Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Patient Choice in Kidney Allocation: The Role of the Queueing Discipline

Xuanming Su, Stefanos Zenios,

To cite this article:
Xuanming Su, Stefanos Zenios, (2004) Patient Choice in Kidney Allocation: The Role of the Queueing Discipline. Manufacturing & Service Operations Management 6(4):280-301. http://dx.doi.org/10.1287/msom.1040.0056

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 6, No. 4, Fall 2004, pp. 280–301
ISSN 1523-4614 | EISSN 1526-5498 | 04 | 0604 | 0280

informs

DOI 10.1287/msom.1040.0056

Patient Choice in Kidney Allocation: The Role of the Queueing Discipline

Xuanming Su
Walter A. Haas School of Business, University of California, Berkeley, California 94720,
xuanming@haas.berkeley.edu

Stefanos Zenios
Graduate School of Business, Stanford University, Stanford, California 94305, stefzen@exch-gsb.stanford.edu

This paper develops and analyzes a queueing model to examine the role of patient choice on the high rate of organ refusals in the kidney transplant waiting system. The model is an M/M/1 queue with homogeneous patients and exponential reneging. Patients join the waiting system and organ transplants are reflected by the service process. In addition, unlike the standard M/M/1 model, each service instance is associated with a variable reward that reflects the quality of the transplant organ, and patients have the option to refuse an organ (service) offer if they expect future offers to be better. Under an assumption of perfect and complete information, it is demonstrated that the queueing discipline is a potent instrument that can be used to maximize social welfare. In particular, first-come-first-serve (FCFS) amplifies patients' desire to refuse offers of marginal quality, and generates excessive organ wastage. By contrast, last-come-first-serve (LCFS) contains the inefficiencies engendered by patient choice and achieves optimal organ utilization. A numerical example calibrated using data from the U.S. transplantation system demonstrates that the welfare improvements possible from a better control of patient choice are equivalent to a 25% increase in the supply of organs.

Key words: kidney allocation; queues; priority; last-come-last-served; stochastic games; efficiency-equity trade-off
History: Received: June 1, 2002; accepted: July 15, 2004. This paper was with the authors 7 months for 3 revisions.

1. Introduction

Kidney transplantation is the preferred treatment for most patients diagnosed with chronic kidney failure. However, there is a significant shortage of organs for transplantation as demonstrated by the expanding transplant waiting list. In the 10-year period between 1992 and 2002, the national waiting list grew from 22,063 to 51,144, and the median waiting time increased from 624 days to 1,144 days (UNOS 2002). Despite these alarming trends, more than 12% of all organs recovered from a donor are eventually discarded, because they are repeatedly refused for transplantation by patients on the waiting list and by their surgeons. The most common reason for these refusals is that these organs are of marginal quality, and hence the patient expects to benefit by waiting for a better organ (UNOS 2002).

Recognizing the inefficiency created by these refusals, the United Network of Organ Sharing (UNOS), which is charged with the management of the transplant waiting system has recently introduced the following modification to the transplant allocation system: Organs from marginal donors are now reserved for patients who declare in advance their willingness to accept such organs. While these patients remain eligible for all other organs, they are likely to be offered an organ from a marginal donor several months or even years before they would expect an offer of a "regular" donor organ. Consequently, UNOS expects to place more of these "marginal" organs and to diminish their wastage by accelerating patient's access to them. Yet, this proposed modification is not free of problems. Specifically, while increasing patient involvement in the transplant allocation system is desirable, there are no guarantees that this will improve the overall system. In fact, readers of this journal are familiar with the classical result in Naor (1969), where consumer choice in a waiting system increases congestion and degrades performance. Is it

possible that patient choice in the transplant waiting list can lead to more wastage and longer waiting times? This paper aims to provide an answer to this question by focusing on the interaction between patient choice and the queueing discipline used to rank patients on the transplant waiting list.

In our view, the effect of patient choice on the transplant allocation system is dictated by two complementary forces. First, depending on their intrinsic characteristics, some patients may be more likely to accept organs of marginal quality than others. Second, the rule used to prioritize patients on the transplant waiting list can encourage patients to be either more or less stringent in their choices. While understanding the interaction between these two forces is desirable, our approach is based on the principle of "divide and conquer." In two companion papers (Su and Zenios 2004a, b) we focus on the role of patient heterogeneity, while suppressing the dynamics of the prioritization rule. By contrast, the focus here is on the effect of the prioritization rule with patient heterogeneity suppressed. The emerging picture is incomplete, but it can be argued that some of the findings obtained from the simplified setting are universally valid.

A queueing model provides the natural abstraction for our investigation. The model incorporates two independent arrival streams, one for patients and a second one for organs, and "service" reflects the transplant operation when a patient and an organ are merged and depart the system. Because organs cannot be stored, and thus cannot be placed in a queue, their arrival can be conveniently captured by the "service" process with the service time equal to the time between organ arrivals. This implies that the basic model is an M/M/1 queue, but with two unique features that deviate from the standard assumptions and reflect the reality of the transplant waiting system. First, each "service" offer is associated with a reward that captures the quality of the organ. And second, patients may refuse an organ offer if they perceive its quality to be unacceptable and expect a future offer to be of better quality. Although the refused organ will then be offered to the next patient in line, it is theoretically possible for the resulting queueing model to "idle" even when the queue is not empty, because the quality of some organs may make them unattractive for all patients on the transplant waiting list. The model also captures patient death through exponential patient reneging, and it assumes that all patients are homogeneous despite their well-documented heterogeneity. As explained above, the latter is a conscious choice made primarily for tractability's sake. To reflect the difference in medical outcomes between waiting and transplantation, the model also assumes that each patient on the waiting list receives a reward per unit time that reflects their quality of life prior to a transplant, and when the patients accept an organ offer, they also receive a reward given by the patient's expected discounted quality-adjusted life years (QALYs) after a transplant.

Patients in this system behave as rational economic agents and determine whether to accept or decline each offer based on the offer's quality. Similarly, the medical planner overseeing the system determines who should be offered each organ. Each patient's objective is to maximize his or her own total expected discounted QALYs, and hence the patient solves an optimal stopping time problem: when to accept an organ offer. On the other hand, the medical planner wishes to maximize the sum of the rewards for all patients and has two policy levers at its disposal: organ rationing and patient prioritization. That is, the planner can influence system outcomes by limiting access to certain organs (rationing) and by dynamically prioritizing the candidates on the waiting list. Of course, in reality, there is not a single planner, but rather a community of stakeholders that collectively determine the rationing and prioritization rule. The monolithic medical planner in our model reflects the collective actions of this community.

Our analysis starts with a frictionless ideal, where the medical planner is a benevolent dictator and the patients do not exercise choice, but instead accept any offer. This identifies the socially efficient (or Pareto optimal) outcome and provides a comparison benchmark for the effects of patient choice. In the context of this system, patient prioritization is not effective because patients are identical, but rationing is effective because there is a tradeoff between waiting and transplanting a not-so-good organ. In fact, when the queue length is small, the planner only offers organs of the best quality. But as the queue length increases,

organs of lower quality are also offered to counterbalance the increased wait.

The next step is to consider the competitive equilibria that will emerge when patients exercise their choice. We assume first that patients are ranked according to the FCFS discipline, and investigate whether, given this priority discipline, rationing increases the planner's objective. It turns out that rationing is ineffective because it is confounded by patient choice. Specifically, in the absence of explicit rationing by the planner, "implicit" rationing will emerge as the equilibrium outcome: Organs of low quality will be discarded because no one on the waiting list would accept them. Further, the quality threshold that separates the acceptable organs from the ones that are declined is decreasing in the queue length. Imposing rationing that raises these thresholds beyond their naturally occurring equilibrium does not benefit the planner in any way. These equilibrium thresholds are higher than the ones obtained in the socially efficient case, mirroring the results in Naor (1969), where customer autonomy in a queueing system caused performance degradation. In our context, patients' quality requirements in the FCFS system are more stringent than is socially optimal, and their behavior causes excessive organ wastage.

