A Model of Generic Drug Shortages: Supply Disruptions, Demand Substitution, and Price Control

Sang-Hyun Kim, Fiona Scott Morton

Yale School of Management, Yale University, 165 Whitney Avenue, New Haven, CT 06511

sang.kim@yale.edu, fiona.scottmorton@yale.edu

January 2015

In recent years, the U.S. health care system has been plagued by shortages of generic sterile injectable drugs. The shortage problem arose suddenly within a short period of time, and remained persistent afterwards. To gain insights into what may have led to this situation, we develop a model that incorporates key characteristics of the generic drug manufacturing industry, including random supply disruptions, demand substitution across different brands, and regulatory control of prices. In our model, firms competitively choose their production capacities anticipating future supply disruptions. These capacity choices in turn determine product availability. Our equilibrium analysis reveals a number of counterintuitive results. First, drug availability *increases* as productions become more prone to disruptions. Second, allowing temporary price increases during shortages may or may not increase drug availability, depending on how severe the shortage is. Based on these insights, we discuss how external factors may have interacted with firm decisions to result in the current drug shortage situation.

1 Introduction

Starting in the late 2000s, there have been significant and persistent shortages of generic drugs in the United States. Number of shortage instances tripled between 2005 and 2010, increasing from 61 to 178 in five years. The majority of the shortages occur in the category of sterile injectable drugs, including cancer drugs, anesthetics, and antibiotics, which accounted for 80% of the shortages that occurred in 2010-2011 despite the fact that they represented only 29% of the entire generics market in volume (U.S. Food and Drug Administration 2011). Because many of these drugs are "medically necessary," shortages can create a serious and even life-threatening situation. Although the situation has shown signs of improvement recently, the number of shortages remains very large (U.S. Government Accountability Office 2014).

These shortages occur mainly because of disruptions in the manufacturing process. It was reported that about half of the shortages in 2010-2011 were due to quality and other production issues (FDA 2011). Notable examples include fungal or bacterial contaminations, introduction of particulate matters in vials, and equipment failures (GAO 2014). Discovery of such issues requires cleanup and restoration, slowing down productions or forcing temporary closure of the entire facility in some cases.

In theory, these disruptions need not result in product shortages; if there are enough excess capacities or inventories that can be utilized as backups, then shortages can be avoided. The fact that shortages have become commonplace indicates that the generics industry's management of capacities has been affected by some external factors.

Responding to the serious nature of the problem, the FDA has instituted a number of remedial procedures that are designed to minimize the impact of shortages. They include: expediting regulatory approval processes, facilitating information sharing among stakeholders, and easing importation of substitutes from overseas (FDA 2011). These measures have been successful to some degree, as the FDA credits them for preventing 38 shortages in 2010 (in the year when 178 shortages were reported). While these efforts help alleviate the impact of shortages, they are stopgap measures that do not address systematic flaw. A permanent solution has proven to be elusive, in spite of numerous published expert opinions, congressional hearings, and a presidential executive order that led to the FDA actions (U.S. House of Representatives 2011, Harris 2011).

The continued shortage situation has prompted many to speculate on the structural causes rooted in market dynamics. Some fault the government's price control policies, characterizing them as arbitrary mechanisms that dampen the firms' incentives to invest in capacity and modernize facilities (see §6 for more discussions). Some blame the FDA, believing that the agency's aggressive regulatory interventions often result in unnecessary production shutdowns and delays (Graham 2012). Others are suspicious about the role of Group Purchasing Organizations (GPO), which act as intermediaries on behalf of hospitals but introduce inefficiencies in the pharmaceutical supply chain. It has also been suggested that an increased number of patent expirations have led the generics manufacturers to shift their limited resources to producing new drugs at the expense of existing ones (U.S. Department of Health and Human Services 2011). Despite many conjectures, none of them has been singled out as the definitive cause of the problem. Instead, it is likely that the problem arises from a combination of factors—economic, operational, and regulatory—that are closely intertwined (FDA 2011).

Among many features of this multifaceted problem, a number of distinctive ones stand out. First, demands for sterile injectable drugs are relatively stable and unresponsive to price changes, because the drugs are medically necessary and insured consumers do not directly bear the price differences. Second, supply disruptions for a given drug occur sporadically, caused by unforeseen events that are often beyond the firms' control. Third, different brands of a given drug are perfect substitutes. Even though some level of differentiation is achieved through packaging and advertising, the underlying products are chemically equivalent and therefore they are clinically indistinguishable. Fourth, the industry is highly concentrated, in most cases with three or fewer competing manufacturers for a

given drug (HHS 2011, FDA 2011). Finally, regulations limit the extent to which drug prices can vary over time. The Medicare Modernization Act of 2003 required that annual price increases be capped at a small percentage, thus creating price stickiness (U.S. House of Representatives 2011).

In this paper we develop a stylized model that captures these unique features and analyze how they interact, generating insights into the dynamics of shortages and the conditions that may exacerbate them. Specifically, we focus on research questions that have important practical implications to the generic drug industry but have not yet been rigorously studied. They are: (1) Knowing that future supply disruptions will reduce production outputs, how do manufacturing firms optimally choose their capacities? (2) How does the "sticky price" policy influence the firms' incentives? (3) What is the net impact of external factors and firm decisions on product availability? (4) Why did the shortage problem arise suddenly during a short period of time, and why has it been so persistent?

The analysis of our model reveals that firms' optimal capacity choices result in an unexpected equilibrium outcome: higher probability of supply disruptions leads to higher product availability. In other words, as productions become more prone to disruptions (i.e., less reliable production process), consumers have a better chance of obtaining the products. This reversal of the usual relationship between reliability and availability also influences how availability changes with other environmental variables. Based on these observations, we propose a hypothesis that explains the current drug shortage situation. We also find that the policy of keeping the price fixed may or may not have a positive effect on product availability; although temporary price increases lower availability in general, there are situations where the opposite is true. We identify the conditions under which this happens, and discuss policy implications.

The rest of the paper is organized as follows. After a brief review of the related literature in §2, in §3 we present model assumptions and derive mathematical expressions that form a basis of the analysis. In §4 we analyze the base model in which product price is assumed to be fixed. This is followed by §5, where we consider an alternative model assumption under which price increases temporarily during shortages. Based on the insights gained from §§4-5, in §6 we propose a hypothesis that may explain the current shortage crisis, concluding in §7. Note that §B.2 and §B.3 in the Appendix contain the results of additional model extensions ("Centralized Firm's Decisions" and "Capacity and Reliability Choices Under Fixed Price") that we refer to throughout the paper.

2 Related Literature

The generic drug industry has been the subject of many studies in economics, as its unique market dynamics provide a fertile ground for testing the theories from the industrial organization literature. Generic versions of a particular drug are introduced to the market after the patent protection enjoyed

by the original drug developer expires, creating opportunities for other pharmaceutical firms to offer chemically equivalent substitutes. These entries usually result in a bifurcated market in which the market share of the branded drug shrinks substantially as price-sensitive buyers opt for its generic counterparts. At the same time, competition among the entrants lowers the price of generics. From an economic perspective, these phenomena can be described by the models of monopoly, market entry and deterrence, vertical differentiation, and oligopolistic competition (Caves et al. 1991, Scherer 1993, Frank and Salkever 1997). These features are further enriched by the fact that the pharmaceutical industry is highly regulated, affecting the firms' incentives and overall efficiency of the market. Over the years researchers have shed light on these issues empirically, such as firms' market entry decisions (Scott Morton 1999), relationship between drug prices and the degree of competition (Reiffen and Ward 2005, Olson and Wendling 2013), and impact of regulations on competition (Danzon and Chao 2000).

Despite the rich body of literature in economics that investigates various aspects of the generic drug industry, very few discuss shortages. This is not surprising, given that the issue has not caught the attention of researchers until recently; although generic drug shortages in the U.S. have always existed to some degree, only in the past few years has the problem reached the status of a national crisis.[1] This development prompted many commentaries from health care researchers and professionals, who diagnose the problem based on conceptual economic arguments (for example, see Jensen and Rappaport (2010), Schweitzer (2013), and Woodcock and Wosinska (2013)). To date, however, there has been a dearth of studies that examine the shortage problem using rigorous analytical frameworks. An exception is an article by Yurukoglu (2012), who develops an econometric model and argues that the current shortage situation is attributed to the implementation of Medicare Modernization Act of 2003, which led to a significant drop in drug prices that "likely reduced capacity and maintenance investments."

Our paper fills this void in the literature. Borrowing the ideas from operations management (OM), we develop a stylized game-theoretic model that captures key characteristics of the generic drug industry and analyze the incentive dynamics under supply disruptions. This approach makes sense especially because supply disruption management has been one of the most active areas of research in OM; for a recent survey of this literature, see Snyder et al. (2014). The majority of works in this area focus on procurement settings in which a buyer has a contractual relationship with one or more suppliers whose production outputs are random due to supply shocks. Representative articles include

[1] A series of *New York Times* articles since 2011 illustrates the persistent and serious nature of the problem: "U.S. Scrambling to Ease Shortage of Vital Medicine" (Harris 2011); "Drug Shortages Persist in U.S., Harming Care" (Thomas 2012); "Drug Shortages Continue to Vex Doctors" (Tavernise 2014).

Tomlin (2006), Dada et al. (2007), Wang et al. (2010), and Yang et al. (2009). Aydin et al. (2012) survey the models of supply disruptions under decentralized decision-making, the category that our paper belongs to. Of these, Deo and Corbett (2009), Tang and Kouvelis (2009), and Kim and Tomlin (2013) are particularly relevant since they consider horizontal competition among firms, similar to what we assume in our paper. These authors, however, focus on other aspects of competition (market entry decisions, correlated disruption risks, etc.) that we do not take into account. In our model, competition is driven by inter-firm demand substitution that occurs after supply disruptions. Not many researchers have studied a similar dynamic, with notable exceptions by Tang and Kouvelis (2009) and Tomlin (2009) who also incorporate demand substitutability but in problem settings quite distinct from ours (centralized decision, sole vs. dual sourcing). The way we represent demand substitution resembles the newsvendor competition model by Netessine and Rudi (2003). Unlike in Netessine and Rudi (2003), however, demand substitution in our model is triggered by supply uncertainty, not by demand uncertainty.

In sum, this paper represents one of the first rigorous studies on the current generic drug shortage situation. We obtain novel insights from the model that combines the ideas from OM and economics that thus far have existed in separate domains, thus contributing to the OM literature on supply disruptions as well as the literature on economics of the pharmaceutical industry.

3 Model

As we previewed in §1, the generic sterile injectable drug industry has a few idiosyncratic characteristics which we highlight in our model. They are: (i) the industry for a given drug typically consists of a small number competing manufacturers; (ii) different brands of a drug are perfectly substitutable; (iii) the demands for a drug are stable over time and exhibit relatively small variability; (iv) production output drops when a random and exogenous shock reduces capacity. To capture these succinctly, we develop a model of duopoly in which two manufacturers facing constant demands and random supply disruptions compete on capacities. The capacity decisions determine product availability as well as the amount of demands that spill over from one firm to the other during disruptions.

Building on this framework, we consider two model variations under different assumptions on price. In §4 we study a base model in which the price is assumed to be fixed. In §5, we modify the base model by assuming that the price varies with total industry capacity when shortages occur. The base model reflects the current government policy that limits short-term price increases. Despite the policy, however, price increases during shortages have been reported, including those from "gray markets" (Link et al. 2012). With the modified model we aim to discover how price variations affect firm decisions and product availability. For expositional clarity, in the remaining space of this section we

describe the model based on fixed-price assumption. Variable-price modification is introduced in §5. We make some simplifying assumptions around this setup in order to construct a parsimonious model that allows tractable analysis.

3.1 Supply Disruptions and Economics

The industry consists of two firms, denoted by the subscript $i \in \{1, 2\}$, that manufacture a homogeneous product whose units are perfectly substitutable (e.g., chemically equivalent cancer drugs). Firm i owns and operates manufacturing facility i. The firms and their facilities have symmetric characteristics, as we elaborate below. The unit price of the product is p. For simplicity, we normalize the unit cost of production to zero and assume that production lead time is negligible. All firms are risk neutral.

Consumer demands for the product arrive at the market over an infinite horizon, starting at time zero. In order to focus on the impact of supply-side uncertainty, we make a simplifying assumption that the demands arrive deterministically at a constant rate of δ per unit time. At each moment, the arriving demands are equally divided and directed to the two firms. Each firm dedicates production capacity at its facility whose rate of output perfectly matches that of the demand stream it receives. We refer to this fully-utilized production capacity as *regular capacity*. Therefore, each facility receives its *allotted demands* at the rate of $\delta/2$ per unit time and produces exactly at that rate using the regular capacity. Without loss of generality, we assume that the costs of setting up capacities are sunk at time zero. In addition to regular capacity, each firm maintains *spare capacity* which is normally idle but may be utilized when production is disrupted. (These distinct capacity types are analogous to "running stocks" and "safety stocks" found in inventory models, e.g., Groenevelt et al. 1992b.) We present the assumptions on disruptions next, followed by spare capacity.

Clearly, all demands are filled when productions on regular capacities run continuously. However, some demands may be lost when a shock arrives at one of the facilities and disables a portion of regular capacity, creating a supply shortage by reducing production output. Examples of such shocks in the drug manufacturing environment include bacterial or particulate contamination of a production line and equipment failure (GAO 2014). We assume that regular capacities at both facilities are subject to independent and repeated shocks. Once a shock arrives, it takes a random amount of time to recover full capacity at each facility. Thus, at any given moment each facility is either in the *undisrupted* state (with full capacity) or in the *disrupted* state (with partial capacity), alternating between the two states as time progresses. Let θ_i be the fraction of time in steady state during which facility i is in the disrupted state; in the remaining $1 - \theta_i$ fraction of time, the facility is in the undisrupted state.[2]

[2] If the states evolve as an alternating renewal process with mean durations $1/\lambda_i$ and $1/\mu_i$, then $\theta_i = \lambda_i/(\lambda_i + \mu_i)$.

Hence, higher θ_i means less reliability of facility i. We refer to θ_i as *disruption probability* at facility i, since it represents the chance that the facility is in the disrupted state at a random point in time.

We assume that the disruption probabilities are sufficiently small so that the approximation $\theta_1\theta_2 \approx 0$ applies, i.e., the probability that both facilities are simultaneously disrupted is negligible. This assumption captures the fact that disruptions are relatively rare events, and it also enables tractable analysis. (A similar assumption is commonly adopted in the spare parts inventory management literature that considers low-probability, high-consequence events; see Sherbrooke (1994) and Muckstadt (2005).) With this approximation, the industry is in one of three states at any given moment: (i) neither facility is disrupted, with probability $(1-\theta_1)(1-\theta_2) \approx 1-\theta_1-\theta_2$; (ii) facility 1 is disrupted and facility 2 is undisrupted, with probability $\theta_1(1-\theta_2) \approx \theta_1$; (iii) facility 1 is undisrupted and facility 2 is disrupted, with probability $(1-\theta_1)\theta_2 \approx \theta_2$.

We assume that, upon arrival, a shock instantaneously reduces the regular capacity at facility i from $\delta/2$ to $\delta/2 \times \epsilon_i$, where $\epsilon_i \in [0,1]$ is a random variable that represents the percentage of regular capacity that survives the shock's impact and can be utilized for production.[3] For parsimony, we refer to ϵ_i simply as *capacity yield*. (The proportional random yield assumption is widely adopted in the OM literature; for example, see Federgruen and Yang (2009), Deo and Corbett (2009), and Tang and Kouvelis (2011).) We assume that ϵ_1 and ϵ_2 are independent and identically distributed with mean $\rho \equiv E[\epsilon_i]$, $i = 1, 2$, sharing the same pdf f and cdf F satisfying $f(0) > 0$, $F(0) = 0$, and $F(1) = 1$. Moreover, we assume that F is logconcave, a mild condition satisfied by many commonly used probability distributions (Bagnoli and Bergstrom 2005) and employed by other OM researchers (e.g., Cachon and Zhang 2006). For expositional ease, we use the notations $\overline{F}(x) \equiv 1 - F(x)$ and $G(x) \equiv \int_0^x F(y)\,dy = E[(x-\epsilon_i)^+]$.

