λ_L for $\lambda_L \in [0, \underline{\lambda_L})$, increasing in λ_L for $\lambda_L \in (\underline{\lambda_L}, \lambda_L^*(0))$ and decreasing in λ_L for $\lambda_L > \lambda_L^*(0)$. Furthermore, when $\frac{v_L}{\tau_L} > \frac{v_H}{\tau_H}$, $(\lambda_L^*(0), 0)$ is the global maximizer of $SW(\lambda_L, \lambda_H)$ since: (i) it satisfies the first order conditions and (ii) $SW(\lambda_L^*(0), 0) = \left(\sqrt{\frac{v_L}{\tau_L}} - \sqrt{c}\right)^2 > SW(0, \lambda_H^*(0)) = \left(\sqrt{\frac{v_H}{\tau_H}} - \sqrt{c}\right)^2$. This implies that, for $\bar{\theta}\Lambda \geq \lambda_L^*(0)$ the socially optimal admission policy, $(\lambda_L^S, \lambda_H^S)$, is $(\lambda_L^*(0), 0)$.

Lemmas 1 and 3 imply that, for $\lambda_L \in (\underline{\lambda_L}, \lambda_L^*(0))$, $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is increasing in λ_L. Therefore, on the frontier $\mathcal{F}$, $SW(\lambda_L, \lambda_H)$ is increasing in λ_L, since $\frac{\partial SW(\lambda_L, \lambda_H^*(\lambda_L))}{\partial \lambda_L} = \frac{dSW(\lambda_L, \lambda_H^*(\lambda_L))}{d\lambda_L}$. This implies that $\lambda_L^*(\lambda_H^*(\lambda_L)) > \lambda_L$. As a result, on any point (λ_L, λ_H), below the frontier $\mathcal{F}$, i.e., $\lambda_H < \lambda_H^*(\lambda_L)$, $SW(\lambda_L, \lambda_H)$ is increasing in both λ_L and λ_H. Hence, for $\bar{\theta}\Lambda \in (\underline{\lambda_L}, \lambda_L^*(0))$ and $\theta\Lambda \leq \lambda_H^*(\bar{\theta}\Lambda)$, the socially optimal admission policy, $(\lambda_L^S, \lambda_H^S)$, is $(\bar{\theta}\Lambda, \theta\Lambda)$.

Above the frontier $\mathcal{F}$, $SW(\lambda_L, \lambda_H)$ is decreasing in λ_H, since $\lambda_H^*(\lambda_L)$ lies on the frontier, and $SW(\lambda_L, \lambda_H)$ is univariate concave in λ_H. Which implies that the local maximum is on the frontier when $\bar{\theta}\Lambda \in (\underline{\lambda_L}, \lambda_L^*(0))$ and $\theta\Lambda \geq \lambda_H^*(\bar{\theta}\Lambda)$. Since $SW(\lambda_L, \lambda_H^*(\lambda_L))$ is increasing in λ_L for this region, $(\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda))$ is a local maximizer.

Lemma A1(b) implies that $SW(\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda)) > SW(\lambda_L^*(\theta\Lambda), \theta\Lambda)$ for $\theta\Lambda < \widetilde{\Lambda}_H(\bar{\theta}\Lambda)$, which implies that $(\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda))$ is the global maximizer of $SW(\lambda_L, \lambda_H)$ when $\bar{\theta}\Lambda \in (\underline{\lambda_L}, \lambda_L^*(0))$ and $\theta\Lambda \in [\lambda_H^*(\bar{\theta}\Lambda), \widetilde{\Lambda}_H(\bar{\theta}\Lambda)]$.

We can write the socially optimal admission policy as:

$$(\lambda_L^S, \lambda_H^S) = \begin{cases} (\bar{\theta}\Lambda, \theta\Lambda) & \text{if } \bar{\theta}\Lambda \in [\underline{\lambda_L}, \lambda_L^*(0)] \\ & \text{and } \theta\Lambda \leq \lambda_H^*(\bar{\theta}\Lambda) \\ (\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda)) & \text{if } \bar{\theta}\Lambda \in [\underline{\lambda_L}, \lambda_L^*(0)] \\ & \text{and } \theta\Lambda \in [\lambda_H^*(\bar{\theta}\Lambda), \widetilde{\Lambda}_H(\bar{\theta}\Lambda)) \\ (\lambda_L^*(0), 0) & \text{if } \bar{\theta}\Lambda \geq \lambda_L^*(0) \end{cases} \tag{24}$$