Next, we turn to investigate whether the ineffectiveness of rationing is an artifact of the FCFS priority system. Our analysis demonstrates that it is not: treatment rationing is always ineffective, and hence, the only meaningful lever for the medical planner to use is the priority discipline. This motivates the following question: What is the priority discipline that maximizes social welfare when patients exercise choice? An examination of the FCFS discipline demonstrates that it is ineffective because future arrivals do not affect the patients already in the system, hence existing patients do not consider the congestion externalities they impose when they decline an organ offer. Motivated by this observation, we then consider the extreme opposite priority rule—LCFS—and demonstrate that, in this rule, patients internalize the externalities of their own decisions, and system performance achieves the socially optimal ideal.

This brings to the forefront the equity-efficiency trade-off that underlies any medical waiting system. Specifically, it is a fundamental premise in medicine that FCFS is fair, and in that respect, LCFS can be blatantly unfair. It is therefore natural to consider a priority rule that is not as inefficient as FCFS but not as unfair as LCFS. To do that, we introduce a family of prioritization schemes, with both FCFS and LCFS belonging to that family, and examine the effect of these schemes. The main finding formalizes the intuition that in systems judged fair according to the FCFS criterion, patients internalize only a small fraction of the externalities caused by their autonomy, but in systems that are unfair according to the same criterion, patients internalize most of the externalities. This casts serious doubts on the validity of the premise that FCFS is a gold standard for fairness. When patients exercise choice, FCFS is the most "unfair" of all policies because the externalities of patient autonomy are borne by everyone *except* the person exercising this autonomy.

The root of the inefficiency identified by this analysis lies in the well-known divergence between individual and social optima in queueing systems. This was first studied in Naor's (1969) seminal paper, which demonstrated that an admission toll is necessary to induce the social optimum. Yechiali (1971) considers the perspective of a profit-maximizing firm, and Lippman and Stidham (1977) and Stidham (1978) study the structural properties of the optimal congestion toll. Mendelson (1985) embeds the queueing model into an economic framework that directly considers the effect of such externalities. The situation becomes more complex when there are multiple customer types, because customers now have the added incentive to misrepresent themselves to obtain a better service level. Mendelson and Whang (1990) derive incentive-compatible priority pricing policies that simultaneously induce truthtelling and socially optimal behavior, and Van Mieghem (2000) combines this with dynamic scheduling policies. Another instance of similar incentive problems arises in the context of multiserver systems. Bell and Stidham (1983) show that when arriving customers are free to choose which server to join, low-cost/high-speed servers will become overcongested.

Compared with all these papers, our work exhibits important differences in modeling methodology, policy implications, and solution techniques. First, while most existing models are concerned with customers'

decisions at arrival epochs (e.g., whether to join the queue, what priority level to purchase, and which server to choose), we are interested in patients' decisions to accept transplantation at service epochs (which correspond to the times when organs become available). Kaplan (1987) studies a similar queueing model in the context of public housing programs, in which registrants may refuse to move into a public housing project if they would rather wait for a more attractive accommodation option. However, while he models tenant choice as an exogenous acceptance probability, the queueing dynamics in our model account for patients' utility-maximizing decisions. Next, the current work also yields entirely different policy implications. While intelligent pricing is essential in previous work, we show that prioritization alone is sufficient to coordinate the system studied here. Bell and Stidham (1983) mention pricing under the LCFS discipline, and Hassin (1985) provides remarks on the optimality of LCFS. However, to the best of our knowledge, no other work has rigorously showed that the socially optimal ideal can be attained by merely changing the priority rule to LCFS. In fact, the paper by Hassin (1985) provides an informal argument favoring LCFS but not a formal proof. Finally, while most papers in the literature assume that queue lengths are not observable and focus on equilibrium analysis, we allow patients to observe queue lengths and analyze this less tractable case using stochastic game techniques. The only other methodologically similar paper that we know of is Altman and Shimkin (1998), which analyzes individual equilibrium decisions to join a processor-sharing system.

In the health economics literature, most work on medical queueing systems is based on economic models that use cost benefit analysis to quantify a socially optimal waiting list configuration (for a comprehensive survey, see Cullis et al. 2000). Incentive considerations in waiting lists focus mainly on the attitudes of physicians. Among others, Yates (1995) expresses concern for the possibility that the pursuit of private practice by consultants in the United Kingdom's National Health Services (NHS) may create a conflict of interest, and Weinstein (2001) contemplates the dual role of physicians as gatekeepers. The incentives of hospital management have also been studied, such as in Feldman and Lobo (1997), although to a lesser extent. The papers by Goddard et al. (1995) and Iversen (1997) examine patient preferences, but differ from the current study in that they consider patients' choices between seeking treatment from the waiting list or the private sector, whereas we are concerned with the behavior of patients while on the waiting list. Furthermore, the majority of papers in this area focus on static equilibrium models. Some exceptions include Goddard and Tavakoli (1994) who present a queueing analysis of the impact of various prioritization regimes, and Van Ackere and Smith (1999) and Smith and Van Ackere (2002) who model the NHS waiting lists, using system dynamics methodology. Most of these papers assume that patients' choices to seek treatment are exogenously given (with the exception of Van Ackere and Smith 1999, where patient arrival rates depend on perceived waiting times). In contrast, we examine patients' utility-maximizing decisions by explicitly modeling their utility functions. We hope that our dynamic analysis of patient choice in the context of kidney transplantation will shed some light on the trade-offs that have not been previously studied.

The kidney allocation problem creates some of modern medicine's most vexing policy dilemmas, and its study would benefit from the rigor provided by queueing models. Extensive simulation studies have been pursued in an effort to clarify the role of different allocation policies on queueing outcomes such as waiting time (see Zenios et al. 2002, Howard 2001, Votruba 2002), and Zenios (1999) provides some analytical results using a multiclass queueing model with reneging. However, these papers suppress patient choice and fail to recognize its critical role.

We now provide an outline for the remainder of this paper. In §2, we provide an overview of kidney transplantation in the United States. A description of the model is presented in §3. Section 4 analyzes the socially optimal outcome when patients do not exercise choice. Section 5 considers the competitive equilibrium that emerges under the FCFS priority discipline. A comparison of the socially optimal outcome to the outcomes from the competitive equilibrium verifies a welfare loss, whose source is examined in detail in §6. The subsequent two sections verify that prioritization is a potent policy instrument. Specifically,

§7 shows that the socially efficient outcome can be recovered by using LCFS, and §8 discusses a family of prioritization schemes that capture the essence of the equity-efficiency trade-off underlying the selection of different priority schemes. Section 9 presents a numerical example that investigates the magnitude of the welfare losses under FCFS and the model parameters that contribute to these losses. Concluding remarks are presented in §10. All proofs are relegated to the appendix.

2. Background on Kidney Transplantation

We now present an overview of kidney transplantation in the United States focusing on its aspects most related to the questions addressed in this paper. Readers who wish to obtain more details should consult UNOS (2002) and Organ Procurement and Transportation Network (OPTN) (2003).

End-Stage Renal Disease (ESRD) or chronic kidney failure is a condition affecting more than 400,000 patients in the United States as of 2002. Most of these patients undergo dialysis therapy, which involves visiting a dialysis center for at least 12 hours each week. The other treatment alternative is kidney transplantation, which is often preferred because it enables patients to resume regular activities. However, to minimize the risk of acute graft rejection (graft is the medical term for the transplant organ), patients who have received a transplant must take immunosuppressive drugs indefinitely. Although some patients are able to obtain a kidney from a living relative, most must rely on cadaveric donors (that is, deceased donors) and continue to receive dialysis treatment while waiting for a suitable organ.

In 2002, there were 23,328 new ESRD patients but only 11,860 cadaveric organs were procured in the same year. UNOS oversees the allocation of organs to transplant candidates by coordinating the activities of 59 organ procurement organizations (OPO) that operate in distinct geographical regions. Individual OPOs are responsible for procuring all organs donated in their region, and patients join the waiting list at their local transplant centers. Typically, when a kidney becomes available, it is first allocated to patients at the local level, then at the regional level, and then at the national level. Furthermore, because post-transplant survival relies heavily on clinical variables such as age and tissue types, UNOS attempts to promote favorable transplant outcomes by providing the OPOs with specific organ allocation guidelines. This includes a point system that prioritizes potential transplant recipients based on points that reflect the quality of the tissue match between the donor and the candidate, as well as points that reflect the candidate's waiting time. The continued shortage of organs and the associated explosion in waiting times has contributed to a convergence of this point system to a system that resembles FCFS.