To make up for reduced capacity, each firm maintains spare capacity that is utilized only during disruptions. (GAO (2014) reports the use of spare capacities by generic drug manufacturers.) To simplify the analysis we assume that spare capacities are unaffected by shocks, unlike the regular capacities. This distinction is reasonable in a situation where spare capacities are maintained in a secure location and utilized only during disruptions as backups, in contrast to fully-utilized regular capacities that are constantly exposed to the sources of shocks that may affect daily production operations. While this dichotomy is an idealized assumption, it is instrumental in deriving key analytical results; note that a similar assumption is commonly made in the spare inventory management models. Firm i sets s_i units of spare capacity at time zero and maintains them afterwards, incurring the holding cost h per

[3] The assumption that capacity is reduced by a fraction is consistent with a report by Woodcock and Wosinska (2013): "In most cases, firms have been able to continue production while improvements were being made, although often at a reduced rate."

unit of capacity per unit time. The holding cost may include the overhead cost of maintenance as well as the opportunity cost of capital; hence, h is similar to its counterpart in inventory models. (Note that the holding cost of regular capacity is normalized to zero, since it does not play a meaningful role in our analysis.) To rule out trivial equilibrium outcomes in which supply shortages never occur, we restrict attention to parameter combinations that satisfy the following relationships: $0 \leq s_i \leq \delta/2$ and $p\theta_i < h$ for $i = 1, 2$. The first condition states that the amount of spare capacity at each facility (s_i) does not exceed the maximum amount of regular capacity ($\delta/2$). The second condition states that the maximum profit that each unit of spare capacity can generate during disruptions at a single facility ($p\theta_i$) is outweighed by a firm's cost of holding the capacity (h).

Notice that spare capacity in our model plays a role similar to that of safety stocks in inventory models. We do not explicitly take inventory into account for two reasons. First, sterile injectables do not have long shelf lives, and therefore manufacturers typically carry only a small amount of safety stocks (HHS 2011). This implies that a volume increase is more likely to be achieved by expanding production capacity, rather than carrying more inventories. Second, an analysis of inventory competition under random supply disruptions and demand substitution is intractable, due to the difficulty of identifying parsimonious mathematical expressions for interdependent inventory levels at the two facilities that evolve according to a random process. In fact, even a model of non-competitive inventory control in the same setting presents analytical challenges, requiring somewhat arbitrary assumptions to gain tractability (see, for example, Groenevelt et al. 1992a,b). Although capacity and inventory are not mapped precisely to one another, their roles are analogous in that spare capacity captures the main tradeoffs of carrying safety stocks.

Finally, we assume that all variables that we introduced thus far, including the distribution functions, are common knowledge.

3.2 Demand Spillovers and Performance Measures

In line with the assumption that product units are perfectly substitutable, we assume that all demands that are allotted to facility i but cannot be filled there are immediately redirected to facility j. If the redirected demand is unfilled again, it is lost. (In their examination of the FDA data, Woodcock and Wosinska (2013) report similar spillover effects.) In the presence of such demand spillovers, then, each firm's sales rate (measured in quantity per unit time) depends on the states of the two facilities at any given moment. Clearly, the sales rate is equal to the allotted demand rate when neither facility is disrupted; since each has ample capacity to fill its allotted demand rate of $\delta/2$, no demand spillover occurs. When one of the facilities is disrupted, however, the sales rates differ from the allotted demand rates as capacity reduction may lead to demand spillover.

To quantify this, suppose that facility 1 is disrupted with percentage yield $\epsilon_1 < 1$ while facility 2 is undisrupted at a random point in time. With regular and spare capacities combined, the realized total capacities that firm 1 and firm 2 possess are $Q_1 = \delta\epsilon_1/2 + s_1$ and $Q_2 = \delta/2 + s_2$, respectively. Demand spillover from facility 1 to facility 2 occurs if Q_1 is smaller than the allotted demand rate $\delta/2$, or equivalently, if $\epsilon_1 < 1 - 2s_1/\delta$. The amount of spillover is equal to $(\delta/2 - Q_1)^+$, where $(\cdot)^+ \equiv \max\{0, \cdot\}$. Hence, firm 1's sales rate is equal to $\min\{\delta/2, Q_1\}$, whereas firm 2's sales rate is equal to $\min\left\{\delta/2 + (\delta/2 - Q_1)^+, Q_2\right\}$, taking into account the additional demands that are spilled over.

Summarizing the outcomes in all three facility states, the sales rate at facility i is equal to

$$R_i \equiv \begin{cases} \frac{\delta}{2} & \text{if neither facility is disrupted,} \\ \min\left\{\frac{\delta}{2}, \frac{\delta}{2}\epsilon_i + s_i\right\} & \text{if facility } i \text{ is disrupted but facility } j \text{ is not,} \\ \min\left\{\frac{\delta}{2} + \left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_j - s_j\right)^+, \frac{\delta}{2} + s_i\right\} & \text{if facility } j \text{ is disrupted but facility } i \text{ is not.} \end{cases} \quad (1)$$

Thus, firm i's long-run average sales rate is

$$E[R_i] = \tfrac{\delta}{2}(1 - \theta_i - \theta_j) + E\left[\min\left\{\tfrac{\delta}{2}, \tfrac{\delta}{2}\epsilon_i + s_i\right\}\right]\theta_i + E\left[\min\left\{\tfrac{\delta}{2} + \left(\tfrac{\delta}{2} - \tfrac{\delta}{2}\epsilon_j - s_j\right)^+, \tfrac{\delta}{2} + s_i\right\}\right]\theta_j, \quad (2)$$

which is obtained by taking expectations of each outcome in (1) and weighing them by their respective steady-state probabilities. (Recall that the above expression is an approximation under the assumption $\theta_i \ll 1$ for $i = 1, 2$ such that $\theta_1\theta_2 \approx 0$.)

The measure of consumer welfare in our model is *availability*, denoted as A, which is defined as the long-run average fill rate at the industry level that accounts for all three states of facilities. When neither facility is disrupted, fill rate is equal to one since there is ample capacity in the industry to satisfy all demands. When one of the facilities is disrupted, however, fill rate (the ratio between expected industry sales rate and total demand rate δ) may fall below one because of capacity reduction; it is equal to $E[\min\{\delta, Q_1 + Q_2\}]/\delta$. Enumerating all three states as above, we can write availability as

$$A = (1 - \theta_i - \theta_j) + \frac{E\left[\min\left\{\delta, \frac{\delta}{2}\epsilon_i + s_i + \frac{\delta}{2} + s_j\right\}\right]}{\delta}\theta_i + \frac{E\left[\min\left\{\delta, \frac{\delta}{2} + s_i + \frac{\delta}{2}\epsilon_j + s_j\right\}\right]}{\delta}\theta_j. \quad (3)$$

As expected, the following relationship can be verified using the expressions in (2) and (3): $A = E[R_1 + R_2]/\delta$. That is, availability is equal to the ratio between the long-run average industry sales rate and the demand rate.

As we prove in Lemma B.1 found in the Appendix, $E[R_i]$ and A defined in (2) and (3) can be expressed in terms of the distribution function F via the notation $G(x) = \int_0^x F(y)\,dy$. First, firm i's

long-run average sales rate $E[R_i]$ can be written as

$$E[R_i] = \frac{\delta}{2} - \frac{\delta}{2}\theta_i G\left(1 - 2\frac{s_i}{\delta}\right) + \frac{\delta}{2}\theta_j G\left(1 - 2\frac{s_j}{\delta}\right) - \frac{\delta}{2}\theta_j G\left((1 - 2\frac{s_i+s_j}{\delta})^+\right). \tag{4}$$

In the context of our problem, the function $G(\cdot)$ represents the expected fraction of demands that cannot be fulfilled because of insufficient capacity (i.e., fractional excess demand).[4] Thus, the expression in (4) can be interpreted as: firm i's expected sales rate in the absence of supply disruptions (the first term on the right-hand side) adjusted downward by the demands that cannot be fulfilled at disrupted facility i (the second term), adjusted upward by the demands that cannot be fulfilled at disrupted facility j and spilled over to undisrupted facility i (the third term), and finally adjusted downward by the demands that cannot be fulfilled either at disrupted facility j or at undisrupted facility i (the last term).

Similarly, availability A can be written as

$$A = 1 - \frac{\theta_i+\theta_j}{2} G\left((1 - 2\frac{s_i+s_j}{\delta})^+\right). \tag{5}$$

This is interpreted as: maximum availability of one (the first term) adjusted downward by the fraction of demands that cannot be fulfilled at either facility when disruptions occur (the second term). The expressions (4) and (5) are used extensively in our analysis.

3.3 Objectives and Decisions

As we outlined at the beginning of this section, we consider two scenarios based on whether or not the price varies with total industry capacity. In each scenario, we study a simultaneous-move game between firm 1 and firm 2 who set their spare capacity levels s_1 and s_2 competitively at time zero. Each firm's payoff is equal to its long-run average profit, denoted by π_i, $i = 1, 2$. The exact expression for π_i is presented in §4 and §5. Firm i sets s_i in order to maximize its payoff π_i. We identify the Nash equilibrium of this capacity game, denoted by the superscript $*$, and study its properties. To deliver clean insights, in several places of our analysis we pay special attention to the symmetric equilibrium that emerges under the assumption $\theta_1 = \theta_2$, i.e., the two facilities have identical disruption probabilities.

[4] For example, the second term on the right-hand side of (4) is derived as follows. Conditional on facility i being disrupted (with probability θ_i) and left with realized total capacity of $\frac{\delta}{2}\epsilon + s_i$, the amount of excess demand at each moment is $\left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon - s_i\right)^+$ since the facility receives the demands at the rate of $\frac{\delta}{2}$. Then the expected excess demand is $E\left[\left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon - s_i\right)^+\right] = \frac{\delta}{2}\int_0^{1-2s_i/\delta}\left(1 - 2\frac{s_i}{\delta} - x\right) f(x)\, dx = \frac{\delta}{2} G\left(1 - 2\frac{s_i}{\delta}\right)$. Therefore, the expected fraction of total demands $\frac{\delta}{2}$ that cannot be fulfilled by reduced supply is equal to $G\left(1 - 2\frac{s_i}{\delta}\right)$.

4 Capacity Choices Under Fixed Price

4.1 Equilibrium of Capacity Game

In this section we assume that unit price of the product is set to a constant value p at time zero and remains unchanged afterwards. Applied to the generic drug industry, this assumption simulates an idealized version of the price control scheme mentioned in §1 which creates price stickiness. With the price fixed at p, firm i's payoff is equal to $\pi_i = -hs_i + pE[R_i]$, i.e., the cost of holding s_i units of spare capacity plus the expected revenue in long-run averages. (Recall that the cost of installing capacity is assumed to be sunk at time zero.) Using the expression of $E[R_i]$ derived in (4), the payoff function can be written as

$$\pi_i = -hs_i + \tfrac{\delta p}{2}\left[1 - \theta_i G\left(1 - 2\tfrac{s_i}{\delta}\right) + \theta_j G\left(1 - 2\tfrac{s_j}{\delta}\right) - \theta_j G\left((1 - 2\tfrac{s_i+s_j}{\delta})^+\right)\right]. \tag{6}$$

As we outlined in §3.3, at time zero firm 1 and firm 2 set their spare capacity levels s_1 and s_2 in a simultaneous-move game. The Nash equilibrium of this game is specified in the following result. (Note that, as we assumed in §3.1, in all results below we restrict attention to parameter combinations that satisfy $p\theta_i < h$ for $i = 1, 2$.)

Proposition 1 (Equilibrium under fixed price) *A unique Nash equilibrium (s_1^*, s_2^*) of the capacity game exists and is identified as follows.*
(a) If $\theta_1 + \theta_2 \leq \frac{h}{p}$, then $(s_1^*, s_2^*) = (0, 0)$.
(b) If $\max\{\theta_1, \theta_2\} < \frac{h}{p} < \theta_1 + \theta_2$, then $(s_1^*, s_2^*) = (\widehat{s}_1, \widehat{s}_2)$ where $\widehat{s}_1 > 0$ and $\widehat{s}_2 > 0$ satisfying $\widehat{s}_1 + \widehat{s}_2 < \frac{\delta}{2}$ is a solution to the system of equations $\theta_i F\left(1 - 2\frac{s_i}{\delta}\right) + \theta_j F\left(1 - 2\frac{s_i+s_j}{\delta}\right) = \frac{h}{p}$ for $i, j \in \{1, 2\}$ with $i \neq j$.

All proofs are found in §A of the Appendix. The proposition identifies the condition under which firms have incentives to hold nonzero spare capacities: $p(\theta_1 + \theta_2) > h$. If this condition is violated, no spare capacity exists in the industry, thus maximally exposing consumers to supply disruptions. The condition states that investing in spare capacity is economically justified if the profit that a unit of spare capacity can generate during disruptions at *any* of the two facilities exceeds the cost of holding it. In other words, not only should the unit price be sufficiently high but also the disruptions *in the entire industry* occur relatively frequently (i.e., relatively large $\theta_1 + \theta_2$). This implies that each firm's capacity investment decision should factor in reliability of the other firm's facility, in addition to that of the firm's own facility. This externality arises because of demand spillovers; if firm j has insufficient amount of capacity during disruptions at its facility, the resulting excess demands are directed to firm

i, creating additional revenue opportunities for the latter and increasing the value of each unit of spare capacity it holds. From this reasoning it is clear that the two firms' spare capacities act as substitutes, which can be mathematically verified using the results in Proposition 1.

As we observe from Proposition 1, the equilibrium decisions (s_1^*, s_2^*) are implicitly specified as a system of equations that is not readily interpretable. For this reason, we first examine a simplified setting.

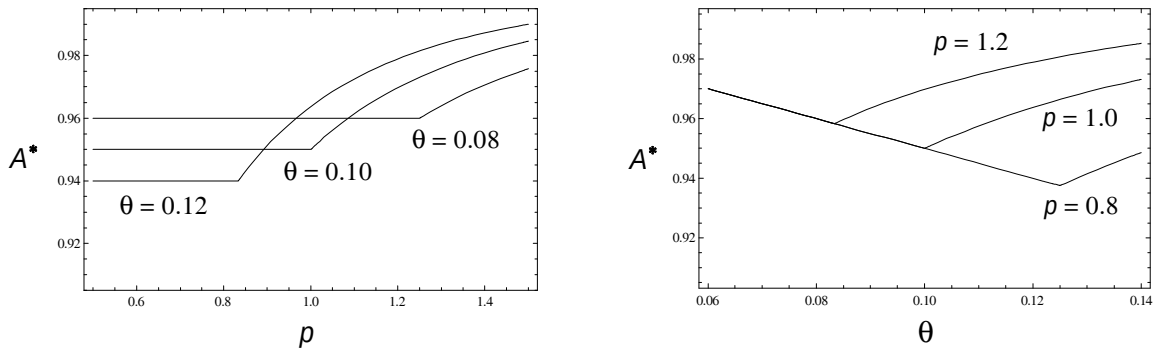
Corollary 1 *Suppose that ϵ_i is uniformly distributed for $i = 1, 2$ and $\theta_1 = \theta_2 = \theta$. Then: (a) $s_1^* = s_2^* = 0$ and $A^* = 1 - \frac{\theta}{2}$ if $p\theta \leq \frac{h}{2}$; (b) $s_1^* = s_2^* = \frac{\delta}{3}\left(1 - \frac{h}{2p\theta}\right) > 0$ and $A^* = 1 - \frac{1}{18\theta}\left(\frac{2h}{p} - \theta\right)^2$ if $\frac{h}{2} < p\theta < h$.*

In this special case where capacity yield ϵ_i is uniformly distributed and the firms have identical disruption probabilities $\theta_1 = \theta_2 = \theta$, we obtain closed-form solutions for the equilibrium spare capacities $s_1^* = s_2^*$ and the corresponding availability A^*. As it turns out, the most important properties of the equilibrium are captured in these simple equations. We discuss them next.

4.2 Properties of Equilibrium

First, the equations for s_1^* and s_2^* in Corollary 1 confirms the intuition that firms increase their spare capacity investments in response to higher price, provided that spare capacities are profitable; from the corollary, we see that $s_i^* = 0$ if price p is small (part (a)) but $s_i^* > 0$ increases in p if p is sufficiently large (part (b)). As a direct consequence, availability A^* is independent of the price p for small p but it increases in p otherwise. In other words, higher profit margin for spare capacity raises availability by providing the firms with more investment incentives; see Figure 1(a) for illustrations.

While this is the expected behavior, an interesting pattern emerges once we examine how it interacts with the disruption probability θ. Consider varying the value of θ. If p is small, it is clear from the expression for A^* in Corollary 1(a) that a higher chance of disruptions (higher θ) leads to lower availability; this is in line with intuition. However, if p is large, the opposite is true. From the expression for A^* in Corollary 1(b), we see that higher θ leads to *higher* availability. In other words, as facilities are expected to encounter more frequent and prolonged disruptions, product availability goes up. The contrast between these two regimes implies that the availability ranking for different values of θ is reversed as the price p increases. (For example, the three examples in Figure 1(a) show that availability is ranked in the order of $\theta = 0.08$, $\theta = 0.10$, and $\theta = 0.12$ when p is small, whereas the order changes to $\theta = 0.12$, $\theta = 0.10$, and $\theta = 0.08$ when p is large.) Figure 1(b) presents a different view of the same observation, clearly showing that availability increases in θ provided that firms have incentives to invest in spare capacities ($s_i^* > 0$).