Proof of Lemma 4: In order to find $\bar{\lambda}_H(\Lambda)$ we first find the honesty constraints for the experts. Recall that, given prices (p_L, p_H) under strategy (χ, β) the experts revenue is given by:

$$R(p_L, p_H, \Lambda, \chi, \beta) = p_L\beta(1-\chi)\Lambda + p_H\beta_H\chi\Lambda.$$

In order for the expert not to cheat, i.e., recommend unnecessary major services, we need the revenues to be non-increasing in χ. Hence we need:

$$\frac{\partial R(p_L, p_H, \Lambda, \chi, \beta)}{\partial \chi} = \Lambda(-p_L\beta_L + p_H\beta_H) \leq 0.$$

Hence the honesty constraint for the expert is $p_L\beta_L \geq p_H\beta_H$. Note that under an honest diagnostic strategy $\chi = \theta$. Therefore, $\beta_H = \frac{\lambda_H}{\theta\Lambda}$ and $\beta_L = \frac{\lambda_L}{(1-\theta)\Lambda}$. Hence we can write the honesty constraint as:

$$\frac{p_L\lambda_L}{(1-\theta)\Lambda} \geq \frac{p_H\lambda_H}{\theta\Lambda}.$$

$p_H > p_L$ implies $\beta_H < \beta_L \leq 1$, due to the honesty constraint. This implies that consumers with major problems have to be indifferent between joining and not joining the queue, hence earn their reservation utilities of 0 from entering the queue. As a result, the only price that satisfies the honesty constraint given (λ_L, λ_H) is $\bar{p}_H(\lambda_L, \lambda_H)$.

We can rewrite the honesty constraint as:

$$\frac{p_L\lambda_L}{1-\theta} \geq \frac{\bar{p}_H(\lambda_L,\lambda_H)\lambda_H}{\theta}.$$

Note that, for $p_L < \bar{p}_L(\lambda_L,\lambda_H)$, the left hand side of the above constraint is increasing in p_L, which implies that the expert can increase his revenues without violating the honesty constraint by raising the price p_L to $\bar{p}_L(\lambda_L,\lambda_H)$. The honesty constraint for a given that the expert serves all minor problem consumers, $\lambda_L = (1-\theta)\Lambda$, is then given by:

$$\frac{\theta\Lambda}{\lambda_H} \geq \frac{\bar{p}_H((1-\theta)\Lambda,\lambda_H)}{\bar{p}_L((1-\theta)\Lambda,\lambda_H)}.$$

The left hand side of the above constraint is decreasing in λ_H and the right hand side of the above inequality is increasing in λ_H, since, $\frac{\bar{p}_H(\lambda_L,\lambda_H)}{\bar{p}_L(\lambda_L,\lambda_H)} = 1 + \frac{(v_H-v_L)-c(\tau_H-\tau_L)}{v_L-cW_q\left(\theta,1,\frac{\lambda_H}{\theta\Lambda}\right)}$ and the waiting time in the queue, W_q, is increasing in λ_H. As a result, the gap in the above constraint shrinks with increasing λ_H and the constraint binds at the maximum number of major problem consumers that can be served under honesty, $\bar{\lambda}_H(\Lambda)$, which proves part 1 of the Lemma.

2. With increasing θ, the left hand side of the honesty constraint, $\frac{\theta\Lambda}{\bar{\lambda}_H(\Lambda)} = \frac{\bar{p}_H((1-\theta)\Lambda,\bar{\lambda}_H(\Lambda))}{\bar{p}_L((1-\theta)\Lambda,\bar{\lambda}_H(\Lambda))}$, increases, while the right hand side decreases since increasing θ leads to lower λ_L and $W_q(\cdot)$ is increasing in λ_L. Increasing $\bar{\lambda}_H(\Lambda)$, decreases the left hand side of the above constraint, while increasing the right hand side as $W_q(\cdot)$ is increasing in λ_H. Together the above facts imply that $\bar{\lambda}_H(\Lambda)$ is increasing in θ.

3. We first show that first best provision cannot be achieved when the optimal policy is to serve all customers, i.e., $(\lambda_L^S,\lambda_H^S) = ((1-\theta)\Lambda,\theta\Lambda)$.