Despite the apparent supply shortage, nearly half of the procured organs are refused by the first-offered transplant candidate. Placement of these refused organs is not an easy task because a substantial number of patients may prefer to remain on dialysis and wait for a better organ offer, rather than obtain an organ of marginal quality. During the search for a recipient, organs accumulate cold ischemia time (during which they are kept frozen), which causes transplant outcomes to deteriorate, and organs are usually discarded if they are not transplanted within 48 hours. As a result, in 2001, about 12% of all recovered kidneys were not transplanted, thus exacerbating the supply shortage. The modified policy described in the introduction was developed in an effort to increase the utilization of these marginal organs; this policy is formally known as the Expanded Criteria Donor program and came into effect in October 2002. However, the effectiveness of this new policy and other policies that encourage patient choice has yet to be examined rigorously.

3. Model Description

We now introduce the queueing model for the local transplant waiting list. Patients arrive according to a time-homogeneous Poisson process with rate λ, and cadaveric organs arrive according to an independent Poisson process with rate μ. Patients can depart from the waiting list either when they receive and accept an organ offer, or when they die after an exponentially distributed amount of time with mean $1/\gamma$. The death process is independent from both patient and organ arrivals. For brevity of notation, it is convenient to normalize the organ arrival rate $\mu \equiv 1$.

The reward structure for the model assumes that patients are assigned a quality of life score that depends on whether they are waiting, they have received a transplant, or they have died. The quality of life score reflects the desirability of each of the three states (dialysis, post-transplant, death) and patient preferences are homogeneous in the sense that all patients have the same quality of life score in each state. Patients on dialysis receive a continuous payoff at a rate h per unit time. Patients who die receive an instantaneous payoff d. In addition, patients receive a payoff from transplantation. To reflect variability in organ types, we assume that there exists a measure of "organ quality" captured by a continuous random variable X, which takes values in $(\underline{x}, \bar{x})$ and has probability density f. When a donor organ arrives, its quality $X = x$ is realized and publicly observed, and this reflects the post-transplant total expected discounted QALYs for the patient receiving the organ. All rewards are continuously discounted at rate β. For future reference, it is also convenient to define the reverse cumulative probability distribution $\bar{F}(x) = \int_x^{\bar{x}} f(x)\,dx$ and also the function $g(x) = \int_x^{\bar{x}} xf(x)\,dx/\bar{F}(x)$, which gives the expected conditional reward from organs with quality that exceeds a threshold x.

Patients have the option to decline an organ offer in anticipation of a potentially better future offer, and their objective is to exercise their "decline" option carefully to maximize their total quality-adjusted life expectancy. A policy for a patient specifies the range of donor organs that are acceptable to the patient at each point in time. Similarly, the medical planner (referred to as "she") must decide how to allocate the organs that become available and her objective is to maximize the total quality-adjusted expected life years of all patients. A policy for the medical planner specifies, at each point in time, the range of donor organs that will be assigned to patients waiting and the rank of all patients on the waiting list. An organ is first assigned to the top-ranked candidate. If she or he declines it, the organ will be offered to the second-ranked candidate. The process will be repeated until the organ is either accepted by someone on the waiting list or it is refused by everyone. The organ is then discarded.

Throughout the analysis, the focus will be on stationary Markovian strategies for all parties. It will also be assumed that each patient and the medical planner have perfect information about the queue length and the quality of organs. All model primitives are known to all parties.

The current specification is quite general and does not require additional assumptions on the parameters. There are no restrictions on the signs of h or d, although finiteness on the bounds $\underline{x}, \bar{x}$ is required. For ease of exposition, we will develop our analysis assuming that the quality random variable X has a continuous density f, but this assumption is not necessary; for example, X could be a discrete random variable. Furthermore, the analysis can be be carried over to the undiscounted case by taking the limit $\beta \to 0$.

Before we proceed further, it is worthwhile to describe the modeling assumptions that deviate from the real-life kidney allocation system reviewed in §2. While it is virtually impossible to develop a tractable model that captures every single aspect of reality, it is nonetheless important to discuss how the simplifications made in the model affect the validity of our findings. The following list is not exhaustive, but it attempts to cover the assumptions that represent the most significant deviations from reality.

3.1. Homogeneous Patients

As explained in the introduction, heterogeneity in patient characteristics is suppressed from our model. The ESRD patient population exhibits wide differences in various clinical attributes (such as age and tissue types) and the outcome from a transplant depends both on the attributes of the organ (the so-called organ type) and the clinical attributes of the patient (the so-called patient type). By suppressing patient heterogeneity, our model assumes that the organ type is a much more important predictor for the outcome of a transplant than the patient type. In fact, survival analysis performed using transplant data from the United States Renal Data System (USRDS) partially supports this assumption: donor age is the factor that most significantly influences post-transplant survival (see Su 2004). Therefore, incorporating only the variability in organ types and not in patient types captures the most important driver of transplant outcomes. This is not to say that variability in patient attributes is any less important. In fact, its interaction with patient choice is studied

in Su and Zenios (2004a), where it is shown that organ types can be partitioned into domains such that organs in each domain are offered to patients of a given type or types. Further, appropriately designing the partition and the assignment of organ domains to patient types minimizes the wastage caused by patient choice.

Beyond the suppressed patient heterogeneity, patients may also be better informed about unobservable risk factors that influence transplant outcomes, and about their preferences among different health states. This information asymmetry introduces another important dimension of strategic behavior that may cause wastage: patients may provide misleading information about their true preferences and health conditions to improve their individual outcomes. Such behavior may exacerbate the problems of patient choice, and its study is left as a topic for future research.

3.2. Perfect Information

The second unrealistic assumption made in our model is that patients have perfect information about the size of the waiting list. Although patients (or at least their physicians) are kept updated about their position on the waiting list, fluctuations in the waiting list because of additions and removals make it difficult to continuously monitor the state of the system. However, as comprehensive information systems are made publicly available over the Internet, the assumption of perfect information will become increasingly relevant. In fact, the website maintained by UNOS provides regular updates about the size of the waiting list, new additions, and any departures. Patients can use this information to intelligently monitor changes in their own wait list position.

3.3. No Organ Deterioration

As mentioned in §2, donor organs accumulate cold ischemia time during the search for a recipient and this causes a degradation in outcomes that is not captured in our model. Theoretically, this is not an important issue because our model permits us to identify the patient (if any) who would accept the organ, and thus placement of the organ can be almost instantaneous. However, contacting that particular individual imposes logistical challenges, and moreover, that individual may not be perfectly rational as assumed in our model, and hence may refuse the offer. Therefore, the reader should assume that our analysis will understate the welfare loss caused by patient choice, and will potentially overestimate the improvements generated by any of our proposals.

3.4. Geographical Factors

In reality, when the local waiting list is exhausted, UNOS will attempt to allocate the kidney to patients on the regional or even national waiting list. However, this incurs substantial time delays (e.g., when transporting the kidney across long distances), which result in organ deterioration as described above. In our model, we simply assume that the kidney is discarded after being refused by all local patients. This captures the negative effects of organ refusal in an imperfect way, and may overstate the benefits of our proposals.

3.5. Social Welfare Function

Our analysis assumes that the medical planner is interested in maximizing the total welfare of all patients participating in the organ allocation system, but is not concerned with other sources of social costs such as the financial cost of treatment. Incorporating health care costs into the analysis provides a different perspective not pursued here.

4. Socially Optimal Outcome

Having described our modeling framework, we now consider the problem of a benevolent medical planner; here, patients do not exercise choice and accept all organ offers. For each queue length n, the planner's policy specifies an organ acceptance threshold $b(n)$ such that only organs with quality exceeding $b(n)$ are offered to a transplant patient. Let $V(n)$ denote the medical planner's optimal total expected discounted reward. Bellman's equation of optimality states that

$$V(0) = \frac{\lambda}{\beta+\lambda} V(1), \tag{1}$$

$$\begin{aligned} V(n) = \sup_{b(n)\in[\underline{x},\bar{x}]} & \frac{1}{\beta+1+\lambda+\gamma n} \\ & \cdot [hn + \bar{F}(b(n))(V(n-1)+g(b(n))) \\ & + (1-\bar{F}(b(n)))V(n) + \lambda V(n+1) \\ & + \gamma n(V(n-1)+d)]. \end{aligned} \tag{2}$$

To interpret (2), notice that the total waiting payoff rate for all patients is hn when there are n patients on the waiting list. This is earned continuously until the next transition, which occurs with rate $1+\lambda+\gamma n$. Then, there are four possible transitions.