(a) Availability as a function of price p

(b) Availability as a function of disruption probability θ

Figure 1: Availability A^* at the symmetric equilibrium under fixed price. The left panel shows A^* as a function of price p for three values of symmetric disruption probability $\theta_1 = \theta_2 = \theta$. The right panel shows A^* as a function of θ for three values of p. In these examples, the capacity yield random variable ϵ_i is assumed to be uniformly distributed with $h = 0.2$ and $\delta = 1$.

In fact, this is a general result—it is not a byproduct of restrictive assumptions in Corollary 1. Even in cases where the capacity yield is non-uniformly distributed and/or the disruption probabilities are asymmetric ($\theta_1 \neq \theta_2$), the same reversal holds. This is proved in the next proposition.

Proposition 2 *At the nonzero equilibrium* $(s_1^*, s_2^*) = (\widehat{s}_1, \widehat{s}_2)$ *specified in Proposition 1(b),* $\frac{\partial(\widehat{s}_1+\widehat{s}_2)}{\partial\theta_i} > 0$ *and* $\frac{\partial A^*}{\partial\theta_i} > 0$ *for* $i = 1, 2$.

Thus, availability A^* evaluated at the equilibrium increases in disruption probability at *either* facility (θ_1 or θ_2), given that the firms find it profitable to invest in spare capacities. This is a more general result than the one inferred from Corollary 1, since it is no longer assumed that disruption probabilities at the two facilities vary in tandem and the only condition imposed on capacity yield random variable is that its cdf F is logconcave.

Summarizing, the equilibrium of the capacity game is established in such a way that, as the disruption probabilities θ_1 and θ_2 gradually increase, the corresponding changes in availability A^* exhibit a sharp reversal . If disruptions are rare and short (small $\theta_1 + \theta_2$), then availability decreases in θ_1 and θ_2. By contrast, if disruptions are frequent and long (large $\theta_1 + \theta_2$), then availability increases in θ_1 and θ_2. The necessary and sufficient condition for this reversal is that spare capacities are profitable for the firms, i.e., the condition $p\,(\theta_1 + \theta_2) > h$ from Proposition 1 is satisfied.

As Proposition 2 reveals, availability increases in disruption probabilities because the firms respond to the latter change by increasing spare capacity levels ($s_1^* + s_2^*$ increases in θ_i); the greater the capacities, the higher the availability. This firm response is intuitive, since the economic value of spare

capacities increases if disruptions occur more often and last longer. However, it is not immediately clear why this endogenous capacity choice determines the direction of availability change, since it represents only an indirect effect of increased θ_i. The direct effect of increased θ_i pushes availability in the opposite direction, because without the firms' interventions, more frequent and prolonged disruptions (i.e., less reliability) lower availability. Proposition 2 states that the net outcome of these two competing forces is such that *the indirect effect dominates the direct effect*. What is significant from this result is that no middle ground is reached, resulting in an unambiguous reversal of the usual relationship between reliability and availability: higher chance of disruptions leads to higher availability.

Interestingly, a key driver of this counterintuitive result is the assumption that capacity yield random variable ϵ_i has a logconcave distribution function. Given that logconcave distributions are ubiquitous (most well-known distributions used in probability modeling belong to this category; see Bagnoli and Bergstrom 2005), we see that the seemingly contradictory relationship between reliability and availability is quite robust. In fact, it can be proved that the same result holds even in alternative scenarios where there is no competition; see Proposition B.2 in the Appendix.

To gain intuition, let us examine a simpler variation of the model in which a single firm managing one facility sets the spare capacity level s given that it faces the disruption probability θ and demands arriving at rate δ. An analysis similar to the one above reveals that, for sufficiently large p and θ, the firm chooses $\widehat{s} > 0$ that satisfies the optimality condition $\theta F\left(1 - \widehat{s}/\delta\right) = h/p$, which maps to the equilibrium condition found in Proposition 1. The left-hand side of this equation represents the long-run average chance of shortage occurrences; conditional on the facility being disrupted with probability θ, shortage occurs if the the total supply $\delta\epsilon + \widehat{s}$ is less than the total demand δ, i.e., with probability $\Pr(\delta\epsilon + \widehat{s} < \delta) = F(1 - \widehat{s}/\delta)$. Rearranging the above equation, we get:

$$1 - \theta F\left(1 - \frac{\widehat{s}}{\delta}\right) = \frac{p - h}{p}. \tag{7}$$

This equation is analogous to the newsvendor formula, as the left-hand side can be interpreted as the in-stock probability (in long-run average) and the right-hand side as the critical ratio. At the optimal choice $\widehat{s}$, availability is

$$A^* = 1 - \theta G\left(1 - \frac{\widehat{s}}{\delta}\right). \tag{8}$$

Notice the similarity between the expressions for in-stock probability and availability appearing in (7) and (8); the only difference between these two service measures is that the distribution function F in (7) is replaced with G in (8), the two linked via the relationship $G\left(x\right) = \int_0^x F\left(y\right) dy$.

Now, consider what happens when disruption probability θ is increased by an infinitesimal amount.

Differentiating (7) with respect to θ and rearranging the terms yields

$$\frac{\theta}{\delta}\frac{\partial \widehat{s}}{\partial \theta} = \frac{F\left(1-\widehat{s}/\delta\right)}{f\left(1-\widehat{s}/\delta\right)}. \tag{9}$$

The left-hand side of this equation denotes the rate of change in the normalized spare capacity level $\widehat{s}/\delta$ in response to a percentage increase in disruption probability. Hence, (9) specifies how the firm adjusts its spare capacity to maintain the optimality condition (7), or equivalently, to preserve payoff-maximizing in-stock probability. To see how this capacity adjustment impacts availability, we differentiate (8) with respect to θ and combine the result with (9) to obtain

$$\frac{\partial A^*}{\partial \theta} = -G\left(1-\frac{\widehat{s}}{\delta}\right) + F\left(1-\frac{\widehat{s}}{\delta}\right)\frac{\theta}{\delta}\frac{\partial \widehat{s}}{\partial \theta} = \frac{-f\left(1-\widehat{s}/\delta\right)G\left(1-\widehat{s}/\delta\right) + F\left(1-\widehat{s}/\delta\right)^2}{f\left(1-\widehat{s}/\delta\right)} > 0, \tag{10}$$

where the inequality follows from logconcavity of F since it implies $F(x)/f(x) > G(x)/F(x)$ (Bagnoli and Bergstrom 2005). Thus, we arrive at the same conclusion as above: even in the single-firm, single-facility case, availability increases in disruption probability provided that the firm has an incentive to invest in spare capacity.[5]

From these discussions we conclude that the availability increase originates from the firm's optimal adjustment of capacity that preserves in-stock probability, which turns out to *overcompensate* for the loss of availability. A reasoning based on the newsvendor model analysis helps understand this result better. Recall that in-stock probability quantifies a binary service outcome, namely the chance that *all* demands are satisfied by supply,[6] whereas availability (or fill rate) quantifies an incremental service outcome, namely the *fraction* of demands satisfied by supply. As these definitions convey, in-stock probability is a less flexible service measure than availability is, and as such, it is more sensitive to a change in environmental variables; a drop in in-stock probability due to increased θ is greater than a corresponding drop in availability. Consequently, compensating for the loss in in-stock probability requires a greater amount of capacity than what is needed to compensate for a similar loss in availability. Logconcavity of the distribution function F ensures that the firm's extra compensation

[5] Alternatively, one can show the same result using another property of logconcave distribution functions. Combining (7) with (8), we can write availability as $A^* = 1 - \frac{c}{p}\frac{G(1-\widehat{s}/\delta)}{F(1-\widehat{s}/\delta)}$, where the ratio $\frac{G(z)}{F(z)}$ can be rewritten as $z - \int_0^z x\frac{f(x)}{F(z)}dx \equiv \eta(z)$ via integration by parts. Bagnoli and Bergstrom (2005) call $\eta(z)$ "mean-advantage-over-inferiors function," and prove that logconcavity of F implies that it is monotone increasing. In the context of our problem, $\eta(z)$ can be interpreted as follows. Suppose that δ customers arrive at the facility and form a queue, obtaining the product on a first-come, first-serve basis. The customer at the z^{th} percentile is at the position δz in the queue. Then the expected sales conditional on a stockout occurrence for the z^{th} percentile-customers is $\int_0^z x\frac{f(x)}{F(z)}dx$, and therefore $\eta(z) = z - \int_0^z x\frac{f(x)}{F(z)}dx$ measures the gap between the queue position of an unfilled customer and the expected number of filled demands. Hence, that $\eta(z)$ is monotone increasing captures the notion that the marginal contribution to expected sales by an additional customer in the queue is diminishing. This monotonicity gives rise to our observation $\frac{\partial A^*}{\partial \theta} > 0$ in (10).

[6] The binary definition of in-stock probability enters into a newsvendor firm's optimality condition because the firm weighs the chance of selling a marginal unit of capacity versus the chance of not selling it.

to preserve the payoff-maximizing in-stock probability results in higher availability than where it started, as illustrated in (10). In essence, overcompensation arises because the shortage risk has a greater impact on the firm's profitability than it does on consumers' chance of obtaining the product.

Returning to the duopoly setting of our model, we see from Proposition 2 that the overcompensation effect survives under competition. In fact, this effect is amplified by competition. This is because decentralization provides each firm with an extra motive to increase its capacity: demand-stealing. As self-interested firms do not fully internalize the negative externality that demand substitution creates, they attempt to profit from each other's spillover demands by inflating their capacities. As a result, competition raises availability (this is proved in Proposition B.2 found in the Appendix). Note that an analogous result is well-known in inventory competition models (e.g., Netessine and Rudi 2003), and it is replicated in our setting where demand substitution occurs because of supply uncertainty, not because of demand uncertainty as assumed by most in the literature. Therefore, the availability increase we observe in Figure 1 is caused by a combination of two effects: firms' inherent tendency to overcompensate for the loss of availability, plus their competitive capacity overinvestment to steal each other's demands. This implies that the reversal of the usual relationship between reliability and availability, which we have discussed thus far, is likely to be prevalent in the generic drug industry where competition is common. In §6 we discuss the implications of this finding in the context of generic sterile injectable drug shortages.

5 Capacity Choices Under Shortage-Induced Price Increase

Building on the insights from the last section, we now relax the fixed price assumption and study how firms' capacity decisions and availability are influenced by price variation. Specifically, we modify the base model by assuming that unit price of the product increases in the amount of capacity shortfall. As we briefly mentioned in §1, such shortage-induced price increases are observed in practice despite the government's effort to contain them. Evidently the main reason for instituting a price control policy is to minimize the health care expenditure and protect consumer welfare. However, the impact of price variation on product availability is not clearly understood; our analysis in this section sheds light on this issue.

We model price variation as follows. Suppose that unit price of the product adjusts instantly to the total industry capacity $Q_1 + Q_2$ present at a given moment, where $Q_i = \delta\epsilon_i/2 + s_i$ is a random variable denoting the capacity that facility i possesses. The industry undergoes a shortage if and only if a disruption reduces the total capacity to $Q_1 + Q_2 < \delta$, i.e., the demand rate exceeds available production capacity. We assume that the price is fixed at a minimum value p_0 if there is ample capacity ($Q_1 + Q_2 \geq \delta$), whereas the price increases in proportion to the amount of shortfall if a shortage occurs

and all capacities are utilized ($Q_1 + Q_2 < \delta$). Note that the fixed minimum price assumption reflects the idea that unutilized capacities bring no economic benefits.

Given this specification, the price is a random variable since its value depends on realization of $Q_1 + Q_2$. This price, denoted by P, is equal to $P = p_0 + b\,(\delta - Q_1 - Q_2)^+$ where the coefficient b represents the rate at which the price increases in proportion to the amount of shortfall $(\delta - Q_1 - Q_2)^+$. We refer to b simply as *price sensitivity*. (Note that we recover the fixed price model of the last section by setting $b = 0$ and $p = p_0$.) As before, we assume that at most one facility is disrupted at a given moment via the assumption $\theta_i \ll 1$ and apply the approximation $(1-\theta_1)(1-\theta_2) \approx 1 - \theta_1 - \theta_2$. Since $Q_1 + Q_2 = \frac{\delta}{2} + \frac{\delta}{2}\epsilon_i + s_1 + s_2$ when facility i is disrupted with capacity yield ϵ_i and facility j is undisrupted, the price P defined above can be written as

$$P = \begin{cases} p_0 & \text{if neither facility is disrupted,} \\ p_0 + b\left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_i - s_1 - s_2\right)^+ & \text{if facility } i \text{ is disrupted but facility } j \text{ is not.} \end{cases} \tag{11}$$

Firm i's payoff is then equal to $\pi_i = -hs_i + E\,[PR_i]$, where the sales rate R_i is defined in (1). Combining (1) and (11) and taking expectations yield

$$\begin{aligned} \pi_i \;=\; & -hs_i + \tfrac{\delta p_0}{2}\left[1 - \theta_i G\left(1 - 2\tfrac{s_i}{\delta}\right) + \theta_j G\left(1 - 2\tfrac{s_j}{\delta}\right) - \theta_j G\left((1 - 2\tfrac{s_i+s_j}{\delta})^+\right)\right] \\ & + \tfrac{\delta^2 b}{4}\left[-\theta_i\left(1 - 2\tfrac{2s_i+s_j}{\delta}\right) G\left((1 - 2\tfrac{s_i+s_j}{\delta})^+\right) + \theta_j\left(1 + 2\tfrac{s_i}{\delta}\right) G\left((1 - 2\tfrac{s_i+s_j}{\delta})^+\right)\right] \\ & + \tfrac{\delta^2 b}{4}\left[\theta_i\left(1 - 2\tfrac{s_i+s_j}{\delta}\right)^2 F\left((1 - 2\tfrac{s_i+s_j}{\delta})^+\right) - \theta_i \textstyle\int_0^{(1-2\frac{s_i+s_j}{\delta})^+} x^2 f(x)\,dx\right]. \end{aligned} \tag{12}$$

Similar to what we assumed in the last section, we restrict attention to cases with $p_0\theta_i < h$ for $i = 1, 2$. The expression for availability A is unchanged from (5), since its definition does not explicitly include price. However, the availability is indirectly affected by price variation, since the firms' optimal choices for capacity levels s_1 and s_2 are influenced by the price.

As before, we assume that the firms set s_1 and s_2 at time zero in a simultaneous-move game, choosing the values that maximize their payoffs π_1 and π_2 given in (12). In order to isolate the impact of price variation, in the following discussions we focus on the symmetric Nash equilibrium of this game by assuming $\theta_1 = \theta_2 = \theta$, i.e., disruption probabilities at the two facilities are identical. The equilibrium is specified in the next proposition.

Proposition 3 (Symmetric equilibrium under shortage-induced price increase) *Suppose* $\theta_1 = \theta_2 = \theta$ *and* $b \le \frac{2p_0}{3\delta}$, *which implies* $\frac{h}{2}\left(p_0 - \frac{\delta b}{4}(3\rho - 1)\right)^{-1} < \frac{h}{p_0}$. *A unique symmetric equilibrium* (s_1^*, s_2^*) *of the capacity game exists and is identified as follows.*

(a) If $\theta \le \frac{h}{2}\left(p_0 - \frac{\delta b}{4}(3\rho - 1)\right)^{-1}$, *then* $s_1^* = s_2^* = 0$.

(b) If $\frac{h}{2}\left(p_0 - \frac{\delta b}{4}(3\rho - 1)\right)^{-1} < \theta < \frac{h}{p_0}$, *then* $s_1^* = s_2^* = \widehat{s} \in \left(0, \frac{\delta}{4}\right)$ *is a solution to the equation* $p_0\theta\left(F\left(1 - 2\frac{s}{\delta}\right) + F\left(1 - 4\frac{s}{\delta}\right)\right) - \frac{\delta b}{2}\theta\left(2F\left(1 - 4\frac{s}{\delta}\right) - 3G\left(1 - 4\frac{s}{\delta}\right)\right) = h$.

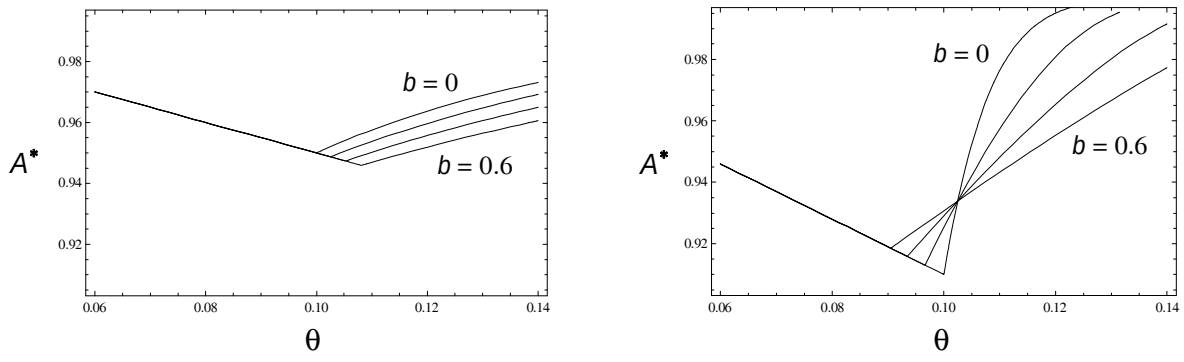
The proposition demonstrates that the basic structure of the equilibrium is unchanged from that specified in Proposition 1, where price was assumed to be fixed. Namely, as in the fixed price case, there exists a threshold value for θ under which firms do not invest in spare capacities ($s_i^* = 0$) and over which they do ($s_i^* > 0$); again, a sufficiently large chance of disruptions provides the firms with incentives to invest in spare capacities.