Recall that, the honesty constraint, $p_L\beta_L \geq p_H\beta_H$, implies that the expert cannot serve all major problem consumers ($\beta_H < 1$) when $p_H > p_L$. Therefore, the only other option to achieve first best provision is to serve all customers under a single price which is equal to $\bar{p}_L((1-\theta)\Lambda,\theta\Lambda)$. However, we show that serving all customers under a single price does not maximize the expert's revenues. If the expert chooses to charge $\bar{p}_H((1-\theta)\Lambda,\bar{\lambda}_H(\Lambda)) > \bar{p}_L((1-\theta)\Lambda,\bar{\lambda}_H(\Lambda))$, then his revenues are equal to $\bar{p}_L((1-\theta)\Lambda,\bar{\lambda}_H(\Lambda))\Lambda$. If the expert chooses to charge a single admission price $\bar{p}_L((1-\theta)\Lambda,\theta\Lambda)$, then his revenues are equal to $\bar{p}_L((1-\theta)\Lambda,\theta\Lambda)\Lambda$. Note that, $\bar{\lambda_H}(\Lambda) < \theta\Lambda$, which implies that $\bar{p}_L((1-\theta)\Lambda,\theta\Lambda) < \bar{p}_L((1-\theta)\Lambda,\bar{\lambda}_H(\Lambda))$, since the prices are decreasing in the cost of waiting and the cost of waiting is increasing in the number of customers joining the queue. Hence, charging a higher price for the major treatment and pricing out some major problem consumers provide higher revenues for the expert, which implies that the first best provision cannot be achieved in equilibrium.

The socially optimal provision can be achieved only under honest service provision as over-provision generates welfare loss. Due to the honesty constraint $\beta_L p_L > \beta_H p_H$, first best provision can be achieved only when $(\lambda_L^S,\lambda_H^S) = (\bar{\theta}\Lambda,\lambda_H^*(\bar{\theta}\Lambda))$ and $\bar{\lambda}_H(\Lambda) > \lambda_H^*(\bar{\theta}\Lambda)$.

We show that there exists a unique threshold $\bar{\Lambda}_s$, such that $\bar{\lambda}_H(\Lambda) \geq \lambda^*_H(\bar{\theta}\Lambda)$ for $\frac{\lambda^*_L(0)}{\bar{\theta}} \geq \Lambda > \bar{\Lambda}_s$. To show this is suffices to prove that $\lambda^*_H(\bar{\theta}\Lambda)$ and $\bar{\lambda}_H(\Lambda)$ cross only once for $\Lambda \in \left[0, \frac{\lambda^*_L(0)}{\bar{\theta}}\right]$. Recall that, $\lambda^*_H(0) > 0$, and $\lambda^*_H(\lambda_L)$ is decreasing in λ_L from Lemma 1. Reordering the honesty constraint in equation 17 as $\bar{\lambda}_H(\Lambda) = \theta\Lambda\frac{\bar{p}_L(\bar{\theta}\Lambda,\bar{\lambda}_H(\Lambda))}{\bar{p}_H(\bar{\theta}\Lambda,\bar{\lambda}_H(\Lambda))}$, and plugging in $\Lambda = 0$, we find that $\bar{\lambda}_H(0) = 0$. Therefore, at $\Lambda = 0$, $\lambda^*_H(\bar{\theta}\Lambda) > \bar{\lambda}_H(\Lambda)$.

$\bar{\lambda}_H(\Lambda) = \frac{\theta\Lambda\bar{p}_L(\bar{\theta}\Lambda,\bar{\lambda}_H(\Lambda))}{\bar{p}_H(\bar{\theta}\Lambda,\bar{\lambda}_H(\Lambda))}$ is increasing in Λ when $\Lambda\bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))$ is increasing in Λ, since, (i) $\bar{p}_H(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))$ is decreasing in Λ, hence increasing Λ increases the right hand side of the above equation, (ii) the left hand side is increasing in $\bar{\lambda}_H(\Lambda)$, and, (iii) the right hand side is decreasing in $\bar{\lambda}_H(\Lambda)$.