(1) With probability $\bar{F}(b(n))/(1+\lambda+\gamma n)$ an organ with quality that exceeds the threshold $b(n)$ arrives, in which case a patient receives that organ and obtains an instantaneous payoff of $g(b(n))$, and the waiting list decreases by one.

(2) With probability $[1-\bar{F}(b(n))]/(1+\lambda+\gamma n)$ an organ with quality below the decision threshold arrives, but this does not affect the state of the system because it is discarded.

(3) A new arrival occurs with probability $\lambda/(1+\lambda+\gamma n)$ and this increases the queue length by one.

(4) A patient dies with probability $\gamma n/(1+\lambda+\gamma n)$, in which case the queue length decreases and a payoff d is accrued.

Equation (1) is a boundary condition, which states that there is no waiting payoff when the system is empty and that the only possible transition is a new patient arrival occurring at a rate λ.

The optimality equation has two important corollaries. First, it can be shown that the optimal decision thresholds $b^*(n)$ are related to the optimal value function as follows (assuming that the constraint $b^*(n) \geq \underline{x}$ is not binding):

$$b^*(n) = V(n) - V(n-1). \tag{3}$$

This expression can be derived from the first-order conditions for the maximization problem in the right-hand side of (2), and it is valid because the function in the right-hand side of (2) is concave in $b(n)$—it has a nonpositive second derivative. More intuitively, this expression follows because the medical planner determines each threshold by weighing two alternatives: (1) approve a transplant and earn a payoff of at least $b(n)+V(n-1)$ or (2) deny the transplant and maintain the continuation payoff $V(n)$. Expression (3) states that the medical planner is indifferent between these two choices on the margin.

The second corollary is that the decision thresholds $b(n)$ are nonincreasing in the queue length n. That is, the medical planner is selective in the choice of organs when the queue length is small, but as the queue length increases, lower quality organs also become acceptable.

Proposition 1. *The socially optimal decision thresholds* $\{b^*(n)\}$ *are nonincreasing in* n.

The result is proven in the appendix using an argument presented in Bertsekas (1995) and extended by George and Harrison (2001).

5. Competitive Equilibrium Under FCFS

We now extend the analysis to explore the competitive equilibrium when patients retain their autonomy and the medical planner uses the FCFS priority discipline. We provide first a precise definition of the medical planner's and of each patient's strategy, introduce the appropriate equilibrium concept and finally provide an algorithm that derives the equilibrium.

Consider first the medical planner. Her strategy consists of the FCFS priority rule and a queue-length-dependent rationing rule $b = \{b(n)\colon\ n = 1, \ldots, \infty\}$. Under this rule, when the queue length is n, only organs with a value no less than $b(n)$ are assigned to patients on the transplant waiting list on a FCFS order. The rationing rule b is common knowledge to all patients. When a patient declines an organ offer, the organ is then offered to the next candidate on the waiting list.

Next, consider the patients on the waiting list. Rather than characterize the strategy of each patient, we define a strategy profile that characterizes the behavior of all patients. Because the underlying system dynamics are exponential, it is natural to focus on Markovian strategies, where each patient's decision on the range of acceptable organs depends only on the current queue length and his position on the waiting list. We also restrict attention to threshold-based acceptance policies, because each patient will accept all organs that yield a higher reward than his continuation payoff from remaining on dialysis. Hence, a strategy profile is characterized by a set of decision thresholds $\{a_k(n)\colon\ n \geq 1, n \geq k \geq 1\}$ with the following interpretation: When the queue length is n, the patient in position k will only accept organs when their quality is no less than the decision threshold $a_k(n)$; otherwise the patient will retain his position in line and pass on the organ to the next patient. For brevity of notation, we let $a = \{a_k(n)\colon\ n \geq 1, n \geq k \geq 1\}$. In addition, given the medical planner's threshold

strategy b, it follows that a strategy profile must be such that

$$a_1(n) \geq a_2(n) \geq \cdots \geq a_n(n) \geq b(n); \tag{4}$$

that is, the threshold for the kth position cannot be less than the threshold for the $k+1$st position, and all thresholds are no less than the medical planner's threshold. This condition follows from the FCFS priority rule, which implies that violations of this condition are not implementable: the decision thresholds for patients that are higher on the priority list determine the feasible thresholds for all patients behind them, and the medical planner's decision threshold places a lower bound on everyone's thresholds.

Embedded in our strategy profile is the assumption that patient strategies are symmetric: although different patients will use different acceptance thresholds at any given time because they will be at different positions along the queue, these thresholds are symmetric in the sense that all patients who are ever in position k when the queue length is n will use the same threshold $a_k(n)$. The assumption of a symmetric equilibrium is common when there are multiple equilibria (see Fudenberg and Tirole 1990, p. 160), because there is no reason for homogeneous patients to behave differently when faced with the same set of circumstances. And even though this assumption does not exclude the existence of other equilibria, it streamlines the analysis considerably.

In this setting, the patients are involved in an infinite horizon dynamic game and the relevant equilibrium concept is that of subgame perfection (Gibbons 1992). In this concept, a strategy profile is a subgame perfect Nash equilibrium if patients cannot gain by unilateral one-stage deviations from the equilibrium strategy. That is, there is no single state, defined by the pair (n, k) of the queue length and patient position, where the kth patient may gain by deviating from the actions prescribed by the strategy profile at this one state. This equilibrium concept involves two assumptions that are worth discussing: (1) each patient can only change one threshold at a time and (2) patients cannot collectively choose their threshold. The latter is a standard assumption because it is practically impossible for patients to "collude" and set their thresholds jointly. The former reflects the assumption that even if a particular patient chooses to change his whole strategy profile, these changes should satisfy Bellman's principle of optimality, and thus, it is sufficient to consider unilateral changes that involve deviations in one threshold at a time.

We will now provide an algorithm that identifies a subgame perfect strategy profile for this game. The first step is to derive expressions for each patient's continuation payoffs (or value function) for any given strategy profile a. Specifically, assuming that all patients comply with strategy profile a, let $V_k^a(n)$ denote the total expected discounted payoff for the patient in position k when the queue length is n, and let V^a denote the collection $\{V_k^a(n)\colon n \geq 1, n \geq k \geq 1\}$. The symmetric assumption is reflected in the fact that $V_k^a(n)$ is the same for all patients who will ever be in position k when the queue length is n. Then, V^a is derived as follows:

$$\begin{aligned} V_1^a(1) = {} & \frac{1}{\beta+1+\lambda+\gamma} \\ & \cdot [h+\bar{F}(a_1(1))g(a_1(1))+(1-\bar{F}(a_1(1)))V_1^a(1) \\ & \quad +\lambda V_1^a(2)+\gamma d], \end{aligned} \tag{5}$$

$$\begin{aligned} V_1^a(n) = {} & \frac{1}{\beta+1+\lambda+\gamma n} \\ & \cdot [h+\bar{F}(a_1(n))g(a_1(n))+(\bar{F}(a_n(n))-\bar{F}(a_1(n))) \\ & \quad \cdot V_1^a(n-1)+(1-\bar{F}(a_n(n)))V_1^a(n) \\ & \quad +\lambda V_1^a(n+1)+\gamma d+(n-1)\gamma V_1^a(n-1)], \end{aligned} \tag{6}$$

$$\begin{aligned} V_n^a(n) = {} & \frac{1}{\beta+1+\lambda+\gamma n} \\ & \cdot [h+\bar{F}(a_{n-1}(n))V_{n-1}^a(n-1)+\bar{F}(a_n(n))g(a_n(n)) \\ & \quad -\bar{F}(a_{n-1}(n))g(a_{n-1}(n))+(1-\bar{F}(a_n(n)))V_n^a(n) \\ & \quad +\lambda V_n^a(n+1)+\gamma d+(n-1)\gamma V_{n-1}^a(n-1)], \end{aligned} \tag{7}$$

$$\begin{aligned} V_k^a(n) = {} & \frac{1}{\beta+1+\lambda+\gamma n} \\ & \cdot [h+\bar{F}(a_{k-1}(n))V_{k-1}^a(n-1)+\bar{F}(a_k(n))g(a_k(n)) \\ & \quad -\bar{F}(a_{k-1}(n))g(a_{k-1}(n))+(\bar{F}(a_n(n))-\bar{F}(a_k(n))) \\ & \quad \cdot V_k^a(n-1)+(1-\bar{F}(a_n(n)))V_k^a(n)+\lambda V_k^a(n+1) \\ & \quad +(k-1)\gamma V_{k-1}^a(n-1)+\gamma d \\ & \quad +(n-k)\gamma V_k^a(n-1)]. \end{aligned} \tag{8}$$

For brevity of notation, it is convenient to express (5)–(8) using the shorthand notation $V^a = T^a V^a$, where the operator T^a is defined by the right-hand side of (5)–(8).