A departure from Proposition 1, however, is that this threshold value, $\theta = \frac{h}{2}\left(p_0 - \frac{\delta b}{4}(3\rho - 1)\right)^{-1}$, now depends on two additional parameters: price sensitivity b and expected capacity yield $\rho = E[\epsilon_i]$. This raises the question: how does the equilibrium adjust in response to changes in these parameters? In particular, do the firms invest more or less in spare capacities if they expect to see a greater price increase following a shortage (i.e., larger b)? The answer is found in the next result.

Proposition 4 *At the nonzero equilibrium* $(s_1^*, s_2^*) = (\widehat{s}, \widehat{s})$ *specified in Proposition 3(b), the following holds.*
(a) $\frac{\partial \widehat{s}}{\partial \theta} > 0$ *and* $\frac{\partial A^*}{\partial \theta} > 0$.
(b) If $\rho \geq \frac{1}{3}$, *then* $\frac{\partial \widehat{s}}{\partial b} < 0$ *and* $\frac{\partial A^*}{\partial b} < 0$. *If* $\rho < \frac{1}{3}$, *on the other hand, there exists a unique cutoff* $\theta = \theta^c$ *such that* $\frac{\partial \widehat{s}}{\partial b} > 0$ *and* $\frac{\partial A^*}{\partial b} > 0$ *for* $\theta < \theta^c$ *while* $\frac{\partial \widehat{s}}{\partial b} < 0$ *and* $\frac{\partial A^*}{\partial b} < 0$ *for* $\theta > \theta^c$.

Part (a) of the proposition parallels Proposition 2 of the fixed price case. The result shows that, even when the price is allowed to vary according to (11), the equilibrium is established in such a way that availability A^* increases in disruption probability θ if the combination of price coefficients p_0 and b makes capacity investment profitable for the firms. This result is analogous to Proposition 2; hence, the main insight from the last section—that the usual relationship between reliability and availability is reversed—is robust to price variation.

Part (b) answers the question above, namely, how price sensitivity b influences the (symmetric) capacity choices $s_1^* = s_2^* = \widehat{s}$ and the corresponding availability A^*. As the result shows, these relationships are not straightforward; $\widehat{s}$ and A^* may or may not increase in b. Moreover, the answer depends on expected capacity yield ρ, or equivalently, expected capacity loss percentage $1-\rho$. Consider the first case $\rho \geq 1/3$, i.e., when capacity loss is expected to be relatively small. The proposition states that both $\widehat{s}$ and A^* decrease in b in this case. (Compare four curves for A^* in Figure 2(a) that correspond to different values of b.) This is intuitive; if the firms anticipate that a shortage will lead to a precipitous price increase (large b), then they have incentives to restrict spare capacity levels so



(a) ρ = 0.5 (b) ρ = 0.1

Figure 2: Availability A^* at the symmetric equilibrium under the variable price specified in (11), as a function of disruption probability $\theta_1 = \theta_2 = \theta$. The four curves in each panel correspond to $b = 0$, $b = 0.2$, $b = 0.4$, and $b = 0.6$ (only the first and last are labeled in the figure). The examples in panel (a) have $\rho = 0.5$ with uniform distribution for capacity yield ϵ_i while the examples in panel (b) have $\rho = 0.1$ with power distribution $F(x) = x^{1/9}$. As in Figure 1, $h = 0.2$ and $\delta = 1$ are assumed.

that they can take advantage of the price jump.[7] As the condition $\rho \geq 1/3$ indicates, however, this intuition is valid insofar as disruptions do not cause severe capacity loss. Indeed, if capacity loss is expected to be significant ($\rho < 1/3$), the opposite can happen: equilibrium spare capacity level $\widehat{s}$ and availability A^* increase in b. That is, firms anticipating a sharp price increase reserve *more* spare capacities, reversing the earlier observation. Figure 2(b) identifies the region in which this reversal happens. From this discussion, we see that a shortage-induced price increase creates subtle dynamics.

An examination of the conditions under which the reversal arises helps us understand this result. According to Proposition 4(b), the reversal happens when disruptions occur relatively infrequently (θ smaller than the cutoff θ^c but large enough to ensure $\widehat{s} > 0$) and the expected capacity loss is severe (small ρ). The first condition discourages firms from holding a large amount of spare capacities, since they will be mostly idle. Because of restricted capacity, then, supply shortfall during shortages is expected to be significant and so is the price jump. The second condition, on the other hand, magnifies the supply shortfall at each facility since a disruption causes severe capacity loss. As a direct consequence, spillover demands increase; more demands originally allotted to a given facility are diverted to the other facility.

In sum, relatively infrequent disruptions and severe capacity loss together lead to a situation where firms are likely to face high price and high volume of spillover demands during disruptions. This creates an opportunity for each firm to increase its profit, achieved by expanding its spare capacity

[7] Yurukoglu (2012) offers the same insight: "If payments are higher in shortage periods, then manufacturers have incentives to create artificial shortages."

and capturing more spillover demands that contribute a high profit margin. It then follows that greater price sensitivity b amplifies this effect, i.e., firms' capacity-expansion incentives become stronger. This explains the reversal observed in Proposition 4(b) and Figure 2(b).

This reasoning reveals that the key to the observed reversal is each firm's demand-stealing motive. This is a purely competitive effect; as we show in Proposition B.4 found in the Appendix, this countervailing effect disappears in an alternative setting where a single firm makes capacity decisions for the entire industry. Hence, we conclude that competition introduces a tradeoff that establishes a nuanced relationship between price variation and availability. Intuition suggests that availability will be lowered if the price increases during shortages, since firms can exploit this market response by creating artificial shortages with restricted capacities to inflate the price. As our analysis reveals, however, firms' incentives to achieve this outcome are muted when capacities are too restricted, causing a significant amount of spillover demands. These secondary demands may be attractive to the firms enough to reverse their capacity restrictions, thus resulting in higher availability.

6 Implications to Generic Sterile Injectable Drug Shortages

Through the analysis in §4 and §5, we uncovered new dynamics that arise from interactions among supply disruptions, demand substitution, competition, and price control. All of them are unique features of the generic drug industry, and therefore the insights we gained from our analysis have important implications to the shortage situation that the industry is currently undergoing. In this section we discuss how our model predictions may explain some of the observed phenomena.

Two facts about the current shortage situation are particularly striking: a sudden, dramatic rise of shortages and their persistency. Together, they indicate that the problem was caused by a structural shift in the industry. Health care experts, policy analysts, and academics have proposed several hypotheses to identify the cause. Some have suggested that industry-wide mergers and consolidations had been the direct cause of the problem (U.S. House of Representatives 2012, Schweitzer 2013), but the study by HHS (2011) found little evidence to support this view. There are also those who point to the industry's adoption of just-in-time (JIT) manufacturing practices as the culprit, since it advocates keeping minimum amount of inventory (Gehrett 2012, Gordon 2013). However, this does not explain the sudden rise of shortages, since JIT had been around for many years prior to the onset of the problem.

Perhaps the most visible and controversial hypothesis is that a change in Medicare Part B reimbursement policy, part of the Medicare Modernization Act (MMA) of 2003, was directly responsible for causing the shortages. This policy change introduced a new basis for calculating drug purchase reimbursement rates, and it had an effect of significantly reducing average payments to service providers

	$\rho = 0.5$			$\rho = 0.1$		
	$\theta = 0.08$	$\theta = 0.10$	$\theta = 0.12$	$\theta = 0.08$	$\theta = 0.10$	$\theta = 0.12$
20% price drop	0.4%	2.1%	1.6%	3.0%	5.1%	0.2%
30% price drop	0.4%	2.6%	3.1%	3.0%	8.9%	2.9%
40% price drop	0.4%	2.6%	4.4%	3.0%	8.9%	10.8%

Table 1: Reduction in equilibrium availability A^* resulting from price drops. In these examples, the price drops from $p = 1.3$ by the percentages specified in the left column. Parameter values $h = 0.2$ and $\delta = 1$ are chosen along with three different values for θ, same as those from Figure 1. Two distributions for capacity yield ϵ_i are shown: $\rho = 0.5$ on the left side of the table is the mean of uniform distribution $F(x) = x$, and $\rho = 0.1$ on the right side is the mean of power distribution $F(x) = x^{1/9}$.

(see Graham (2012) for details; Yurukoglu (2012) reports a median drop of 50% in Medicare payments). The claim is that reduced reimbursements translate into lower drug prices for manufacturers, whose rational response is to underinvest in capacities used for producing these drugs (U.S. House of Representatives 2012). Yurukoglu (2012), based on his econometric analysis, concludes that empirical evidence supports this view. However, others have voiced skepticism based on practitioners' feedbacks and other counterexamples (Graham 2012, Woodcock and Wosinska 2013, GAO 2014).

While attributing the current shortage situation to MMA may be debatable, the basic economic argument is intuitive: reduced price lessens the incentives for capacity investment, which then lowers drug availability.[8] Not surprisingly, our model predicts this. Figure 1(a), for example, clearly shows that availability decreases as price drops. The same figure, however, provides richer information: comparing the three curves in Figure 1(a), we see that availability reduction is more pronounced for higher disruption probability θ. In other words, it is not simply that a price drop lowers availability, but that it tends to lower availability more precipitously if facilities are less reliable. This is a direct consequence of our findings from §4.2. There, we discovered that availability increases in disruption probability θ if price is high, because firms invest in capacities that overcompensate for the loss of availability. By contrast, availability decreases in θ if price is too low, which discourages capacity investments. It then follows that, if a price drop is significant enough to cross these two regimes, the corresponding reduction in availability becomes magnified when θ is large, i.e., if facilities are less reliable. The numerical examples in Table 1 confirm this point. For example, when disruptions are so severe that 90% of regular capacity is expected to be disabled ($\rho = 0.1$), a 40% price drop lowers availability by 3.0% if $\theta = 0.08$ but the same drop lowers availability much more substantially (10.8%) if $\theta = 0.12$.

This observation suggests a possible explanation for the sudden rise and persistency of drug short-

[8] Independent of the debates on the role of MMA, a report by HHS (2011) notes "among the group of drugs that eventually experience a shortage, average prices decreased in every year leading up to a shortage."

ages: *price reduction unmasked large disruption risks that had been hidden under capacity buffers.* Suppose that price is sufficiently high so that firms invest in capacities. Then, according to the insight we gained about how reliability impacts availability through capacity choices, the larger the underlying disruption risks, the less do they materialize into shortages because firms tend to overinvest in capacities. This buffer role of capacity, however, is a double-edged sword; just as the firms aggressively build capacity buffers that conceal the disruption risks when price is high, the buffers come off quickly after a significant price drop.

Numerous reports indicate that sterile injectable drug manufacturing processes are indeed exposed to potentially large disruption risks. Woodcock and Wosinska (2013), for example, identifies aging facilities as one of the main contributors to the current shortage problem, noting that many facilities have been running continually since the 1960s with minimal upgrades. High costs of facility upgrades may have shifted the firms' attention away from reliability improvement, directing them instead toward capacity expansion which, as we asserted above, inflates availability but makes it susceptible to price changes. This reasoning is supported by our analysis in Appendix §B.3, where we extend the base model by endogenizing firms' reliability decisions in addition to their capacity choices. We demonstrate that, under a few simplifying assumptions, higher cost of improving reliability leads the firms to lower reliability and increase capacities, resulting in higher availability; see Proposition B.6. This confirms our conjecture above—an equilibrium can be established where firms focus more on mitigating shortages (with capacity investments) than preventing them (with reliability improvement), thus concealing potentially large disruption risks that may surface if economic conditions change.

It is worth noting that a report by HHS (2011) does not mention the role of MMA. Instead, it puts weight on the view that more diversified portfolio of products was the main contributor to the shortages. According to the report, there was an unusually large number of patent expirations between 2008 and 2010, prompting many generic manufacturers to start producing these new drugs at the expense of existing ones. This is a plausible hypothesis, and our model prediction does not rule it out. Although our model does not explicitly take into account multiple products, the impact of introducing new products can be simulated by increasing the capacity holding cost h, which includes the opportunity cost of capital. From the expressions in Proposition 1 and Corollary 1, it is easy to see that increasing h is mathematically equivalent to reducing the price p. Hence, diverting excess capacity to new products has an effect similar to what price reduction achieves.

Finally, our analysis in §5 sheds light on how the current policy of limiting short-term price increase affects shortages. As we discovered, competition is a key factor in answering this question. If competition were absent, it is best to keep the price fixed; this is because a shortage-induced price

increase motivates the firms to create artificial shortages by restricting their capacities. Competition may mute such a perverse incentive, but it is effective only under certain conditions, namely, when disruptions do not occur very often but once they occur, they disable a large portion of production capacity. In such instances, competing firms' demand-stealing incentives overtake shortage-inducing incentives, thus raising availability. Hence, policy makers may take advantage of the generic drug manufacturers' fierce competition to raise availability, accomplished by allowing temporary price increases during shortages under some circumstances. While adhering to the principle of stabilizing prices has merits from a cost control perspective, it also has nontrivial implications to drug availability. Our analysis suggests that a balanced approach may bring more benefits to the consumers.

7 Conclusions

Motivated by the shortages of generic sterile injectable drugs that have plagued the U.S. health care system in recent years, in this paper we develop an analytical model that captures the unique characteristics of the generic drug industry in order to gain insights into what may have contributed to this problem. Our model assumes that two firms competitively set the levels of their spare capacities, i.e., production capacities that are utilized during disruptions to mitigate shortages. Complicating the decisions is the fact that unmet demands from one firm spills to the other during disruptions, since generic drugs are perfectly substitutable. The equilibrium of this game determines product availability, which varies with the changes in environmental variables such as probability of supply disruptions and short-term price increases.

We find that disruption probability and availability are nontrivially linked. Namely, the firms' equilibrium choices of spare capacities reverse the usual relationship between the two variables: a higher chance of disruptions leads to higher availability. Clearly, an immediate impact of increased disruption probability is lower product availability and reduced sales. This direct effect, however, is countered by an indirect effect—firms' capacity expansions to mitigate sales losses—which pushes availability in the opposite direction. Interestingly, the indirect effect dominates the direct effect, hence the reversal. Our analysis reveals that this overcompensation arises because the shortage risk has a greater impact on the firms' profitability than it does on the consumers' chance of obtaining the product.

This finding points to a possible explanation for the sudden rise of the current drug shortages. Prior to the onset of shortages, high drug availability may have been supported by the firms' investments in production capacities rather than investments in improving reliability of their manufacturing facilities. That is, large disruption risks may have been hidden under the capacity buffers that prevented shortages. Our model predicts that, in such situations, a substantial reduction in price (or an increase in

capacity holding cost) can trigger a precipitous drop in availability as the firms react to this change by quickly removing the inflated capacity buffers that concealed large disruption risks. Reports indicate that the generic sterile injectable drug industry had experienced significant price pressure and product proliferations in a short period of time leading up to the onset of current shortages, while the aging facilities had received few upgrades. According to our model, these are the ingredients for a "perfect storm" of sudden shortages.

Various suggestions have been made to correct the shortage problem, but many require structural changes of the industry that may take years to implement. One potential remedy is to make pricing more flexible. Generic drug prices are highly regulated, and there are mechanisms that limit price increases. In our analysis of the extended model, we investigate whether allowing a short-term price increase during shortages could raise availability. Such a price increase could give rise to perverse incentives, whereby firms restrict their capacities to create artificial shortages and receive high price. Availability will be lowered as a result. However, competition may lead to an equilibrium where the opposite outcome emerges. Since firms can increase their profits by capturing more spillover demands originating from their competitors, under certain circumstances, such demand-stealing incentives prompt capacity expansions that raise availability. Therefore, we find that pricing flexibility has a potential to alleviate the shortage problem, provided that certain conditions are met.