For $\lambda_H \leq \lambda^*_H(\lambda_L)$ and $\lambda_L \leq \lambda^*_L(0)$, $\frac{\partial\bar{p}_L(\lambda_L,\lambda_H)}{\partial\lambda_L} = \frac{\partial SW(\lambda_L,\lambda_H)}{\partial\lambda_L} - \frac{\partial(\lambda_H\bar{p}_H(\lambda_L,\lambda_H))}{\partial\lambda_L} > \frac{\partial SW(\lambda_L,\lambda_H)}{\partial\lambda_L}$, since $\bar{p}_H(\lambda_L, \lambda_H)$ is decreasing in λ_L. We are analyzing the region where the socially optimal policy is of the form $(\bar{\theta}\Lambda, \lambda^*_H(\bar{\theta}))$. Recall that, in this region, $SW(\lambda_L, \lambda_H)$ is increasing in λ_L for $\lambda_L \leq \lambda^*_L(0)$ and $\lambda_H \leq \lambda^*_H(\lambda_L)$. This implies that $\Lambda\bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))$ is increasing in Λ for $\bar{\lambda}_H(\Lambda) \leq \lambda^*_H(\bar{\theta}\Lambda)$. As a result, we know that $\bar{\lambda}_H(\Lambda)$ is increasing in Λ upto $\bar{\Lambda}_S$. Above $\bar{\Lambda}_S$, we know that $\bar{\lambda}_H(\Lambda)$ does not cross $\lambda^*_H(\bar{\theta}\Lambda)$ for $\Lambda \leq \frac{\lambda^*_L(0)}{\bar{\theta}}$ since, $\lambda^*_H(\bar{\theta}\Lambda)$ is decreasing in Λ and $\bar{\lambda}_H(\Lambda)$ is increasing in Λ for $\bar{\lambda}_H(\Lambda) \leq \lambda^*_H(\bar{\theta}\Lambda)$.

The above result proves that, the first best provision can be achieved only if $\Lambda \geq \bar{\Lambda}_S$.

Proof of Lemma 5: 1. We first show that for $\theta > \frac{c(\tau_H-\tau_L)}{v_H-v_L}$, as Λ approaches 0, providing false major treatment recommendations, $\chi = 1$, provides higher revenues than providing honest service recommendations, $\chi = \theta$. Also, we know from Lemma 4 that for $\Lambda \geq \bar{\Lambda}_s$ the expert can achieve first best service provision and extract all the welfare through prices, hence maximize his revenues. Therefore, for $\Lambda \geq \bar{\Lambda}_s$ providing honest service recommendations, $\chi = \theta$, provides higher revenues for the expert.

Let us define $\Pi_h(\Lambda)$ as the maximum revenues that the expert can earn when he provides honest service recommendations, $\chi = \theta$, for $\Lambda \leq \bar{\Lambda}_s$.

$$\Pi_h(\Lambda) = \bar{\theta}\Lambda\bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda)) + \bar{\lambda}_H(\Lambda)\bar{p}_H(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda)).$$

Recall that, $\bar{\lambda}_H(\Lambda) = \theta\Lambda\frac{\bar{p}_L(\bar{\theta}\Lambda,\bar{\lambda}_H(\Lambda))}{\bar{p}_H(\theta\Lambda,\bar{\lambda}_H(\Lambda))}$, therefore we can write $\Pi_h(\Lambda)$ as:

$$\Pi_h(\Lambda) = \Lambda\bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda)).$$

As Λ approaches to 0, $\Pi_c(\Lambda) > \Pi_h(\Lambda)$:

$\lim_{\Lambda\to 0}\frac{\Pi_c(\Lambda)}{\Pi_h(\Lambda)} = \frac{\Lambda\bar{p}_C(\Lambda)}{\Lambda\bar{p}_L(\bar{\theta}\Lambda,\bar{\lambda}_H(\Lambda))} = \frac{\bar{p}_C(\Lambda)}{\bar{p}_L(\bar{\theta}\Lambda,\bar{\lambda}_H(\Lambda))}$,

$= \lim_{\Lambda\to 0}\frac{\theta(v_H-v_L)-c(\tau_H-\tau_L)+v_L-c\tau_L-cW_q\left(1,0,\frac{\Lambda}{\Lambda}\right)}{v_L-c\tau_L-cW_q\left(\theta,\frac{\bar{\theta}\Lambda}{\theta\Lambda},\frac{\bar{\lambda}_H(\Lambda)}{\theta\Lambda}\right)} = \frac{\theta(v_H-v_L)-c(\tau_H-\tau_L)+v_L-c\tau_L}{v_L-c\tau_L} > 1$, when $\theta > \frac{c(\tau_H-\tau_L)}{v_H-v_L}$,

since the waiting time in the queue, $W_q(\cdot)$, goes to 0 as $\Lambda \to 0$.