The expressions in (5)–(8) are structurally similar to the expressions in the single-party, dynamic decision-making problem (1)–(2). The differences arise because (5)–(8) capture the objective of a single patient and not the objective of the medical planner. We will now interpret these expressions by examining (8), which gives the most general case; the interpretation for Equations (5)–(7) follows a similar line of thinking.

To explore the individual components of (8) notice that $h/(\beta+1+\lambda+\gamma n)$ is the expected waiting payoff until the next transition, which occurs at a rate $1+\lambda+\gamma n$. Then, there are three possible transitions: organ arrival (scaled to rate 1), patient arrival (rate λ), and patient death (rate γn). We shall systematically consider each possible type of transition, its probability of occurrence, and its associated continuation payoff. The following are the possible cases.

(1) *Organ Arrival.* When an organ arrives, there are four possibilities from the perspective of the kth patient: (a) Its value exceeds $a_{k-1}(n)$, which implies that it will be accepted by one of the first $k-1$ patients leading into a reduction in the queue length and a reduction in the kth patient's position. (b) Its value will be between $a_{k-1}(n)$ and $a_k(n)$, implying that it will be accepted by the kth patient who will receive the reward and depart. (c) Its value will exceed $a_n(n)$ but will be less than $a_k(n)$, implying that it will be accepted by one of the patients behind the kth patient leading into a reduction in the queue length. (d) Its value will be less than the threshold $a_n(n)$, implying that the organ is discarded and the state of the system is unchanged.

(2) *Patient Arrival.* This transition occurs with probability $\lambda/(1+\lambda+\gamma n)$, it leaves the position of the kth patient unchanged, and it increases the queue length by one.

(3) *Patient Death.* This occurs with rate $n\gamma$ and there are three distinct possibilities: (a) One of the $(k-1)$ patients in front of the kth patient dies, advancing the position of the kth patient and decreasing the queue length by one (this transition occurs with probability $(k-1)\gamma/(1+\lambda+\gamma n)$. (b) The kth patient dies with probability $\gamma/(1+\lambda+\gamma n)$ and leaves the system with instantaneous payoff d. (c) One of the $(n-k)$ patients behind the kth patient dies, decreasing the queue length by one, but leaving the kth patient's position unchanged (this transition occurs with probability $(n-k)\gamma/(1+\lambda+\gamma n)$).

Having completed the derivation of the value function V^a and the operator T^a, we are in a position to formally define the equilibrium concept as follows: A strategy profile a^F is a subgame perfect Nash equilibrium if it attains the following supremum (where the supremum is taken over one component of the strategy profile at a time)

$$V^F = \sup_{\{a_k(n):\, a_1(n)\geq a_2(n)\geq\cdots\geq a_n(n)\geq b(n)\}} T^a V^F. \tag{9}$$

The fixed point V^F is the continuation payoff for the equilibrium strategy a^F; the superscript F here represents an abuse of notation and indicates that the priority rule is FCFS. The notation in (9) represents in a shorthand notation the following substitutions into (5)–(8). First, the value function V^a is now replaced by V^F. Second, the right-hand sides of (5)–(8) are maximized as follows: (5) is maximized over $a_1(1)$, (6) is maximized over $a_1(n)$, (7) is maximized over $a_n(n)$, and (8) is maximized over $a_k(n)$. The reader may verify that these right-hand expressions are quasi-concave, and thus each has a unique maximizer. For example, differentiating (5) with respect to $a_1(1)$ yields $f(a_1(1))[V_1^a(1)-a_1(1)]$, demonstrating that (5) increases with $a_1(1)$ for $a_1(1) < V_1^a(1)$ and decreases with $a_1(1)$ for $a_1(1) > V_1^a(1)$. Therefore, the theory of dynamic programming guarantees existence and uniqueness of the equilibrium value function V^F, and quasi-concavity ensures that the corresponding equilibrium strategy a^F is also unique.

In equilibrium, when the queue length is n, the patient in position k will accept organs with quality no less than $a_k^F(n)$. Further, that patient is free to change the decision threshold unilaterally, but such changes will not improve his expected payoff. He is also free to unilaterally change his policy in every subsequent state, but again, this will not provide any improvement as long as the strategies of all other patients are unchanged. An improvement can be realized only when several patients agree to change their decision thresholds simultaneously.

Having completed the characterization of the equilibrium strategy profile, we are now in a position to derive a simple expression relating the strategy profile to the continuation payoffs. Specifically, as in the socially optimal case, it is straightforward to confirm that the first-order optimality conditions are necessary and sufficient (ignoring momentarily the boundary constraints $a_k(n) \geq a_{k+1}(n) \geq b(n)$), which implies that the following conditions hold:

$$a_k^F(n) = V_k^F(n-1), \tag{10}$$

$$a_n^F(n) = V_n^F(n). \tag{11}$$

Intuitively, these conditions state that a patient is indifferent between either accepting the marginal organ $a_k^F(n)$ or passing it along to the next patient on the waiting list. If he passes it to the next patient, then that patient will accept it, and hence the queue length decreases by one leaving him with a continuation payoff $V_k^F(n-1)$. Similar interpretations are valid in the boundary cases, where the patient is either the only one waiting or the last one waiting. In these boundary cases, the patient's position and queue length remain unchanged if the marginal organ is refused.

Having characterized the equilibrium strategy profile that emerges as a response to the medical planner's organ allocation policy, we can now take a step back and return to the following question: Given our prediction for the equilibrium strategy profile, what is the medical planner's optimal treatment rationing strategy b that would maximize the planner's objective?

The following proposition, proven in the appendix, provides the answer.

Proposition 2. *It is optimal to impose no treatment rationing by choosing* $b(n) \equiv \underline{x}$ *for all* n.

The main observation employed in the proof is that the medical planner's decision thresholds $\{b(n)\}$ appear in the equilibrium characterization (9) only as a constraint on the feasible decision thresholds for the patients. Therefore, explicitly restricting the range of available organs is ineffective, because its only impact would be to discard kidneys that would not have been otherwise discarded. Interpreted more negatively, this result states that rationing is not a useful policy instrument when patients are autonomous. Rationing corresponds to a form of service rate control, because restricting the range of acceptable organs for transplant effectively modifies the "service rate" of the organ allocation system. In conventional queueing models with adjustable service rates, it is optimal to dynamically increase the service rate (at a cost) when excessive work load builds up. Even in the system studied here "service rate" controls are effective when patients are not autonomous (see the results in §4). However, these controls becomes ineffective when patients are autonomous. The implicit thresholds implemented by the patients' equilibrium strategy are as effective as any explicit thresholds that can be imposed by the medical planner.

From a computational standpoint, Proposition 2 simplifies the derivation of the equilibrium strategy profile. Specifically, when the medical planner imposes no control on the range of feasible donor organs, patients no longer have to consider the size of the waiting list in their decision problem. The only relevant information is their position on the waiting list. This is in stark contrast to the case where the medical planner imposes controls, because then queue length matters because of its effect on the range of organs available to the patients. This implies that $V_k^F(n)$ is independent of the queue length and depends only on the patient's position k, thus we let $V_k^F(n) = V^F(k)$ and $a_k^F(n) = a^F(k)$. With this simplification, it follows that the equilibrium strategy profile is derived as follows:

$$V^F(1) = \sup_{a(1)\in[\underline{x},\bar{x}]} \frac{1}{\beta+1+\gamma} \cdot [h+\bar{F}(a(1))g(a(1))+(1-\bar{F}(a(1)))V^F(1)+\gamma d], \tag{12}$$

$$V^F(k) = \sup_{a(k)\in[\underline{x},a(k-1)]} \frac{1}{\beta+1+\gamma k} \cdot [h+\bar{F}(a(k-1))V^F(k-1)+\bar{F}(a(k))g(a(k)) - \bar{F}(a(k-1))g(a(k-1))+(1-\bar{F}(a(k)))V^F(k) + \gamma d+(k-1)\gamma V^F(k-1)]. \tag{13}$$

This characterization of the equilibrium strategy profile also reveals the main shortcoming of FCFS. First, note that absent any explicit control by the planner, the system with autonomous patients achieves an implicit decision threshold such that only organs with quality greater than $a^F(n)$ are accepted when the queue length is n. However, these equilibrium

thresholds obtained by (12)–(13) will deviate from the socially optimal thresholds $b^*(n)$. A more detailed examination of this finding will be pursued in the next section.