There are other factors that may influence the drug shortages that we did not discuss in this paper. For example, controversies surround the role of GPOs, the procurement intermediaries that are unique in pharmaceutical supply chains. Clarifying how GPOs may have helped or hurt the shortage situation would be a fruitful direction for future research. In addition, some have advocated strengthening the failure-to-supply clauses in drug purchase contracts that penalize manufacturers for product nondelivery (HHS 2011). These clauses appear to have flaws, since they are invoked only when alternative supplies exist; an industry-wide shortage does not trigger the penalties (Woodcock and Wosinska 2013). A mechanism design analysis that incorporates some of the unique features of the generic drug industry could reveal better contracting approaches.

References

[1] Aydin G, Babich V, Beil D, Yang Z (2012) Decentralized supply risk management. Kouvelis P, Dong L, Boyabatli O, Li R, eds. *Handbook of Integrated Risk Management in Global Supply Chains* (John Wiley & Sons, Hoboken, NJ), 389-424.

[2] Bagnoli M, Bergstrom T (2005) Log-concave probability and its applications. *Economic Theory.* 26(2): 445-469.

[3] Cachon G, Netessine S (2004) Game theoretic applications in supply chain analysis. *Supply Chain Analysis in the eBusiness Era*, eds. Simchi-Levi D, Wu SD, Shen Z-J. (Kluwer).

[4] Cachon G, Zhang F (2006) Procuring fast delivery: Sole sourcing with information asymmetry. *Management Science*, 52(6): 881-896.

[5] Caves RE, Whinston MD, Hurwitz MA, Pakes A, Temin P (1991) Patent expiration, entry, and competition in the U.S. pharmaceutical industry. *Brookings Papers on Economic Activity. Microeconomics*, Vol. 1991: 1-66.

[6] Dada M, Petruzzi NC, Schwarz LB (2007) A newsvendor's procurement problem when suppliers are unreliable. *Manufacturing & Service Operations Management*. 9(1): 9-32.

[7] Danzon PM, Caho L-W (2000) Does regulation drive out competition in pharmaceutical markets? *Journal of Law and Economics*, 43(2): 311-358.

[8] Deo S, Corbett CJ (2009) Cournot competition under yield uncertainty: The case of the U.S. influenza vaccine market. *Manufacturing & Service Operations Management.* 11(4): 563-576

[9] Federgruen A, Yang N (2009) Competition under generalized attraction models: Applications to quality competition under yield uncertainty. *Management Science.* 55(12): 2028-2043.

[10] Frank RG, Salkever DS (1997) Generic entry and the pricing of pharmaceuticals. *Journal of Economics & Management Strategy*. 6(1): 75-90.

[11] Gehrett BK (2012) A prescription for drug shortages. *Journal of the American Medical Association*. 307(2): 153-154.

[12] Gordon S (2013) 80 percent of cancer docs have faced drug shortages: Survey. HealthDay (December 18, 2013). http://consumer.healthday.com/cancer-information-5/breast-cancer-news-94/drug-shortages-continue-to-threaten-cancer-care-683163.html. (Accessed December 15, 2014.)

[13] Graham JR (2012) The shortage of generic sterile injectable drugs: Diagnosis and solutions. Policy Brief, Mackinac Center for Public Policy. http://www.mackinac.org/archives/2012/s2012-04SterileInjectables.pdf. (Accessed December 22, 2014.)

[14] Groenevelt H, Pintelon L, Seidmann A (1992a) Production lot sizing with machine breakdowns. *Management Science*. 38(1): 104-123.

[15] Groenevelt H, Pintelon L, Seidmann A (1992b) Production batching with machine breakdowns and safety stocks. *Operations Research*. 40(5): 959-971.

[16] Harris G (2011) U.S. scrambling to ease shortage of vital medicine. The New York Times (August 19). http://www.nytimes.com/2011/08/20/health/policy/20drug.html. (Accessed Dember 20, 2014.)

[17] Harris G (2011) Obama tries to speed response to shortages in vital medicines. *The New York*

Times (October 31). http://www.nytimes.com/2011/10/31/health/policy/medicine-shortages-addressed-in-obama-executive-order.html. (Accessed December 16, 2014.)

[18] Jensen V, Rappaport BA (2010) The reality of drug shortages—the case of the injectable agent Propofol. *New England Journal of Medicine*. 363(9): 806-807.

[19] Kim S-H, Tomlin B (2013) Guilt by association: Strategic failure prevention and recovery capacity investments. *Management Science*. 59(7): 1631-1649.

[20] Link MP, Hagerty K, Kantarjian HM (2012) Chemotherapy drug shortages in the United States: Genesis and potential solutions. *Journal of Clinical Oncology*. 30(7): 692-694.

[21] Muckstadt JA (2005) *Analysis and Algorithms for Service Parts Supply Chains*. Springer, New York.

[22] Netessine S, Rudi N (2003) Centralized and competitive inventory models with demand substitution. *Operations Research*. 51(2): 329-335.

[23] Olson LM, Wendling BW (2013) The effect of generic drug competition on generic drug prices during the Hatch-Waxman 180-Day exclusivity period. Working paper, Bureau of Economics, Federal Trade Commission.

[24] Reiffen D, Ward MR (2005) Generic drug industry dynamics. *Review of Economics and Statistics*, 87(1): 37-49.

[25] Scherer FM (1993) Pricing, profits, and technological progress in the pharmaceutical industry. *Journal of Economic Perspectives*. 7(3): 97-115.

[26] Schweitzer SO (2013) How the U.S. Food and Drug Administration can solve the prescription drug shortage problem. *American Journal of Public Health*, 103(5):e10-e14.

[27] Scott Morton FM (1999) Entry decisions in the generic pharmaceutical industry. *RAND Journal of Economics*, 30(3):421-440.

[28] Sherbrooke CC (1992) *Optimal Inventory Modeling of Systems: Multi-Echelon Techniques*. John Wiley & Sons, New York.

[29] Snyder LV, Atan Z, Peng P, Rong Y, Schmitt AJ, Sinsoyasal B (2014) OR/MS models for supply chain disruptions: A review. Working paper.

[30] Tang SY, Kouvelis P (2011) Supplier diversification strategies in the presence of yield uncertainty and buyer competition. *Manufacturing & Service Operations Management.* 13(4): 439-451.

[31] Tavernise S (2014) Drug shortages continue to vex doctors. *The New York Times* (February 10). http://www.nytimes.com/2014/02/11/health/shortages-of-critical-drugs-continue-to-vex-doctors-study-finds.html. (Accessed Dember 20, 2014.)

[32] Thomas K (2012) Drug shortages persist in U.S., harming care. *The New York Times* (Novem-

ber 16). http://www.nytimes.com/2012/11/17/business/drug-shortages-are-becoming-persistent-in-us.html. (Accessed Dember 20, 2014.)

[33] Tomlin B (2006) On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Science*. 52(5): 639-657.

[34] Tomlin B (2009) Disruption-management strategies for short life-cycle products. *Naval Research Logistics*. 56(4): 318-347.

[35] U.S. Department of Health and Human Services (2011) ASPE Issue Brief: Economic analysis of the causes of drug shortages. Office of the Assistant Secretary for Planning and Evaluation. http://aspe.hhs.gov/sp/reports/2011/drugshortages/ib.pdf. (Accessed December 15, 2014.)

[36] U.S. Food and Drug Administration (2011) A review of FDA's approach to medical product shortages. http://www.fda.gov/downloads/AboutFDA/ReportsManualsForms/Reports/UCM277755.pdf. (Accessed December 15, 2014.)

[37] U.S. House of Representatives (2011) Drug shortage crisis: Lives are in the balance. Hearing before the Subcommittee on Health Care, District of Columbia, Census and the National Archives of the Committee on Oversight and Government Reform, House of Representatives, 112 Congress. http://oversight.house.gov/hearing/drug-shortage-crisis-lives-are-in-the-balance/. (Accessed December 15, 2014.)

[38] U.S. Government Accountability Office (2014) Drug Shortages: Public health threat continues, despite efforts to help ensure product availability. GAO Report No. GAO-14-194. http://www.gao.gov/assets/670/660785.pdf. (Accessed December 15, 2014.)

[39] Wang Y, Gilland W, Tomlin B (2010) Mitigating supply risk: Dual sourcing or process improvement? *Manufacturing & Service Operations Management*. 12(3): 489-510.

[40] Woodcock J, Wosinska M (2013) Economic and technological drivers of generic sterile injectable drug shortages. *Nature*. 93(2): 170-176.

[41] Yang Z, Aydin G, Babich V, Beil DR (2009) Supply disruptions, asymmetric information, and a backup production option. *Management Science*, 55(2): 192-209.

[42] Yurukoglu A (2012) Medicare reimbursements and shortages of sterile injectable pharmaceuticals. Working paper, Stanford University.

Appendix

A Proofs

Proof of Proposition 1. Recall the assumption $0 \leq s_i \leq \frac{\delta}{2}$, which implies $0 \leq s_i + s_j \leq \delta$. Differentiating π_i from (6) with respect to s_i yields $\frac{\partial \pi_i}{\partial s_i} = -h + p\theta_i F\left(1 - 2\frac{s_i}{\delta}\right) + p\theta_j F\left(1 - 2\frac{s_i+s_j}{\delta}\right)$ if

$0 \leq s_i + s_j < \frac{\delta}{2}$ while $\frac{\partial \pi_i}{\partial s_i} = -h + p\theta_i F\left(1 - 2\frac{s_i}{\delta}\right) < 0$ if $\frac{\delta}{2} \leq s_i + s_j \leq \delta$, where the inequality $\frac{\partial \pi_i}{\partial s_i} < 0$ in the latter case follows from the assumption $\theta_i < \frac{h}{p}$. The inequality implies that the maximizer of π_i exists in the region $0 \leq s_i + s_j < \frac{\delta}{2}$, where $\frac{\partial^2 \pi_i}{\partial s_i^2} = -\frac{2}{\delta} p\theta_i f\left(1 - 2\frac{s_i}{\delta}\right) - \frac{2}{\delta} p\theta_j f\left(1 - 2\frac{s_i + s_j}{\delta}\right) < 0$ and $\frac{\partial^2 \pi_i}{\partial s_i \partial s_j} = -\frac{2}{\delta} p\theta_j f\left(1 - 2\frac{s_i + s_j}{\delta}\right) < 0$. Existence of the equilibrium is established by concavity of π_i (Cachon and Netessine 2004, Theorem 1). In addition, since $\left|\frac{\partial^2 \pi_i}{\partial s_i \partial s_j}\right| < \left|\frac{\partial^2 \pi_i}{\partial s_i^2}\right|$ for $i, j \in \{1, 2\}$ with $i \neq j$, the equilibrium is unique via contraction mapping (Cachon and Netessine 2004, Theorem 5). At the lower and upper bounds of s_i defined in the region $0 \leq s_i < \frac{\delta}{2} - s_j$, we have $\lim_{s_i \to 0} \frac{\partial \pi_i}{\partial s_i} = -h + p\theta_i + p\theta_j F\left(1 - 2\frac{s_j}{\delta}\right)$ and $\lim_{s_i \to \frac{\delta}{2} - s_j} \frac{\partial \pi_i}{\partial s_i} = -h + p\theta_i F\left(2\frac{s_j}{\delta}\right) < 0$, where the last inequality follows from the assumption $\theta_i < \frac{h}{p}$. Consider two cases: $\theta_i + \theta_j \leq \frac{h}{p}$ and $\theta_i + \theta_j > \frac{h}{p}$. If $\theta_i + \theta_j \leq \frac{h}{p}$, then $\lim_{s_i \to 0} \frac{\partial \pi_i}{\partial s_i} \leq -h + p\left(\theta_i + \theta_j\right) \leq 0$ for $i, j \in \{1, 2\}$ with $i \neq j$ and therefore the equilibrium is $s_1^* = s_2^* = 0$. Next, consider $\theta_i + \theta_j > \frac{h}{p}$. We show by contradiction that $s_1^* > 0$ and $s_2^* > 0$ in this case. Suppose that firm j finds it optimal to set $s_j = 0$, which requires $\lim_{s_j \to 0} \frac{\partial \pi_j}{\partial s_j} \leq 0$ given concavity. Then, since $\lim_{s_i \to 0} \left.\frac{\partial \pi_i}{\partial s_i}\right|_{s_j = 0} = -h + p\left(\theta_i + \theta_j\right) > 0$, firm i responds by setting $s_i = s_i^0$ where s_i^0 solves the first-order condition $\left.\frac{\partial \pi_i}{\partial s_i}\right|_{s_j = 0} = -h + p\left(\theta_i + \theta_j\right) F\left(1 - 2\frac{s_i}{\delta}\right) = 0$, or equivalently $F\left(1 - 2\frac{s_i^0}{\delta}\right) = \frac{h}{p(\theta_i + \theta_j)}$. This interior solution, however, contradicts the earlier assumption that $s_j = 0$ is optimal since $\lim_{s_j \to 0} \left.\frac{\partial \pi_j}{\partial s_j}\right|_{s_i = s_i^0} = -h + p\theta_j + p\theta_i F\left(1 - 2\frac{s_i^0}{\delta}\right) = -h + p\theta_j + h\frac{\theta_i}{\theta_i + \theta_j} = \theta_j\left(p - \frac{h}{\theta_i + \theta_j}\right) > 0$, i.e., π_j is maximized at $s_j > 0$. Thus, $s_1^* > 0$ and $s_2^* > 0$ when $\theta_i + \theta_j > \frac{h}{p}$ and their values are determined from the first-order conditions $\frac{\partial \pi_i}{\partial s_i} = -h + p\theta_i F\left(1 - 2\frac{s_i}{\delta}\right) + p\theta_j F\left(1 - 2\frac{s_i + s_j}{\delta}\right) = 0$ for $i, j \in \{1, 2\}$ with $i \neq j$. ■

Proof of Proposition 2. For notational convenience, let $\widehat{z}_i \equiv 1 - 2\frac{\widehat{s}_i}{\delta}$ and $\widehat{z}_{12} \equiv 1 - 2\frac{\widehat{s}_1 + \widehat{s}_2}{\delta}$, $i = 1, 2$. Since $\widehat{s}_i > 0$, we have $\widehat{z}_i > \widehat{z}_{12}$. Recall from Proposition 1 that the equilibrium is identified from the system of equations $\theta_1 F\left(\widehat{z}_1\right) + \theta_2 F\left(\widehat{z}_{12}\right) = \frac{h}{p}$ and $\theta_2 F\left(\widehat{z}_2\right) + \theta_1 F\left(\widehat{z}_{12}\right) = \frac{h}{p}$. Implicitly differentiating the two equations with respect to θ_1 and combining them, we get

$$\frac{\partial\left(\widehat{s}_1 + \widehat{s}_2\right)}{\partial \theta_1} = \frac{\delta}{2} \frac{\theta_1 f\left(\widehat{z}_1\right) F\left(\widehat{z}_{12}\right) + \theta_2 f\left(\widehat{z}_2\right) F\left(\widehat{z}_1\right)}{\theta_1^2 f\left(\widehat{z}_1\right) f\left(\widehat{z}_{12}\right) + \theta_1 \theta_2 f\left(\widehat{z}_1\right) f\left(\widehat{z}_2\right) + \theta_2^2 f\left(\widehat{z}_2\right) f\left(\widehat{z}_{12}\right)} > 0. \tag{A.1}$$