As a result we know that $\Pi_c(0) > \Pi_h(0)$ and $\Pi_c(\bar{\Lambda}_s) < \Pi_h(\bar{\Lambda}_s)$. To complete the proof we need to show that $\Pi_c(\Lambda)$ crosses $\Pi_h(\Lambda)$ only once within the interval $[0, \bar{\Lambda}_s]$.

To prove this we show that $\frac{\partial \Pi_c(\Lambda)}{\partial \Lambda} < \frac{\partial \Pi_h(\Lambda)}{\partial \Lambda}$ for all Λ in the interval. Note that Λ is a common term both in $\Pi_c(\Lambda)$ and $\Pi_h(\Lambda)$, hence it suffices to focus only on $\bar{p}_c(\Lambda)$ and $\bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda_H}(\Lambda))$ and show that $\frac{\partial \bar{p}_c(\Lambda)}{\partial \Lambda} < \frac{\partial \bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda_H}(\Lambda))}{\partial \Lambda}$.

$\frac{\partial \bar{p}_c(\Lambda)}{\partial \Lambda} = \frac{\theta v_H + \bar{\theta} v_L - c\tau_H - cW_q(1,0,1)}{\partial \Lambda} = -c\frac{\partial W_q(1,0,1)}{\partial \Lambda}$.

$\frac{\partial \bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\partial \Lambda} = \frac{v_L + -c\tau_L - cW_q\left(\theta, 1, \frac{\bar{\lambda}_H(\Lambda)}{\theta\Lambda}\right)}{\partial \Lambda} = -c\frac{\partial W_q\left(\theta, 1, \frac{\bar{\lambda}_H(\Lambda)}{\theta\Lambda}\right)}{\partial \Lambda}$. Hence, it suffices to show that:

$$\frac{\partial W_q\left(\theta, 1, \frac{\bar{\lambda}_H(\Lambda)}{\theta\Lambda}\right)}{\partial \Lambda} < \frac{\partial W_q\left(1, 0, 1\right)}{\partial \Lambda}.$$

Using the definition of $W_q(\cdot)$ from equation 4, we find that:

$$W_q\left(\theta, 1, \frac{\bar{\lambda}_H(\Lambda)}{\theta\Lambda}\right) = \frac{\bar{\theta}\Lambda\tau_L^2 + \bar{\lambda}_H(\Lambda)\tau_H^2}{1 - \bar{\theta}\Lambda\tau_L - \bar{\lambda}_H(\Lambda)\tau_H},$$

and,

$$W_q\left(1, 0, 1\right) = \frac{\Lambda\tau_H^2}{1 - \Lambda\tau_H}.$$

We now define the waiting time in the queue as a function of the number of customers receiving major and minor treatment, λ_H and λ_L. Let

$$w_q(\lambda_L, \lambda_H) = \frac{\lambda_H\tau_H^2 + \lambda_L\tau_L^2}{1 - \lambda_L\tau_L - \lambda_H\tau_H}.$$

Note that, $w_q(\cdot)$ is a convex and increasing function of both λ_L and λ_H. Furthermore, $w_q(\cdot)$ is super-modular, i.e., $\frac{\partial^2 w_q(\lambda_L, \lambda_H)}{\partial \lambda_L \partial \lambda_H} > 0$.

Using the definition of $w_q(\cdot)$ we get:

$$W_q\left(\theta, 1, \frac{\bar{\lambda}_H(\Lambda)}{\theta\Lambda}\right) = w_q(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))$$

and

$$W_q\left(1, 0, 1\right) = w_q(0, \Lambda).$$

To prove the result we need to show that $\frac{\partial w_q(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\partial \Lambda} < \frac{\partial w_q(0,\Lambda)}{\partial \Lambda}$.