6. Welfare Loss from Patient Autonomy

We will now compare the socially optimal decision thresholds $b^*(n)$ to the competitive equilibrium thresholds $a^F(k)$ identified in §5. It will be shown that the latter can be obtained by solving a dynamic programming recursion analogous to that solved in the socially optimal case (1)–(2), but with patient arrivals ignored. Hence, the competitive thresholds are inefficient. The argument proceeds by considering the aggregate value function $\bar{V}^F(k) = \sum_{i=1}^{k} V^F(i)$, and it shows that $\bar{V}^F(k)$ and $a^F(k)$ are obtained by solving (1)–(2), but with $\lambda = 0$. This statement is made precise in the following proposition, which is proven in the appendix.

Proposition 3. *Under FCFS, the solution* $\{a^F(k)\}$ *to the optimality Equations* (12)–(13) *also solves* (1)–(2), *but with the arrival rate* $\lambda = 0$.

This result suggests that if the social planner is faced with a *hypothetical* system with an arrival rate of zero, the optimal controls she would choose for this hypothetical system would coincide with the thresholds that arise in competitive equilibrium under FCFS. That is, the welfare loss in the competitive system arises because the self-serving behavior of current patients ignores the welfare of future patients. It can also be shown that the competitive thresholds are higher than the socially optimal ones

$$a^F(n) \geq b^*(n), \tag{14}$$

implying that the competitive outcome is inefficient because patients are too stringent in their choices, and thus generate congestion externalities. A formal proof for this result will be presented in §8, where a general class of priority rules and their corresponding equilibria will be analyzed.

In summary, our analysis has shown that treatment rationing is an ineffective way to control patient behavior, and that the system is inefficient under the commonly used FCFS rule. However, we shall show in the next section that social efficiency can be achieved if the priority rule is LCFS.

7. The Role of Prioritization

The analysis for the LCFS discipline follows the steps developed in §5 with the exception that each new patient now arrives at the top of the line, shifting back the position of all other patients by one. Using the approach of §5, we can show that the medical planner will never restrict the range of organs offered to the patients, and hence, the value function for a patient in position k does not depend on the queue length. If we now let $V^L(k)$ denote the value function for a patient in position k and $a^L(k)$ denote that patient's equilibrium acceptance threshold, then the dynamic programming recursion for the competitive equilibrium is as follows:

$$\begin{aligned} V^L(1) = \sup_{a(1)\in[\underline{x},\bar{x}]} & \frac{1}{\beta+1+\lambda+\gamma} \\ & \cdot[h+\bar{F}(a(1))g(a(1))+(1-\bar{F}(a(1)))V^L(1) \\ & +\lambda V^L(2)+\gamma d], \end{aligned} \tag{15}$$

$$\begin{aligned} V^L(k) = \sup_{a(k)\in[\underline{x},a(k-1)]} & \frac{1}{\beta+1+\lambda+\gamma k} \\ & \cdot[h+\bar{F}(a(k-1))V^L(k-1)+\bar{F}(a(k))g(a(k)) \\ & -\bar{F}(a(k-1))g(a(k-1))+(1-\bar{F}(a(k)))V^L(k) \\ & +\lambda V^L(k+1)+\gamma d+(k-1)\gamma V^L(k-1)]. \end{aligned} \tag{16}$$

Unlike the FCFS system, in the LCFS system, arrivals are taken into account because they cause each patient's position to increase by one. One can proceed by considering the aggregate value functions and comparing the competitive acceptance thresholds under LCFS to the socially optimal thresholds. The following result shows that the aggregate value function is obtained by solving the dynamic programming recursion (1)–(2), and hence, with the LCFS priority rule, the medical waiting system with autonomous patients is socially efficient.

Proposition 4. *Under LCFS, the optimal strategy for the medical planner is to exercise no treatment rationing. The competitive equilibrium that emerges,* $\{a^L(k)\}$, *is socially optimal.*

We shall defer the proof of this proposition until a more general version of this result is presented in §8.

This result establishes that with LCFS the externalities caused by patient choice are completely internalized, because a patient who refuses an offer will

drop on the waiting list when a new patient arrives. The threat of this reduction in position is sufficient to align the patient's behavior with the behavior desired by the medical planner, and thus the socially optimal ideal can be attained.

Most of the existing literature on queueing models with hidden information relies on monetary payments to distinguish between customers of different (unobserved) types. The observation that prioritization alone is sufficient to coordinate such queueing systems was first made informally by Hassin (1985), and to the best of our knowledge, Proposition 4 is the first case in which this insight is rigorously proved. Further, previous work has focussed mainly on static equilibrium analysis in which individual decisions are made based on long-run average quantities (such as expected delay). Because these quantities are independent of the queueing discipline by virtue of Little's Law, static models do not fully capture the power of priorities (within a single customer class). In this regard, our dynamic analysis provides the additional insight that an appropriately chosen queueing discipline can indeed completely eradicate incentive problems among identical customers, as long as real-time system information is readily available and intelligently utilized.

Our analysis of the two polar extremes of LCFS and FCFS allows us to conclude that the former is efficient, whereas the latter is not. However, the social efficiency of LCFS should be treated with caution because of strategic difficulties associated with its implementation: without any form of monitoring, any person in line has the motivation to balk and reenter the system at the top of the line (see Hassin 1985). Although we shall assume that this is not permitted, we also acknowledge the massive administrative costs involved with preventing such behavior, which could partly explain why LCFS systems are rarely observed in practice.

On the other hand, the result that FCFS is inefficient appears incompatible with the observation that it is commonly used in practice. In fact, apart from minor provisions made for exceptional cases, it is the primary prioritization scheme being used. While it is recognized in Larson (1987) that FCFS prioritization, being a symbol of justice and equity, enjoys many advantages that go beyond economic welfare, our model highlights the inherent inefficiency caused by the inability of FCFS to contain the externalities generated by patients' self-serving behavior.

8. The Equity-Efficiency Trade-off

The result that LCFS achieves a socially efficient competitive equilibrium, while FCFS suffers deadweight losses caused by externality problems brings to the forefront the trade-off between efficiency and equity: The priority rule that maximizes system efficiency is the one that deviates the most from FCFS, the acceptable standard for equity. While this tradeoff has attracted considerable attention, previous studies have focused on patient heterogeneity as the main driver behind it. Specifically, while an efficient policy would allocate organs to patients most likely to benefit from transplantation, this would be unjust because it would create disparities in access to transplantation between different ethnic and racial groups (see Zenios et al. 2000). By contrast, the results developed in this paper demonstrate that a trade-off between efficiency and equity can exist even when patients are homogeneous, because of patient choice and its interaction with the queueing priority rule.

To quantify this trade-off, we now consider a continuum of priority rules called randomized absolute priority rules. In these rules, new patients will either receive absolute priority and skip to the head of the line or join the end of the line. Each new patient is granted absolute priority independently with probability $p \in [0, 1]$; we let $\{a^p(k)\}$ denote the equilibrium decision thresholds under this priority system. The case of $p = 0$ corresponds to FCFS and $p = 1$ corresponds to LCFS. These prioritization schemes are probabilistic hybrids of the FCFS and LCFS rules and are thus natural candidates for analysis.

The reader could almost predict our next result, which states that under the randomized absolute priority rule with parameter p, the arrival rate when patients exercise choice is effectively λp. The proof, presented in the appendix, relies on the aggregation argument described in §6.

Proposition 5. *Under randomized absolute priority with parameter p, the competitive equilibrium thresholds $\{a^p(k)\}$ can be obtained by solving the optimality Equations* (1)–(2) *with the arrival rate set at λp.*

This proposition presents a continuum of cases covering both FCFS and LCFS prioritization, and confirms the antithetical relationship between absolute priority and externalities. Absolute priority embodies nuances of social injustice, while negative externalities, manifested through distortions in offer acceptance rates, lead to economic inefficiency. The relationship between the absolute priority parameter p and the effective arrival rate λp can thus be interpreted as a quantitative representation of the equity-efficiency trade-off, because absolute priorities provide a mechanism to minimize the impact of the negative externalities.