To prove $\frac{\partial A^*}{\partial \theta_i} > 0$, we first show $\frac{\partial(\widehat{s}_1 + \widehat{s}_2)}{\partial \theta_1} > \frac{\delta}{2(\theta_1 + \theta_2)} \frac{F(\widehat{z}_{12})}{f(\widehat{z}_{12})}$. Dividing the numerator and the denominator of the right hand-side of (A.1) by $f\left(\widehat{z}_1\right) f\left(\widehat{z}_2\right)$ and rearranging the terms, it can be shown that $\frac{\partial(\widehat{s}_1 + \widehat{s}_2)}{\partial \theta_1} > \frac{\delta}{2(\theta_1 + \theta_2)} \frac{F(\widehat{z}_{12})}{f(\widehat{z}_{12})}$ if and only if $\varphi > 0$ where $\varphi \equiv \theta_1\left(\frac{f(\widehat{z}_{12})}{f(\widehat{z}_2)} + \frac{F(\widehat{z}_1)}{f(\widehat{z}_1)} \frac{f(\widehat{z}_{12})}{F(\widehat{z}_{12})} - 1\right) + \theta_2\left(\frac{F(\widehat{z}_1)}{f(\widehat{z}_1)} \frac{f(\widehat{z}_{12})}{F(\widehat{z}_{12})} - \frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)}\right)$. Consider two cases in turn: $\frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)} \geq 1$ and $\frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)} < 1$. Suppose $\frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)} \geq 1$. Since $\widehat{z}_1 > \widehat{z}_{12}$ and the cdf F is an increasing function, we have $\frac{F(\widehat{z}_1)}{F(\widehat{z}_{12})} > 1$. Hence, $\varphi > \theta_1\left(\frac{f(\widehat{z}_{12})}{f(\widehat{z}_2)} + \frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)} - 1\right) + \theta_2\left(\frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)} - \frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)}\right) = \theta_1\left(\frac{f(\widehat{z}_{12})}{f(\widehat{z}_2)} + \frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)} - 1\right) \geq \theta_1 \frac{f(\widehat{z}_{12})}{f(\widehat{z}_2)} \geq 0$, where we used the condition $\frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)} \geq 1$. Next, suppose $\frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)} < 1$. Since F is logconcave, the ratio $\frac{F(x)}{f(x)}$ is an

increasing function. With $\widehat{z}_1 > \widehat{z}_{12}$, we thus have $\frac{F(\widehat{z}_1)}{f(\widehat{z}_1)} > \frac{F(\widehat{z}_{12})}{f(\widehat{z}_{12})}$ or equivalently $\frac{F(\widehat{z}_1)}{f(\widehat{z}_1)}\frac{f(\widehat{z}_{12})}{F(\widehat{z}_{12})} > 1$. It then follows that $\varphi > \theta_1\left(\frac{f(\widehat{z}_{12})}{f(\widehat{z}_2)} + 1 - 1\right) + \theta_2\left(1 - \frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)}\right) > \theta_1\frac{f(\widehat{z}_{12})}{f(\widehat{z}_2)} \geq 0$, where we used the condition $\frac{f(\widehat{z}_{12})}{f(\widehat{z}_1)} < 1$. In both cases we have $\varphi > 0$, which implies $\frac{\partial(\widehat{s}_1+\widehat{s}_2)}{\partial\theta_1} > \frac{\delta}{2(\theta_1+\theta_2)}\frac{F(\widehat{z}_{12})}{f(\widehat{z}_{12})}$ as we set out to prove. Differentiating availability A^* from (5) and applying this inequality, we have $\frac{\partial A^*}{\partial\theta_1} = -\frac{1}{2}G\left(1 - 2\frac{\widehat{s}_1+\widehat{s}_2}{\delta}\right) + \frac{\theta_1+\theta_2}{\delta}F\left(1 - 2\frac{\widehat{s}_1+\widehat{s}_2}{\delta}\right)\frac{\partial(\widehat{s}_1+\widehat{s}_2)}{\partial\theta_1} > -\frac{1}{2}G\left(\widehat{z}_{12}\right) + \frac{1}{2}\frac{F(\widehat{z}_{12})^2}{f(\widehat{z}_{12})} > 0$, where the last inequality follows from logconcavity of F; to see this, note that logconcavity of F implies logconcavity of G (Bagnoli and Bergstrom 2005, Theorem 1), which implies $\frac{d^2}{dx^2}\ln G\left(x\right) = \frac{f(x)G(x)-F(x)^2}{G(x)^2} < 0$. ■

Proof of Proposition 3. Since $0 \leq \rho \leq 1$ and $b \leq \frac{2p_0}{3\delta}$, we have $\frac{h}{2(p_0-\delta b(3\rho-1)/4)} \leq \frac{h}{2(p_0-\delta b/2)} \leq \frac{3h}{4p_0} < \frac{h}{p_0}$, proving the assertion in the proposition. Suppose $\frac{\delta}{2} \leq s_i + s_j \leq \delta$. Then firm i's payoff in (12) reduces to $\pi_i = -hs_i + \frac{\delta p_0}{2}\left[1 - \theta_i G\left(1 - 2\frac{s_i}{\delta}\right) + \theta_j G\left(1 - 2\frac{s_j}{\delta}\right)\right]$. Differentiating this with respect to s_i yields $\frac{\partial\pi_i}{\partial s_i} = -h + p_0\theta_i F\left(1 - 2\frac{s_i}{\delta}\right) < 0$, where the inequality follows from the assumption $\theta_i < \frac{h}{p_0}$, $i = 1, 2$. Since π_i is decreasing in s_i in the interval $\frac{\delta}{2} \leq s_i + s_j \leq \delta$, the equilibrium does not exist there. Next, suppose $0 \leq s_i + s_j < \frac{\delta}{2}$. For notational convenience, let $z_i \equiv 1 - 2\frac{s_i}{\delta}$ and $z_{ij} \equiv 1 - 2\frac{s_i+s_j}{\delta}$. Differentiating π_i in (12) and setting $\theta_i = \theta_j = \theta$ yield $\frac{\partial\pi_i}{\partial s_i} = -h + p_0\theta\left[F\left(z_i\right) + F\left(z_{ij}\right)\right] + \frac{3\delta b}{2}\theta G\left(z_{ij}\right) - \delta b\theta\left(1 + \frac{s_i-s_j}{\delta}\right)F\left(z_{ij}\right)$ and $\frac{\partial^2\pi_i}{\partial s_i^2} = -\frac{2p_0}{\delta}\theta\left[f\left(z_i\right) + f\left(z_{ij}\right)\right] - 4b\theta F\left(z_{ij}\right) + 2b\theta\left(1 + \frac{s_i-s_j}{\delta}\right)f\left(z_{ij}\right)$. Since $0 \leq s_i \leq \frac{\delta}{2}$ and $0 \leq s_j \leq \frac{\delta}{2}$ by assumption, we have $-\frac{\delta}{2} \leq s_i - s_j \leq \frac{\delta}{2}$. This, together with the assumption $b \leq \frac{2p_0}{3\delta}$ stated in the proposition, implies $\frac{p_0}{\delta} - b\left(1 + \frac{s_i-s_j}{\delta}\right) \geq 0$. Then $\frac{\partial^2\pi_i}{\partial s_i^2} = -\frac{2p_0}{\delta}\theta f\left(z_i\right) - 4b\theta F\left(z_{ij}\right) - 2\theta\left[\frac{p_0}{\delta} - b\left(1 + \frac{s_i-s_j}{\delta}\right)\right]f\left(z_{ij}\right) < 0$, i.e., π_i is concave. Given symmetry, this implies that a symmetric Nash equilibrium exists if $b \leq \frac{2p_0}{3\delta}$ (Cachon and Netessine 2004). At the symmetric equilibrium $s_i = s_j = s$ in the considered interval $0 \leq s_i + s_j < \frac{\delta}{2}$ or equivalently $0 \leq s < \frac{\delta}{4}$, we have $\left.\frac{\partial\pi_i}{\partial s_i}\right|_{s_i=s_j=s} = \chi\left(s\right)$, where $\chi\left(s\right) \equiv -h + p_0\theta\left(F\left(1 - 2\frac{s}{\delta}\right) + F\left(1 - 4\frac{s}{\delta}\right)\right) - \frac{\delta b}{2}\theta\left(2F\left(1 - 4\frac{s}{\delta}\right) - 3G\left(1 - 4\frac{s}{\delta}\right)\right)$. Observe the following properties of $\chi\left(s\right)$; (i) $\chi'\left(s\right) = -\frac{2p_0\theta}{\delta}f\left(1 - 2\frac{s}{\delta}\right) - 4\theta\left(\frac{p_0}{\delta} - b\right)f\left(1 - 4\frac{s}{\delta}\right) - 6b\theta F\left(1 - 4\frac{s}{\delta}\right) < 0$, where the inequality follows from the assumption $b \leq \frac{2p_0}{3\delta}$; (ii) $\chi\left(0\right) = -h + 2p_0\theta - \frac{\delta b}{2}\theta\left(3\rho - 1\right)$, where we used the identity $G\left(1\right) = \int_0^1 F\left(x\right)dx = 1 - E\left[\epsilon_i\right] = 1 - \rho$; (iii) $\chi\left(\frac{\delta}{4}\right) = -h + p_0\theta F\left(\frac{1}{2}\right) < 0$, where the inequality follows from the assumption $\theta < \frac{h}{p_0}$. Therefore, $\chi\left(s\right)$ is a decreasing function that starts from $-h + 2p_0\theta - \frac{\delta b}{2}\theta\left(3\rho - 1\right)$ at $s = 0$ and approaches a negative number as $s \to \frac{\delta}{4}$. This implies that the symmetric equilibrium is unique and specified as follows: $s^* = 0$ if $\chi\left(0\right) = -h + 2p_0\theta - \frac{\delta b}{2}\theta\left(3\rho - 1\right) \leq 0$ or equivalently $\theta \leq \frac{h}{2(p_0-\delta b(3\rho-1)/4)}$, while $s^* = \widehat{s} > 0$ where $\widehat{s}$ is the unique solution to the equation $\chi\left(s\right) = 0$ if $\frac{h}{2(p_0-\delta b(3\rho-1)/4)} < \theta < \frac{h}{p_0}$. ■

Proof of Proposition 4. For notational convenience, let $\widehat{z}_1 \equiv 1 - 2\frac{\widehat{s}}{\delta}$ and $\widehat{z}_{12} \equiv 1 - 4\frac{\widehat{s}}{\delta}$.

(a) Implicitly differentiating the equilibrium condition $h = p_0\theta\left(F\left(\widehat{z}_1\right) + F\left(\widehat{z}_{12}\right)\right) - \frac{\delta b\theta}{2}\left(2F\left(\widehat{z}_{12}\right) - 3G\left(\widehat{z}_{12}\right)\right)$

specified in Proposition 3 with respect to θ and collecting the terms, we get

$$\frac{\partial \widehat{s}}{\partial \theta} = \frac{\delta}{4\theta} \frac{2p_0 F(\widehat{z}_1) + 2(p_0 - \delta b) F(\widehat{z}_{12}) + 3\delta b G(\widehat{z}_{12})}{p_0 f(\widehat{z}_1) + 2(p_0 - \delta b) f(\widehat{z}_{12}) + 3\delta b F(\widehat{z}_{12})} > 0, \tag{A.2}$$

where the inequality follows from the condition $b \leq \frac{2p_0}{3\delta}$ (see Proposition 3), which implies $p_0 - \delta b \geq \frac{p_0}{3} > 0$. To prove $\frac{\partial A^*}{\partial \theta} > 0$, where $A^* = 1 - \theta G(\widehat{z}_{12})$ is evaluated at the symmetric equilibrium (see (5)), we first show $\frac{\partial \widehat{s}}{\partial \theta} > \frac{\delta}{4\theta} \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})}$. Dividing the numerator of the right-hand side of (A.2) by $G(\widehat{z}_{12})$ and the denominator by $F(\widehat{z}_{12})$ and rearranging the terms, it can be shown that $\frac{\partial \widehat{s}}{\partial \theta} > \frac{\delta}{4\theta} \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})}$ if and only if $\varphi > 0$ where $\varphi \equiv p_0 \left(\frac{2F(\widehat{z}_1)}{G(\widehat{z}_{12})} - \frac{f(\widehat{z}_1)}{F(\widehat{z}_{12})} \right) + 2(p_0 - \delta b) \left(\frac{F(\widehat{z}_{12})}{G(\widehat{z}_{12})} - \frac{f(\widehat{z}_{12})}{F(\widehat{z}_{12})} \right)$. The first term of φ is positive since $\frac{2F(\widehat{z}_1)}{G(\widehat{z}_{12})} - \frac{f(\widehat{z}_1)}{F(\widehat{z}_{12})} = \frac{F(\widehat{z}_1)}{G(\widehat{z}_{12})} + \frac{f(\widehat{z}_1)}{G(\widehat{z}_{12})} \left(\frac{F(\widehat{z}_1)}{f(\widehat{z}_1)} - \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} \right) > \frac{F(\widehat{z}_1)}{G(\widehat{z}_{12})} + \frac{f(\widehat{z}_1)}{G(\widehat{z}_{12})} \left(\frac{F(\widehat{z}_{12})}{f(\widehat{z}_{12})} - \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} \right) > 0$. The two inequalities follow from logconcavity of F, since it implies $\frac{F(z)}{f(z)}$ is an increasing function and $\frac{F(z)}{f(z)} > \frac{G(z)}{F(z)}$ (Bagnoli and Bergstrom 2005). Similarly, the second term of φ is positive since $\frac{F(z)}{f(z)} > \frac{G(z)}{F(z)}$ and $p_0 - \delta b > 0$. Since both terms of φ are positive, $\varphi > 0$ and it follows that $\frac{\partial \widehat{s}}{\partial \theta} > \frac{\delta}{4\theta} \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})}$. This inequality implies $\frac{\partial A^*}{\partial \theta} > 0$, since $\frac{\partial A^*}{\partial \theta} = -G(\widehat{z}_{12}) + \frac{4\theta}{\delta} F(\widehat{z}_{12}) \frac{d\widehat{s}}{d\theta} > -G(\widehat{z}_{12}) + \frac{4\theta}{\delta} F(\widehat{z}_{12}) \left(\frac{\delta}{4\theta} \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} \right) = 0$.

(b) Implicitly differentiating the equilibrium condition $h = p_0 \theta (F(\widehat{z}_1) + F(\widehat{z}_{12})) - \frac{\delta b \theta}{2} (2F(\widehat{z}_{12}) - 3G(\widehat{z}_{12}))$ specified in Proposition 3 with respect to b and collecting terms, we get

$$\frac{\partial \widehat{s}}{\partial b} = \frac{\delta^2}{4} \frac{-2F(\widehat{z}_{12}) + 3G(\widehat{z}_{12})}{p_0 f(\widehat{z}_1) + 2(p_0 - \delta b) f(\widehat{z}_{12}) + 3\delta b F(\widehat{z}_{12})}. \tag{A.3}$$

Since $p_0 - \delta b > 0$, the denominator of the right-hand side of (A.3) is positive. Thus, the sign of $\frac{\partial \widehat{s}}{\partial b}$ is determined by the sign of the numerator: $\frac{\partial \widehat{s}}{\partial b} < 0$ if $\frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} < \frac{2}{3}$ and $\frac{d\widehat{s}}{db} > 0$ if $\frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} > \frac{2}{3}$. Note that the function $\frac{G(z)}{F(z)}$ is increasing by logconcavity of F (Bagnoli and Bergstrom 2005). Thus, $\frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})}$ with $\widehat{z}_{12} = 1 - 4\frac{\widehat{s}}{\delta}$ decreases in $\widehat{s}$, starting from $\lim_{\widehat{s} \to 0} \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} = \frac{G(1)}{F(1)} = 1 - \rho$ and approaching $\lim_{\widehat{s} \to \delta/4} \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} = \lim_{z \to 0} \frac{G(z)}{F(z)} = \lim_{z \to 0} \frac{F(z)}{f(z)} = 0$, where we used l'Hopital's rule (note it has been assumed $f(0) > 0$; see §3.1). Suppose $\rho \geq \frac{1}{3}$. Then $\frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} < \frac{2}{3}$ for all $\widehat{s} > 0$ since $\lim_{\widehat{s} \to 0} \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} = 1 - \rho \leq \frac{2}{3}$ and $\frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})}$ decreases in $\widehat{s}$. From (A.3), we see that this implies $\frac{d\widehat{s}}{db} < 0$ for all $\widehat{s} > 0$. On the other hand, if $\rho < \frac{1}{3}$ then $\lim_{\widehat{s} \to 0} \frac{G(\widehat{z}_{12})}{F(\widehat{z}_{12})} = 1 - \rho > \frac{2}{3}$ and therefore there exists a unique cutoff $\widehat{s}^c \in \left(0, \frac{\delta}{4}\right)$ such that $\frac{\partial \widehat{s}}{\partial b} > 0$ for $\widehat{s} < \widehat{s}^c$ and $\frac{\partial \widehat{s}}{\partial b} < 0$ for $\widehat{s} > \widehat{s}^c$. Given that $\widehat{s}$ is monotone increasing in θ, as proved in (A.2), this in turn implies that there exists a unique value θ^c that maps to $\widehat{s}^c$ such that $\frac{\partial \widehat{s}}{\partial b} > 0$ for $\theta < \theta^c$ and $\frac{\partial \widehat{s}}{\partial b} < 0$ for $\theta > \theta^c$. Summarizing, we have $\frac{\partial \widehat{s}}{\partial b} < 0$ if $\rho \geq \frac{1}{3}$ while $\frac{\partial \widehat{s}}{\partial b} > 0$ for $\theta < \theta^c$ and $\frac{\partial \widehat{s}}{\partial b} < 0$ for $\theta > \theta^c$ if $\rho < \frac{1}{3}$. Since availability at the symmetric equilibrium is $A^* = 1 - \theta G(\widehat{z}_{12})$, it is clear that $\frac{\partial A^*}{\partial b} < 0$ if and only if $\frac{\partial \widehat{s}}{\partial b} < 0$; the statement in the proposition follows from these observations. ■

B Additional Results

B.1 Performance Measure Evaluations

Lemma B.1 *With price p fixed, firm i's long-run average sales and availability defined in (2) and (3) are evaluated as (4) and (5).*