$\frac{\partial w_q(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\partial \Lambda} = \frac{\partial w_q(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\partial \lambda_L}\frac{\partial(\bar{\theta}\Lambda)}{\partial \Lambda} + \frac{\partial w_q(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\partial \lambda_H}\frac{\partial \bar{\lambda}_H(\Lambda)}{\partial \Lambda}$

$< \frac{\partial w_q(\bar{\theta}\Lambda, \theta\Lambda)}{\partial \Lambda} = \frac{\partial w_q(\bar{\theta}\Lambda, \theta\Lambda)}{\partial \lambda_L}\frac{\partial(\bar{\theta}\Lambda)}{\partial \Lambda} + \frac{\partial w_q(\bar{\theta}\Lambda, \theta\Lambda)}{\partial \lambda_H}\frac{\partial(\theta\Lambda)}{\partial \Lambda} = \frac{\bar{\theta}\tau_L^2 + \theta\tau_H^2}{(1 - \bar{\theta}\Lambda - \theta\Lambda)^2}$, due to the following:

(i) $\frac{\partial w_q(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\partial \lambda_L} < \frac{\partial w_q(\bar{\theta}\Lambda, \theta\Lambda)}{\partial \lambda_L}$ since, $w_q(\lambda_L, \lambda_H)$ is super-modular and $\bar{\lambda}_H(\Lambda) < \theta\Lambda$.

(ii) $\frac{\partial w_q(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\partial \lambda_H} < \frac{\partial w_q(\bar{\theta}\Lambda, \theta\Lambda)}{\partial \lambda_H}$ since, $w_q(\cdot)$ is convex in λ_H and $\bar{\lambda}_H(\Lambda) < \theta\Lambda$, and,

(iii) $\frac{\partial \bar{\lambda}_H(\Lambda)}{\partial \Lambda} < \theta = \frac{\partial(\theta\Lambda)}{\partial \Lambda}$. Since $\frac{\bar{\lambda}_H(\Lambda)}{\theta\Lambda} = \frac{\bar{p}_L}{\bar{p}_H}$ and $\frac{\bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\bar{p}_H(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}$ is decreasing in $W_q(\cdot)$ and $W_q(\cdot)$ is increasing in Λ since $\bar{\lambda}_H(\Lambda)$ is increasing in Λ. $\frac{\partial \bar{\lambda}_H(\Lambda)}{\partial \Lambda} \geq \theta$ implies that $\frac{\bar{\lambda}_H(\Lambda)}{\theta\Lambda}$ is increasing, which implies that $\frac{\bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\bar{p}_H(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}$ is increasing, which is a contradiction. Therefore, $\frac{\partial \bar{\lambda}_H(\Lambda)}{\partial \Lambda} < \theta$.

$\frac{\partial w_q(0,\Lambda)}{\partial \Lambda} = \frac{\tau_H^2}{(1-\Lambda\tau_H)^2} \geq \frac{\bar{\theta}\tau_L^2 + \theta\tau_H^2}{(1-\bar{\theta}\Lambda-\theta\Lambda)^2} > \frac{\partial w_q(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))}{\partial \Lambda}$ for $\theta \in (0,1)$. Which proves the result.

Proof of Proposition 2: 1. From Proposition 1, we know that for $\Lambda \geq \frac{\widehat{\lambda}_L}{1-\bar{\theta}}$, the socially optimal policy is to only admit consumer with minor problems, i.e., $\lambda_H^S = 0$. By setting the price for major treatment, $p_h = \infty$, the expert not only prices out all major type consumers but also credibly signals his honesty to the market, since a consumer receiving a major treatment diagnosis will not join the queue.

For $\Lambda \in \left[\frac{\widehat{\lambda}_L}{1-\bar{\theta}}, \frac{\lambda_L^*(0)}{1-\bar{\theta}}\right]$, the socially optimal admission policy is to admit all minor problem consumers, i.e., $(\lambda_L^S, \lambda_H^S) = (\bar{\theta}\Lambda, 0)$. By charging $\bar{p}_L(\bar{\theta}\Lambda, 0)$ the expert can serve all minor problem consumers and achieve the socially optimal service provision. Since the price $\bar{p}_L(\bar{\theta}\Lambda, 0)$ extracts consumers' welfare, the expert also maximizes revenues.

For $\Lambda > \frac{\lambda_L^*(0)}{1-\bar{\theta}}$, the socially optimal policy is $(\lambda_L^S, \lambda_H^S) = (\lambda_L^*(0), 0)$. The expert can achieve the socially optimal service provision by pricing out all major problem consumers and serving $\lambda_L^*(0)$ consumers with minor problems by charging $\bar{p}_L(\lambda_L^*(0), 0)$. As in the previous case, $\bar{p}_L(\lambda_L^*(0), 0)$ extracts the consumers' welfare and maximizes the expert's revenues.