The effect of the priority parameter p on controlling the externalities can also be confirmed in the following result, which states the impact of the parameter on the equilibrium thresholds.

PROPOSITION 6. *Let* $0 \le p \le p' \le 1$. *Then, for each queue length* n,

$$a^p(n) \ge a^{p'}(n). \tag{17}$$

This proposition states that patients become less selective as the priority parameter increases and, consequently, as the threat of a reduction in their priority following an organ refusal becomes more severe. Therefore, the social planner is able to mitigate the externality effects with randomized absolute priority rules. Such regimes can vary in intensity according to the parameter p, and more importantly, can be justified in several practical cases. (For example, emergency cases and pediatric patients may arguably deserve absolute priority.) On a broader level, our interpretations also suggest that these externality problems can be kept under control using a more general class of preemptive regimes that go beyond randomized absolute priority rules—regimes in which patients who decline an organ offer would expect a decrease in their priority position.

9. Numerical Study

In this section, we present results from a numerical study that has two main objectives. The first one is to illustrate that patient choice can significantly degrade the performance of the transplant waiting list as measured by average waiting time and expected patient reward, and to identify key system parameters that either exacerbate or alleviate such performance degradation. The second objective is to study the effect of absolute priorities on system performance. In particular, it will be demonstrated that a system where a small fraction of the patients are granted absolute priorities, while the rests are prioritized according to FCFS, can recover most of the losses caused by the FCFS system.

The parameters for the study are estimated from kidney transplant data obtained from UNOS (2002). The arrival rate is set at $\lambda = 200$ patients per year and the service rate is set at $\mu = 100$ organs per year. These were obtained by scaling down (and rounding to the closest hundred) the corresponding figures for the national waiting list by a factor of 100, so they represent the waiting list maintained by a small OPO (there are 59 OPOs in the United States). Patient death rate is $\gamma = 0.124$, calculated from mortality statistics of transplant candidates as reported in USRDS (2002) and in Wolfe et al. (1999). The following reward structure is used as a basis for decision making. Following Zenios et al. (2000), a reward of $h = 0.6$ per year is assigned to patients on the waiting list. This implies that the relative quality of life on the waiting list is 60% of that for healthy individuals. The corresponding reward per unit time for patients who have received a transplant is taken to be 0.75. Patient reward after death is $d = 0$, measured in QALYs.

We next proceed to estimate the payoffs obtained from receiving a transplant. Using a USRDS data set of 37,756 kidney transplants performed between November 1994 and December 1999, we fit an exponential survival model with covariates that include, for example, patient and donor tissue types, age, sex, and race. (See Venables and Ripley (1997) for an overview of the relevant survival analysis techniques.) For the average patient, we use this model to estimate survival under two extreme scenarios (best-possible kidney and worst-possible kidney) and then obtain the quality-adjusted life expectancy assuming a quality of life with transplant of 0.75 as above. Then, the best-case estimate is 8.98 years and the worst-case estimate is 4.05 years. Given these estimates, we assume that the random variable X representing the reward from a random organ offer is uniformly distributed between $\underline{x} = 4.00$ and $\bar{x} = 9.00$. All rewards are discounted at a continuous rate $\beta =$ 3%. Notice that with this reward structure, a patient's expected discounted QALYs from waiting indefinitely

is $h/(\beta+\gamma)=3.89$, hence even the worst kidney offer is attractive.

First, we address the effect of patient choice by comparing the performance of the waiting system under both FCFS and LCFS. FCFS is a proxy for the allocation system that now prevails in the United States. Our analysis allows us to compute system transition rates, stationary queue-length distributions, and value functions for both systems, and these quantities can then be used to compute the average waiting time, average queue length, average life expectancy, and welfare loss defined as the percentage difference in QALY between FCFS and the socially optimal LCFS. Table 1 summarizes the key performance metrics. The results demonstrate that patient choice in the FCFS system increases the average queue length and average waiting time until treatment relative to LCFS by 17.8% and 20.4%, respectively. This is because a fraction of the organs allocated under LCFS are discarded under FCFS because they are refused by the patients on the waiting list. This organ wastage culminates into a 5.9% decrease in each patient's quality-adjusted life expectancy. This seemingly small welfare loss translates into an annual loss of more than 7,000 patient life years (because more than 23,328 patients join the waiting list each year and the life expectancy of each patient is reduced by 0.3 years). The economic consequences are also significant: Because each of the approximately 9,636 patients receiving cadaveric transplants each year has to spend an additional 1.1 years on dialysis with FCFS and because the cost of treatment on the waiting list is \$45,000 per patient per year, this translates into an approximate increase in health care costs of more than \$476 million per year. These figures are calculated based on waiting list statistics for 2002 reported in OPTN (2003) and dialysis costs reported in Pastan and Bailey (1998), which are all expected to increase in the future.

Table 1 Welfare Loss in Terms of Performance Metrics

	Mean queue length	Waiting time until transplant (years)	Discarded kidneys (%)	Expected survival (QALY)
First-best LCFS	806	5.48	0	5.20
FCFS	950	6.60	15.8	4.89
% change under FCFS	+17.8	+20.4	—	−5.90

Next, we examine the impact of different system parameters on welfare losses. We obtain the quality-adjusted life expectancy under four additional scenarios, representing: (1) a bigger OPO, (2) increased variability in organ quality, (3) improved overall organ quality, and (4) increased organ supply. The results are summarized in Table 2.

The following four observations can be made.

(1) A 10-fold increase in arrival and service rates is associated with an increase in welfare loss from 5.90% to 7.98%. Further, increased scale is associated with improved overall performance in LCFS (QALYs increase from 5.20 to 5.32), but not in FCFS. This suggests that the statistical economies of scale that are frequently present in standard queueing systems and that explain the improved performance of the LCFS system in this scenario, also cause an increase in welfare losses under the FCFS system. Increased scale with FCFS makes organ refusals more appealing for the patients on the top of the waiting list. The increased arrival rate elevates these patients' likelihood of receiving a better future offer and drives their decision to refuse organs of low quality.

(2) In the scenario representing increased variability, the organ quality distribution is changed to $U[4,10]$; an alternative is to use $U[3,10]$, which has the same mean as the baseline case and higher variance, but this creates boundary problems when the value of the organ is less than 3.89—the value of waiting in perpetuity. We will compare this case to the

Table 2 Dependence of Welfare Loss on System Parameters

			LCFS	FCFS	Welfare loss (%)
Baseline scenario:	$\lambda=200, \mu=100$	$X\sim U[4,9]$	5.20	4.89	5.90
Increased scale:	$\lambda=2{,}000, \mu=1{,}000$	$X\sim U[4,9]$	5.32	4.89	7.98
Increased variability:	$\lambda=200, \mu=100$	$X\sim U[4,10]$	5.45	5.08	6.79
Increased mean:	$\lambda=200, \mu=100$	$X\sim U[4.5,9.5]$	5.45	5.11	6.28
Increased supply:	$\lambda=200, \mu=125$	$X\sim U[4,9]$	5.52	5.16	6.57

scenario representing increased mean quality (with organ quality distribution of $U[4.5, 9.5]$), because the mean organ quality is the same for both scenarios. We see that the welfare loss is slightly higher in an environment of higher variability (6.79% compared to 6.28%), because the patients' option to wait becomes more valuable.

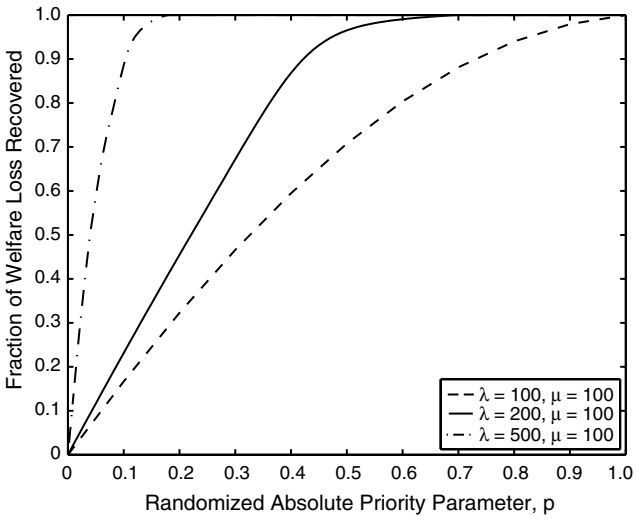
(3) An improvement in overall organ quality by increasing post-transplant QALYs uniformly by 0.5 years improves quality-adjusted life expectancy in FCFS from 4.89 years to 5.11 years, but it also increases the total welfare loss. Further, the increased QALY under FCFS (5.11 years) falls short of the socially optimal performance of the baseline system (5.20 years). This implies that remedying the externality problems in the original system would have a larger welfare impact compared to an improvement in average organ quality.