Proof. The expectation in the second term of (2) is evaluated as $E\left[\min\left\{\frac{\delta}{2}, \frac{\delta}{2}\epsilon_i + s_i\right\}\right] = \int_0^{1-2\frac{s_i}{\delta}} \left(\frac{\delta}{2}x + s_i\right) f(x)\, dx + \frac{\delta}{2}\overline{F}\left(1-2\frac{s_i}{\delta}\right) = \frac{\delta}{2} - \frac{\delta}{2}\int_0^{1-2\frac{s_i}{\delta}} F(x)\, dx$ using integration by parts. To evaluate the expectation in the last term of (2), consider two cases separately: $s_i + s_j < \frac{\delta}{2}$ and $s_i + s_j \geq \frac{\delta}{2}$. Suppose $s_i + s_j < \frac{\delta}{2}$, which implies $0 < 1 - 2\frac{s_i+s_j}{\delta} \leq 1 - 2\frac{s_j}{\delta} \leq 1$. Then $\min\left\{\frac{\delta}{2} + \left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_j - s_j\right)^+, \frac{\delta}{2} + s_i\right\}$ is equal to: (i) $\frac{\delta}{2} + s_i$ if $0 \leq \epsilon_j < 1 - 2\frac{s_i+s_j}{\delta}$; (ii) $\delta - \frac{\delta}{2}\epsilon_j - s_j$ if $1 - 2\frac{s_i+s_j}{\delta} \leq \epsilon_j < 1 - 2\frac{s_j}{\delta}$; (iii) $\frac{\delta}{2}$ if $1 - 2\frac{s_j}{\delta} \leq \epsilon_j \leq 1$. Therefore, $E\left[\min\left\{\frac{\delta}{2} + \left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_j - s_j\right)^+, \frac{\delta}{2} + s_i\right\}\right] = \left(\frac{\delta}{2} + s_i\right) F\left(1 - 2\frac{s_i+s_j}{\delta}\right) + \int_{1-2\frac{s_i+s_j}{\delta}}^{1-2\frac{s_j}{\delta}} \left(\delta - \frac{\delta}{2}x - s_j\right) f(x)\, dx + \frac{\delta}{2}\overline{F}\left(1-2\frac{s_j}{\delta}\right)$, which is evaluated as $\frac{\delta}{2} + \frac{\delta}{2}G\left(1-2\frac{s_j}{\delta}\right) - \frac{\delta}{2}G\left(1-2\frac{s_i+s_j}{\delta}\right)$ using integration by parts. Next, suppose $\frac{\delta}{2} \leq s_i + s_j$, which implies $1 - 2\frac{s_i+s_j}{\delta} \leq 0 \leq 1 - 2\frac{s_j}{\delta} \leq 1$. Following the steps similar to above, it can be shown that $E\left[\min\left\{\frac{\delta}{2} + \left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_j - s_j\right)^+, \frac{\delta}{2} + s_i\right\}\right] = \frac{\delta}{2} + \frac{\delta}{2}G\left(1-2\frac{s_j}{\delta}\right)$ in this case. Substituting the evaluated expressions in (2) and rearranging the terms yield (4). Evaluating the expectations in (3) similarly and rearranging the terms yield the availability expression in (5). ■

Lemma B.2 *With price P specified in (11), firm i's expected payoff is evaluated as (12).*

Proof. Recall that the steady-state probability that neither facility is disrupted is $1 - \theta_i - \theta_j$ while the probability that facility i is disrupted is θ_i. With the sales rate R_i and the price P specified as in (1) and (11), firm i's payoff $\pi_i = -hs_i + E[PR_i]$ can be written as $\pi_i = -hs_i + p_0\sigma_i + b\theta_i\xi_i + b\theta_j\zeta_i$ where σ_i is equal to (2) and $\xi_i \equiv E\left[\left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_i - s_i - s_j\right)^+ \min\left\{\frac{\delta}{2}, \frac{\delta}{2}\epsilon_i + s_i\right\}\right]$ and $\zeta_i \equiv E\left[\left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_j - s_i - s_j\right)^+ \min\left\{\frac{\delta}{2} + \left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_j - s_j\right)^+, \frac{\delta}{2} + s_i\right\}\right]$. Note σ_i has been evaluated in (4). If $\frac{\delta}{2} \leq s_i + s_j \leq \delta$, then $\left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_i - s_i - s_j\right)^+ = \left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_j - s_i - s_j\right)^+ = 0$ and therefore $\xi_i = \zeta_i = 0$. Suppose $0 \leq s_i + s_j < \frac{\delta}{2}$. First, consider ξ_i. Since $\min\left\{\frac{\delta}{2}, \frac{\delta}{2}\epsilon_i + s_i\right\} = \frac{\delta}{2}\epsilon_i + s_i$ if $\epsilon_i \leq 1 - 2\frac{s_i+s_j}{\delta} \leq 1 - 2\frac{s_j}{\delta}$, we can rewrite ξ_i as $\xi_i = E\left[\left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_i - s_i - s_j\right)^+ \left(\frac{\delta}{2}\epsilon_i + s_i\right)\right]$. Evaluating this using integration by parts yields $\xi_i = \frac{\delta^2}{4}\left(1 - 2\frac{s_i+s_j}{\delta}\right)^2 F\left(1 - 2\frac{s_i+s_j}{\delta}\right) - \frac{\delta^2}{4}\left(1 - 2\frac{2s_i+s_j}{\delta}\right) G\left(1 - 2\frac{s_i+s_j}{\delta}\right) - \frac{\delta^2}{4}\int_0^{1-2\frac{s_i+s_j}{\delta}} x^2 f(x)\, dx$. Next, consider ζ_i. Since $\min\left\{\frac{\delta}{2} + \left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_j - s_j\right)^+, \frac{\delta}{2} + s_i\right\} = \frac{\delta}{2} + s_i$ if $0 \leq \epsilon_j \leq 1 - 2\frac{s_i+s_j}{\delta} \leq 1 - 2\frac{s_j}{\delta}$, we can rewrite ζ_i as $\zeta_i = E\left[\left(\frac{\delta}{2} - \frac{\delta}{2}\epsilon_j - s_i - s_j\right)^+ \left(\frac{\delta}{2} + s_i\right)\right]$. Evaluating this using integration by parts yields $\zeta_i = \frac{\delta^2}{4}\left(1 + 2\frac{s_i}{\delta}\right) G\left(1 - 2\frac{s_i+s_j}{\delta}\right)$. Combining the evaluated expressions of ξ_i and ζ_i for the two cases $\frac{\delta}{2} \leq s_i + s_j \leq \delta$ and $0 \leq s_i + s_j < \frac{\delta}{2}$ and substituting them in $\pi_i = -hs_i + p_0\sigma_i + b\theta_i\xi_i + b\theta_j\zeta_i$, we get (12). ■

B.2 Centralized Firm's Decisions

In this section we consider the centralized firm case in which a single firm makes capacity decisions for two facilities that it operates. Due to perfect substitutability, precise allocation of spare capacities s_1 and s_2 is immaterial in this case; the centralized firm needs only to choose the total spare capacity $S = s_1 + s_2$ and allocate it arbitrarily between the two facilities. For this reason, we regard S as the firm's decision variable. Let the superscript $\dagger$ denote the optimal decision. In what follows, we only focus on the symmetric case with $\theta_1 = \theta_2 = \theta$. First consider the case where price is fixed at p as in §4. In this case, the centralized firm's payoff is $\Pi = -hS + \delta p\left[1 - \theta G\left((1 - 2\frac{S}{\delta})^+\right)\right]$, and availability is $A = 1 - \theta G\left((1 - 2\frac{S}{\delta})^+\right)$.

Proposition B.1 *(Optimum under fixed price) The centralized firm chooses spare capacities as follows. If $\theta \leq \frac{h}{2p}$, then $S^\dagger = 0$ and $A^\dagger = 1 - \theta(1-\rho)$. If $\frac{h}{2p} < \theta < \frac{h}{p}$, then $S^\dagger = \frac{\delta}{2}\left(1 - F^{-1}\left(\frac{h}{2p\theta}\right)\right) < \frac{\delta}{2}$ and $A^\dagger = 1 - \theta G\left(F^{-1}\left(\frac{h}{2p\theta}\right)\right)$.*

Proof. Proof is similar to that of Proposition 1 and is omitted. ■

Proposition B.2 *At the nonzero optimum specified in Proposition B.1 for $\frac{h}{2p} < \theta < \frac{h}{p}$: (a) $\frac{dA^\dagger}{d\theta} > 0$; (b) $A^\dagger < A^*$ where A^* is the equilibrium availability in the decentralized case.*

Proof. Differentiating $A^\dagger = 1 - \theta G\left(F^{-1}\left(\frac{h}{2p\theta}\right)\right)$ from Proposition B.1 yields $\frac{\partial A^\dagger}{\partial \theta} = -G(y) + \frac{F(y)^2}{f(y)} > 0$, where $y = F^{-1}\left(\frac{h}{2p\theta}\right)$ and the inequality follows from logconcavity of F. From Proposition 1 and Proposition B.1 along with (5), we see that $A^\dagger = A^* = 1 - \theta(1-\rho)$ if $\theta \leq \frac{h}{2p}$. If $\frac{h}{2p} < \theta < \frac{h}{p}$, on the other hand, $A^\dagger = 1 - \theta G\left(F^{-1}\left(\frac{h}{2p\theta}\right)\right)$ and $A^* = 1 - \theta G\left(1 - 4\frac{\widehat{s}}{\delta}\right)$ where $\widehat{s}$ is the solution of the equation $F\left(1 - 2\frac{s}{\delta}\right) + F\left(1 - 4\frac{s}{\delta}\right) = \frac{h}{p\theta}$. Observe $\frac{h}{p\theta} = F\left(1 - 2\frac{\widehat{s}}{\delta}\right) + F\left(1 - 4\frac{\widehat{s}}{\delta}\right) > 2F\left(1 - 4\frac{\widehat{s}}{\delta}\right)$ or equivalently $1 - 4\frac{\widehat{s}}{\delta} < F^{-1}\left(\frac{h}{2p\theta}\right)$. Hence, $A^\dagger = 1 - \theta G\left(F^{-1}\left(\frac{h}{2p\theta}\right)\right) < 1 - \theta G\left(1 - 4\frac{\widehat{s}}{\delta}\right) = A^*$. ■

Next, consider the case where price varies as in §5, following (11). The centralized firm's payoff Π is equal to the sum of the decentralized firms' payoffs π_1 and π_2. Using the expression for π_i from (12) and setting $\theta_1 = \theta_2 = \theta$, we evaluate Π as follows: if $0 \leq S < \frac{\delta}{2}$, then $\Pi = -hS + \delta p_0(1 - \theta G(z)) + \frac{\delta^2 b}{2}\theta\left(2(1-z)G(z) + z^2 F(z) - \int_0^z x^2 f(x)\,dx\right)$ where $z = 1 - 2\frac{S}{\delta}$; if $\frac{\delta}{2} \leq S \leq \delta$, on the other hand, $\Pi = -hS + \delta p_0$.

Proposition B.3 *(Optimum under shortage-induced price increase) Suppose $b \leq \frac{2p_0}{3\delta}$ as in Proposition 3. Then the optimal capacity is determined as follows. If $\theta \leq \frac{h}{2}(p_0 - \delta b\rho)^{-1}$, then $S^\dagger = 0$. If $\frac{h}{2}(p_0 - \delta b\rho)^{-1} < \theta < \frac{h}{p_0}$, then $S^\dagger \in \left(0, \frac{\delta}{2}\right)$ is the unique solution of the equation $2p_0\theta F\left(1 - 2\frac{S}{\delta}\right) - 2\delta b\theta\left(F\left(1 - 2\frac{S}{\delta}\right) - G\left(1 - 2\frac{S}{\delta}\right)\right) = h$.*

Proof. Differentiating Π with respect to S yields $\frac{\partial \Pi}{\partial S} = -h + 2p_0\theta F(z) - 2\delta b\theta (F(z) - G(z))$ if $0 \leq S < \frac{\delta}{2}$ and $\frac{\partial \Pi}{\partial S} = -h < 0$ if $\frac{\delta}{2} \leq S \leq \delta$. Moreover, $\lim_{S\to 0} \frac{\partial \Pi}{\partial S} = -h + 2p_0\theta - 2\delta b\theta\rho$ and $\lim_{S\to \delta/2} \frac{\partial \Pi}{\partial S} = -h < 0$. From these observations we see that Π is maximized in the interval $0 \leq S < \frac{\delta}{2}$. The second derivative in this interval is $\frac{\partial^2 \Pi}{\partial S^2} = -4b\theta F(z) - 4\theta \left(\frac{p_0}{\delta} - b\right) f(z) < 0$, where the inequality follows from the assumption $b \leq \frac{2p_0}{3\delta}$ that implies $p_0 > \delta b$. Hence, Π is concave. Then the optimal S is found at $S = 0$ if $\lim_{S\to 0} \frac{\partial \Pi}{\partial S} \leq 0$ or equivalently $\theta \leq \frac{h}{2(p_0 - \delta b\rho)}$. If $\frac{h}{2(p_0-\delta b\rho)} < \theta < \frac{h}{p_0}$, on the other hand, the optimal S is determined from the first-order condition $\frac{\partial \Pi}{\partial S} = 0$ or equivalently $p_0 F(z) - \delta b (F(z) - G(z)) = \frac{h}{2\theta}$. ■

Proposition B.4 *At the nonzero optimum specified in Proposition B.3 for* $\frac{h}{2}(p_0 - \delta b\rho)^{-1} < \theta < \frac{h}{p_0}$*: (a)* $\frac{\partial S^\dagger}{\partial \theta} > 0$ *and* $\frac{\partial A^\dagger}{\partial \theta} > 0$*; (b)* $\frac{\partial S^\dagger}{\partial b} < 0$ *and* $\frac{\partial A^\dagger}{\partial b} < 0$.

Proof. An analysis similar to the one in the proof of Proposition 4 proves part (a). To prove part (b), differentiate the optimality condition in Proposition B.3 implicitly with respect to b to obtain $\frac{\partial S^\dagger}{\partial b} = -\frac{\delta}{2} \frac{F(z^\dagger) - G(z^\dagger)}{(p_0 - \delta b) f(z^\dagger) + \delta b F(z^\dagger)}$ where $z^\dagger = 1 - 2\frac{S^\dagger}{\delta}$. The denominator of the derived expression is positive since the condition $b \leq \frac{2p_0}{3\delta}$ in Proposition B.3 implies $p_0 > \delta b$. The numerator is also positive since $G(z) = \int_0^z F(x)\,dx \leq \int_0^z F(z)\,dx = zF(z) < F(z)$ for all $z \in (0,1)$. Combining, we have $\frac{\partial S^\dagger}{\partial b} < 0$. It then follows from $A^\dagger = 1 - \theta G\left(1 - 2\frac{S^\dagger}{\delta}\right)$ that $\frac{\partial A^\dagger}{\partial b} < 0$. ■

B.3 Capacity and Reliability Choices Under Fixed Price

In this section we extend the capacity competition model of §4 by adding a new feature: firms' efforts to reduce disruption probabilities. We call this *reliability improvement effort* and denote it as a_i, representing a reduction in firm i's disruption probability from $\overline{\theta}$ to $\theta_i = \overline{\theta} - a_i$. The upper bound $\overline{\theta}$ represents the "default" disruption probability inherent in each facility when no effort is made, and it is assumed to be identical across the two facilities. To simplify analysis, we assume that the cost of effort is equal to $r\left(1/\theta_i - 1/\overline{\theta}\right)$ where r is the coefficient of effort cost; the larger r, the more costly it is to exert the same magnitude of effort. Since there is one-to-one mapping between a_i and the realized disruption probability θ_i, we adopt the convention that firm i sets θ_i. Furthermore, we assume that the firms exert effort at time zero, when they make their decisions on spare capacity. Thus, firm i sets θ_i and s_i simultaneously, taking into account a similar move by firm j. The resulting Nash equilibrium of this reliability-capacity game is denoted by $((\theta_1^*, s_1^*), (\theta_1^*, s_1^*))$, which determines the availability A^* at that point.

Firm i's payoff is identical to (6) except for the addition of effort cost. Hence,

$$\pi_i = -r\left(\tfrac{1}{\theta_i} - \tfrac{1}{\overline{\theta}}\right) - hs_i + \tfrac{\delta p}{2}\left[1 - \theta_i G\left(1 - 2\tfrac{s_i}{\delta}\right) + \theta_j G\left(1 - 2\tfrac{s_j}{\delta}\right) - \theta_j G\left((1 - 2\tfrac{s_i + s_j}{\delta})^+\right)\right]. \quad \text{(B.1)}$$

As it turns out, analyzing the equilibrium of the reliability-capacity game with general distribution F presents tractability challenges. For this reason, in what follows we make a simplifying assumption that the yield percentage random variables ϵ_1 and ϵ_2 are both uniformly distributed and seek a symmetric equilibrium. Furthermore, we focus on the range of r satisfying $\delta p\overline{\theta}^2/16 < r < \delta p\overline{\theta}^2/4$, which rule out trivial solutions.