2. From Lemma 5, we know that the expert is better off by providing honest diagnosis for $\Lambda > \bar{\Lambda}_c$. From Lemma 4, we know that for $\Lambda > \bar{\Lambda}_s$, the expert can achieve the first best service provision.

Therefore, when $\Lambda \in \left[\bar{\Lambda}_s, \frac{\widehat{\lambda}_L}{(1-\bar{\theta})}\right]$, the expert can charge $\bar{p}_L(\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda))$ for the minor treatment, $\bar{p}_H(\bar{\theta}\Lambda, \lambda_H^*(\bar{\theta}\Lambda))$ for the major treatment, and, achieve the socially optimal service provision and maximize his revenues.

For $\Lambda \in \left[\bar{\Lambda}_c, \bar{\Lambda}_s\right]$, from Lemma 4 and Lemma 5 we know that the expert cannot achieve the socially optimal (first best) service provision, but providing honest service recommendations, $\chi = \theta$, is still optimal (from Lemma 5). By charging $\bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))$ and $\bar{p}_H(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))$ the expert achieves the maximum revenues under honest service provision, since $\bar{\lambda}_H(\Lambda)$ is the maximum number of major problem consumers that can be served under honesty ($\chi = \theta$) when the potential demand is Λ, and the revenues are increasing in both the number of minor and major problem consumers served since $\bar{\lambda}_H(\Lambda) < \lambda_H^*(\bar{\theta}\Lambda)$ and $\bar{\theta}\Lambda < \lambda_L^*(0)$. From Lemma 5, we know that the maximum revenues under honest service provision is higher than the maximum revenues under over-provision ($\chi = 1$). Therefore, the expert maximizes his revenues by charging $\bar{p}_L(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))$ for the minor treatment, $\bar{p}_H(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))$ for the major treatment and serving all minor problem consumers, $\bar{\theta}\Lambda$, and $\bar{\lambda}_H(\Lambda)$ major problem consumers.

3. From Lemma 5, we know that over-provision, $\chi = 1$, provides higher revenues for the expert compared to honest service recommendation $\chi = \theta$ when $\Lambda < \bar{\Lambda}_c$. The experts revenue function

under over provision is given by $\Pi_c(\Lambda)$. Note that,

$$\Pi_c(\Lambda) = \mathbb{I}(\Lambda > \lambda_c^*) \left(\sqrt{\frac{\theta v_H + (1-\theta) v_L}{\tau_H}} - \sqrt{c} \right)^2 + \mathbb{I}(\Lambda \leq \lambda_c^*) \Lambda \left((\theta v_H + (1-\theta) v_L) - c W_H(1,0,1) \right)$$

is (i) increasing and concave in Λ for $\Lambda \leq \lambda_c^*$ since $W_H(1,0,1)$ is convex in Λ and λ_c^* is the unique maximizer of $\Lambda \left((\theta v_H + (1-\theta) v_L) - c W_H(1,0,1) \right)$, and, (ii) constant in Λ for $\Lambda > \lambda_c^*$. Given that the expert will choose over-provision, $\chi = 1$, it is optimal to treat all customers if $\Lambda < \lambda_c^*$ and only λ_c^* customers otherwise.

For $\Lambda < \bar{\Lambda}_c$, there are two possibilities:

I. If $\Lambda < \min\{\bar{\Lambda}_c, \lambda_c^*\}$, the revenue maximizing admission rate is Λ. By charging $\bar{p}_c(\Lambda)$ for the major treatment, the expert signals to the market that he will diagnose all customers with a major problem, $\chi = 1$, and treat them with a major treatment, since $\bar{p}_c(\Lambda)$ is less than $\bar{p}_H(\bar{\theta}\Lambda, \bar{\lambda}_H(\Lambda))$ which is the lowest price for the major treatment satisfying the honesty constraint. By charging any p_L that is less than $\bar{p}_c(\Lambda)$ the expert ensures that he will choose not to provide the minor treatment, which implies that $\lambda_L^e = 0$.

II. If $\lambda_c^* < \Lambda < \bar{\lambda}_c$, the revenue maximizing admission rate is λ_c^*. Again, by charging $\bar{p}_c(\lambda_c^*)$ for the major treatment, the expert signals to the market that he will diagnose all customers with a major problem, $\chi = 1$, and treat them with a major treatment. Charging any p_L that is less than $\bar{p}_c(\Lambda)$ ensures that the expert will choose not to provide the minor treatment, which implies that $\lambda_L^e = 0$.