Proposition B.5 *Suppose that ϵ is uniformly distributed and assume $\overline{\theta} < \frac{h}{p}$ and $\underline{\underline{r}} < r < \overline{\overline{r}}$, where $\underline{\underline{r}} \equiv \frac{\delta p\overline{\theta}^2}{16}$ and $\overline{\overline{r}} \equiv \frac{\delta p\overline{\theta}^2}{4}$. Then the symmetric equilibrium $\theta_1^* = \theta_2^* = \theta^*$ and $s_1^* = s_2^* = s^*$ of the reliability-capacity game exists and is identified as follows: (a) If $\overline{\theta} \le \frac{h}{2p}$, then $\theta^* = 2\sqrt{\frac{r}{\delta p}} < \overline{\theta}$ and $s^* = 0$; (b) If $\frac{h}{2p} < \overline{\theta} < \frac{h}{p}$, then $\underline{\underline{r}} < \underline{r} < \overline{r} < \overline{\overline{r}}$ where $\underline{r} \equiv \frac{\delta h^2}{16p}$ and $\overline{r} \equiv \frac{\delta}{p}\left(\frac{h+p\overline{\theta}}{6}\right)^2$. Moreover: (i) $\theta^* = 2\sqrt{\frac{r}{\delta p}} < \overline{\theta}$ and $s^* = 0$ if $\underline{\underline{r}} < r \le \underline{r}$; (ii) $\theta^* = 6\sqrt{\frac{r}{\delta p}} - \frac{h}{p} < \overline{\theta}$ and $s^* = \delta\left(\frac{2-(h/2)\sqrt{\delta/(rp)}}{6-h\sqrt{\delta/(rp)}}\right) > 0$ if $\underline{r} < r < \overline{r}$; (iii) $\theta^* = \overline{\theta}$ and $s^* = \frac{\delta}{3}\left(1 - \frac{h}{2p\overline{\theta}}\right) > 0$ if $\overline{r} \le r < \overline{\overline{r}}$.*

Proof. For notational convenience, let $z_i \equiv 1 - 2\frac{s_i}{\delta}$ and $z_{ij} \equiv 1 - 2\frac{s_i+s_j}{\delta}$. With uniform distribution, $f(x) = 1$, $F(x) = x$, and $G(x) = \frac{x^2}{2}$ for $x \in [0,1]$. Therefore, (B.1) reduces to $\pi_i = -r\left(\frac{1}{\theta_i} - \frac{1}{\overline{\theta}}\right) - hs_i + \frac{\delta p}{2}\left(1 - \frac{\theta_i}{2}z_i^2 + \frac{\theta_j}{2}z_j^2 - \frac{\theta_j}{2}z_{ij}^2\right)$ if $0 \le s_i + s_j \le \frac{\delta}{2}$ and $\pi_i = -r\left(\frac{1}{\theta_i} - \frac{1}{\overline{\theta}}\right) - hs_i + \frac{\delta p}{2}\left(1 - \frac{\theta_i}{2}z_i^2 + \frac{\theta_j}{2}z_j^2\right)$ if $\frac{\delta}{2} < s_i + s_j \le \delta$. Differentiating this with respect to s_i yields $\frac{\partial \pi_i}{\partial s_i} = -h + p\theta_i z_i + p\theta_j z_{ij}$ if $0 \le s_i + s_j \le \frac{\delta}{2}$ and $\frac{\partial \pi_i}{\partial s_i} = -h + p\theta_i z_i < 0$ if $\frac{\delta}{2} < s_i + s_j \le \delta$, where the last inequality follows from the assumptions $\theta_i \le \overline{\theta}$ and $\overline{\theta} < \frac{h}{p}$. The inequality implies that the maximizer of π_i does not exist in the region $\frac{\delta}{2} < s_i + s_j \le \delta$. Hence, if the equilibrium exists, it is found in the region $0 \le s_i + s_j \le \frac{\delta}{2}$ where $\frac{\partial \pi_i}{\partial a_i} = -\frac{r}{\theta_i^2} + \frac{\delta p}{4}z_i^2$ and $\frac{\partial^2 \pi_i}{\partial a_i^2} = -\frac{2r}{\theta_i^3} < 0$. Observe that $\frac{\partial \pi_i}{\partial a_i}$ is independent of firm j's decision variables a_j and s_j, implying that firm i's optimal choice for a_i interacts only with his own capacity s_i. This implies that the strategic interaction between the two firms occurs via s_i and s_j only, and as a result, the simultaneous-move game with two sets of decision variables (a_i, s_i) and (a_j, s_j) is reduced to the game with one set of variables (s_i, s_j) with the optimal value for a_i determined by s_i. Since π_i is concave in $a_i \in [0, \overline{\theta})$ for any given s_i and $\lim_{a_i \to \overline{\theta}} \frac{\partial \pi_i}{\partial a_i} = -\infty < 0$, it is optimal for the firm to set $a_i = 0$ (or equivalently $\theta_i = \overline{\theta}$) if $\lim_{a_i \to 0} \frac{\partial \pi_i}{\partial a_i} = -\frac{r}{\overline{\theta}^2} + \frac{\delta p}{4}z_i^2 \le 0$ and set $a_i > 0$ (or $\theta_i < \overline{\theta}$) that solves the first-order condition $\frac{\partial \pi_i}{\partial a_i} = 0$ otherwise. Rewriting these conditions, we see that for a given value of s_i, firm i sets $a_i = \overline{\theta} - \theta_i$ such that

$$\theta_i = \begin{cases} \frac{2}{z_i}\sqrt{\frac{r}{\delta p}} < \overline{\theta} & \text{if} \quad 0 \le s_i < \overline{s}, \\ \overline{\theta} & \text{if} \quad \overline{s} \le s_i \le \frac{\delta}{2}, \end{cases} \tag{B.2}$$

where $\overline{s} \equiv \frac{\delta}{2} - \frac{1}{\overline{\theta}}\sqrt{\frac{\delta r}{p}}$. Note that $\overline{s} < \frac{\delta}{4}$ under the assumption $\frac{\delta p\overline{\theta}^2}{16} = \underline{\underline{r}} < r$ stated in the proposition.

Substituting (B.2) in π_i yields the reduced payoff $\widetilde{\pi}_i$ as a function of s_i and s_j, defined separately for the following four regions: (i) $0 \leq s_i < \overline{s}$ and $0 \leq s_j < \overline{s}$, (ii) $0 \leq s_i < \overline{s}$ and $\overline{s} \leq s_j \leq \frac{\delta}{2}$, (iii) $\overline{s} \leq s_i \leq \frac{\delta}{2}$ and $0 \leq s_j < \overline{s}$, and (iv) $\overline{s} \leq s_i \leq \frac{\delta}{2}$ and $\overline{s} \leq s_j \leq \frac{\delta}{2}$. Differentiating $\widetilde{\pi}_i$ with respect to s_i, we get (i) $\frac{\partial \widetilde{\pi}_i}{\partial s_i} = -h + 2\sqrt{\frac{rp}{\delta}} + 2\sqrt{\frac{rp}{\delta}}\frac{z_{ij}}{z_j}$, (ii) $\frac{\partial \widetilde{\pi}_i}{\partial s_i} = -h + 2\sqrt{\frac{rp}{\delta}} + p\overline{\theta}z_{ij}$, (iii) $\frac{\partial \widetilde{\pi}_i}{\partial s_i} = -h + p\overline{\theta}z_i + 2\sqrt{\frac{rp}{\delta}}\frac{z_{ij}}{z_j}$, and (iv) $\frac{\partial \widetilde{\pi}_i}{\partial s_i} = -h + p\overline{\theta}z_i + p\overline{\theta}z_{ij}$ for each of the four regions above. Differentiating this again proves that $\frac{\partial^2 \widetilde{\pi}_i}{\partial s_i^2} < 0$ everywhere. Hence, $\widetilde{\pi}_i$ is concave and therefore the Nash equilibrium exists (Cachon and Netessine 2004). From the above expressions, we see that the interior symmetric equilibrium $s_i^* = s_j^* = s^*$ satisfying $0 < s^* < \frac{\delta}{4}$ is identified by the solution of the first-order condition $w(s) \equiv \left.\frac{\partial \widetilde{\pi}_i}{\partial s_i}\right|_{s_i = s_j = s} = 0$, where $w(s) = -h + 2\sqrt{\frac{rp}{\delta}} + 2\sqrt{\frac{rp}{\delta}}\frac{1-4s/\delta}{1-2s/\delta}$ if $0 \leq s < \overline{s}$ and $w(s) = -h + 2p\overline{\theta}\left(1 - 3\frac{s}{\delta}\right)$ if $\overline{s} \leq s \leq \frac{\delta}{4}$. It is easily verified that $w(s)$ is a decreasing function. At the upper bound $s = \frac{\delta}{4}$, we have $w\left(\frac{\delta}{4}\right) = -h + \frac{p\overline{\theta}}{2} < 0$ by the assumption $\overline{\theta} < \frac{h}{p}$, which implies that the solution of the equation $w(s) = 0$ is nonzero if and only if $w(0) > 0$. Hence, the symmetric equilibrium s^* is found in the interior of the defined interval $\left[0, \frac{\delta}{4}\right]$ if and only if $w(0) = -h + 4\sqrt{\frac{rp}{\delta}} > 0$; otherwise the equilibrium is $s^* = 0$. We now consider the cases stated in the proposition.

(a) Suppose $\overline{\theta} \leq \frac{h}{2p}$. Combining this condition with the assumption $r < \overline{\overline{r}} = \frac{\delta p \overline{\theta}^2}{4}$, we see that $w(0) = -h + 4\sqrt{\frac{rp}{\delta}} < -h + 2p\overline{\theta} \leq 0$, which implies $s^* = 0$ by the argument above. At this value, $\theta^* = 2\sqrt{\frac{r}{\delta p}} < \overline{\theta}$ (see (B.2)).

(b) Suppose $\frac{h}{2p} < \overline{\theta} < \frac{h}{p}$. The condition $\overline{\theta} < \frac{h}{p}$ can be rewritten as $\underline{\underline{r}} = \frac{\delta p \overline{\theta}^2}{16} < \frac{\delta h^2}{16p} = \underline{r}$. Moreover, the condition $\frac{h}{2\overline{\theta}} < p$ implies $\frac{h}{4} < \frac{h + p\overline{\theta}}{6}$, which can be rewritten as $\underline{r} = \frac{\delta h^2}{16p} < \frac{\delta}{p}\left(\frac{h+p\overline{\theta}}{6}\right)^2 = \overline{r}$. Finally, the condition $\frac{h}{2\overline{\theta}} < p$ implies $8\left(p\overline{\theta} - \frac{h}{2}\right)\left(p\overline{\theta} + \frac{h}{4}\right) = 8p^2\overline{\theta}^2 - 2hp\overline{\theta} - h^2 > 0$, which can be rewritten as $\overline{r} = \frac{\delta}{p}\left(\frac{h+p\overline{\theta}}{6}\right)^2 < \frac{\delta p \overline{\theta}^2}{4} = \overline{\overline{r}}$. In summary, we have $\underline{\underline{r}} < \underline{r} < \overline{r} < \overline{\overline{r}}$. Consider the three cases listed in part (b) in turn. If $\underline{\underline{r}} < r \leq \underline{r} = \frac{\delta h^2}{16p}$, we have $w(0) = -h + 4\sqrt{\frac{rp}{\delta}} \leq 0$, which implies $s^* = 0$ by the argument above. At this value, $\theta^* = 2\sqrt{\frac{r}{\delta p}} < \overline{\theta}$ (see (B.2)). If $\frac{\delta h^2}{16p} = \underline{r} < r < \overline{r} = \frac{\delta}{p}\left(\frac{h+p\overline{\theta}}{6}\right)^2$, we have $w(0) = -h + 4\sqrt{\frac{rp}{\delta}} > 0$ and $w(\overline{s}) = -h - p\overline{\theta} + 6\sqrt{\frac{rp}{\delta}} < 0$, which implies that $w(s)$ crosses zero in the interval $(0, \overline{s})$. From the definition of $w(s)$ we see that the equilibrium is found at a value of s that solves the equation $w(s) = -h + 2\sqrt{\frac{rp}{\delta}} + 2\sqrt{\frac{rp}{\delta}}\frac{1-4s/\delta}{1-2s/\delta} = 0$, from which we get $s^* = \delta\left(\frac{2 - (h/2)\sqrt{\delta/(rp)}}{6 - h\sqrt{\delta/(rp)}}\right)$. At this value, $\theta^* = 6\sqrt{\frac{r}{\delta p}} - \frac{h}{p}$ (see (B.2)). Finally, if $\frac{\delta}{p}\left(\frac{h+p\overline{\theta}}{6}\right)^2 = \overline{r} \leq r < \overline{\overline{r}} = \frac{\delta p \overline{\theta}^2}{4}$, we have $w(\overline{s}) = -h - p\overline{\theta} + 6\sqrt{\frac{rp}{\delta}} > 0$, which together with the earlier observation $w\left(\frac{\delta}{4}\right) < 0$ that $w(s)$ crosses zero in the interval $\left(\overline{s}, \frac{\delta}{4}\right)$. From the definition of $w(s)$ we see that the equilibrium is found at a value of s that solves the equation $w(s) = -h + 2p\overline{\theta}\left(1 - 3\frac{s}{\delta}\right) = 0$, from which we get $s^* = \frac{\delta}{3}\left(1 - \frac{h}{2p\overline{\theta}}\right)$. At this value, $\theta^* = \overline{\theta}$ (see (B.2)). ■

Corollary B.1 *At the symmetric equilibrium identified in Proposition B.5, availability is evaluated*

as follows: (a) If $\overline{\theta} \leq \frac{h}{2p}$*, then* $A^* = 1 - \sqrt{\frac{r}{\delta p}}$*; (b) If* $\frac{h}{2p} < \overline{\theta} < \frac{h}{p}$*, (i)* $A^* = 1 - \sqrt{\frac{r}{\delta p}}$ *if* $\underline{\underline{r}} < r \leq \underline{r}$*; (ii)* $A^* = 1 - \frac{1}{2}\sqrt{\frac{r}{\delta p}}\frac{\left(h\sqrt{\delta/(rp)}-2\right)^2}{6-h\sqrt{\delta/(rp)}}$ *if* $\underline{r} < r < \overline{r}$*; (iii)* $A^* = 1 - \frac{\overline{\theta}}{18}\left(\frac{2h}{p\overline{\theta}} - 1\right)^2$ *if* $\overline{r} \leq r < \overline{\overline{r}}$*.*

Proposition B.6 *Under the conditions in Proposition B.5 that result in the symmetric interior equilibrium with* $\theta^* < \overline{\theta}$ *and* $s^* > 0$*,* $\frac{\partial A^*}{\partial r} > 0$*.*

Proof. From Proposition B.5, we see that the conditions that result in the symmetric equilibrium are $\frac{h}{2p} < \overline{\theta} < \frac{h}{p}$ and $\underline{r} < r < \overline{r}$ where $\underline{r} \equiv \frac{\delta h^2}{16p}$ and $\overline{r} \equiv \frac{\delta}{p}\left(\frac{h+p\overline{\theta}}{6}\right)^2$. In this case $A^* = 1 - \frac{1}{2}\sqrt{\frac{r}{\delta p}}\frac{\left(h\sqrt{\delta/(rp)}-2\right)^2}{6-h\sqrt{\delta/(rp)}}$, according to Corollary B.1. Differentiating A^* with respect to r yields $\frac{\partial A^*}{\partial r} = \frac{\delta}{2r^3}\left(\frac{r}{\delta p}\right)^{3/2}\left(6 - h\sqrt{\frac{\delta}{rp}}\right)^{-2}\omega(r)$, where $\omega(r) \equiv \delta h^2 - 12rp + 4h\sqrt{\delta rp}$. We prove that $\omega(r) > 0$ in the considered interval $[\underline{r}, \overline{r}]$. Evaluating $\omega(r)$ at the lower and upper bounds, we get $\omega(\underline{r}) = \frac{5}{4}\delta h^2 > 0$ and $\omega(\overline{r}) = \frac{\delta h^2}{3}\left(4 - \left(\frac{p\overline{\theta}}{h}\right)^2\right) > \delta h^2 > 0$, where we used the condition $\overline{\theta} < \frac{h}{p}$ to prove the inequality. Moreover, $\omega'(r) = -12p + 2h\sqrt{\frac{\delta p}{r}}$ with $\omega'(\underline{r}) = -4p < 0$ and $\omega''(r) < 0$. Together, these results show that $\omega(r)$ starts from a positive number at $r = \underline{r}$, concave decreasing until it reaches another positive number at $r = \overline{r}$. Therefore, $\omega(r) > 0$ for all $r \in [\underline{r}, \overline{r}]$. It then follows that $\frac{\partial A^*}{\partial r} > 0$. ■