(4) In the same vein, increasing the organ supply by a practically impossible 25% leads to an increase in quality-adjusted life expectancy from 4.89 years to 5.16 years in the FCFS system. This improvement has the same order of magnitude as the welfare gain from LCFS in the baseline system. That is, the welfare gain caused by adopting a system that aligns patient choice with social efficiency is equivalent to a welfare gain that could be achieved by a 25% increase in organ supply.

In summary, these results suggest that the kidney transplantation system suffers from substantial welfare losses caused by patient choice. While most efforts to improve the performance focus on supply-side strategies, our analysis suggests that the effectiveness of supply-side interventions pales in comparison to that of the demand-based priority rules proposed here.

How much absolute priority should be granted? While the previous analysis suggests that the welfare gains obtained by switching from FCFS to LCFS are substantial, there are several hurdles that would prevent the adoption of LCFS. The analysis in §8 suggests that randomized absolute priority systems can recover part of the welfare losses, because patients internalize some of the externalities caused by their decision to decline an organ offer. We will now investigate the relationship between the fraction of patients receiving absolute priorities, and the fraction of welfare loss that can be recovered.

Figure 1 System Performance Under Various Degrees of Absolute Priority



We start with the baseline scenario and investigate how the performance of the system (measured by the quality-adjusted life expectancy) changes as the fraction p of patients who receive absolute priority also changes. The solid curve in Figure 1 summarizes our results: the parameter p is reflected on the horizontal axis and the fraction of welfare losses recovered is plotted on the vertical axis. As expected, the fraction of welfare loss recovered increases from 0 to 1 as we increase p from 0 (FCFS) to 1 (LCFS). Further, this curve is concave, indicating that the first few units of absolute priority are more effective than subsequent units. We note that 90% of the welfare loss is recovered with $p = 0.41$, and the socially efficient outcome is attained with $p = 0.69$.

Next, we repeat the experiment and vary the supply-to-demand ratio. We consider the following two cases ($\lambda = 100$, $\mu = 100$ and $\lambda = 500$, $\mu = 100$), and represent the results using dotted lines in Figure 1. When demand is five times as much as supply, 90% of the welfare loss is already recovered with $p = 0.10$, and the socially efficient outcome is attained with $p = 0.17$. When demand and supply are balanced, a random absolute priority parameter of $p = 0.72$ is required to recover 90% of the welfare loss, and the socially efficient outcome is attained only when $p = 0.98$. This comparison demonstrates that as the supply-demand imbalance increases, the initial units of absolute priority become more effective. Therefore, the limited

use of randomized priority can improve social welfare in overloaded organ allocation systems, which are expected to become more common in the future.

10. Concluding Remarks

Despite the continued shortage of organs for transplantation, approximately 12% of all kidneys procured from deceased donors are routinely discarded. This wastage contributes to the already long waiting times and to the high cost of treatment for patients on the waiting list. It is also widely recognized that patient choice is a main driver behind this wastage. In fact, a recent report by UNOS (2002) states that "a subset of deceased organs of marginal quality has a high discard rate after procurement because their placement is arduous and prolonged." This observation has motivated the development of an Expanded Criteria Donor policy that aims to expedite the placement of marginal organs by exploiting patient choice.

This paper attempts to clarify the effect of patient choice on the organ allocation system by examining a stylized queueing model. It demonstrates that the priority discipline used to rank patients on the transplant waiting list can exacerbate the wastage of organs. In fact, FCFS, an established standard of equity, increases patients' desire to refuse organs of marginal quality and aggravates the organ discard rate. By contrast, LCFS eliminates all externality problems and achieves optimal organ utilization. This finding highlights a new dimension in the equity-efficiency dilemma. The most equitable FCFS policy is inefficient because it exaggerates the externalities of patient choice. This efficiency loss is over and above the well-studied inefficiencies caused by FCFS's failure to allocate organs to the patients more likely to benefit from them (see Zenios et al. (2000) and references therein).

However, our findings are still several steps away from providing immediately applicable practical recommendations beyond the general warning that "FCFS exacerbates the organ wastage caused by patient choice." Specifically, LCFS is practically infeasible, and the absence of patient heterogeneity in our model may create suspicion about the validity of our findings. A companion paper (Su and Zenios 2004a) explores the role of heterogeneity in more detail and shows that the inefficiency of the FCFS rule remains relevant in that context. In another work (Su and Zenios 2004b), we study a variant of the "multiserver FCFS" queueing discipline: kidneys are classified into quality grades, and there is a separate FCFS queue for each grade of kidneys. The grades are designed so that better kidneys involve longer waits, and patients balance the trade-off between the quality of the kidney and the waiting time by declaring which grade of kidney they wish to receive. Theoretical analysis demonstrates that this system mitigates the inefficiencies of FCFS, and a simulation model that captures the realistic complexities of the kidney allocation environment (see Su et al. 2004), demonstrates that the same system accelerates the placement of kidneys and reduces organ wastage.

One of the objectives of this paper is to determine the impact of the priority discipline on the number of organs discarded because of patient choice. The numerical results demonstrate that under a FCFS discipline, as many as 15% of organs may be wasted because of patient refusals. This estimate compares favorably to the actual percentage of 12% reported in clinical studies, and supports the hypothesis that a significant number of organs are refused because of the prioritization system. However, the actual number of organs discarded cannot be estimated precisely by our model. This is because the local waiting systems that make up the national waiting list share organs, and thus the U.S. waiting system looks like a massive single waiting list—much different that the smaller waiting systems analyzed in §9. One may then argue that organ refusals will be rare in such a massive system, because the effects of the priority discipline will be dwarfed by the scale of the waiting system. While compelling, this argument is flawed: In a national waiting system, refusals cause an organ to be transported across regions, incurring transportation delays that cause a deterioration in the quality of the organ. So even if the organ is not discarded following refusals, it will be substantially inferior. However even if one ignores the transportation delays, the numerical results in §9 suggest that the welfare losses because of organ refusals increase with the scale of the system, and thus, they may be more profound in a massive national system compared to a smaller local system. Nevertheless, we do not wish to argue that our findings make a convincing case that much of the currently refused organs are because of

the priority discipline. While our informal discussions with clinicians in two transplant centers suggest that patients frequently refuse organs because they recognize that once they reach the top of the waiting list they can afford to be selective, one of the reviewers shared with us a different conversation with a transplant team that contradicted our informal discussions. Absent any primary unbiased data, the results in this paper support the hypothesis postulated above, but do not convincingly prove it.

In summary, while it is desirable to expand the role of patient choice in the organ allocation system, poorly designed initiatives can have the opposite of the intended effect. This paper highlights these "unintended consequences" and identifies containment mechanisms that help improve the utilization of scarce organ resources.

Acknowledgments
The authors have benefited from stimulating discussions with Haim Mendelson and Korok Ray. Partial financial support was provided by the National Science Foundation Grant SBER-9982446. This paper has benefited substantially from the comments and suggestions of the Senior Editor and three anonymous referees.

Appendix. Proofs

Proof of Proposition 1. The proof shall proceed in three steps.

Step 1. In this step, we use value iteration to establish that the relative value function defined as $\Delta(n) \equiv V(n) - V(n-1)$ is nonincreasing in n. Let $V^k(n)$ denote the kth iterate for the value function, with the supremums attained by $\{b^k(n)\}$; the superscript here represents an abuse of notation because in the main part of this paper, superscripts in the value function are used to represent different priority rules. That is, starting arbitrarily with $V^0(n) \equiv 0$, we have, for every $n \geq 0$

$$\begin{aligned} V^{k+1}(0) &= \frac{\lambda}{\beta+\lambda} V^k(1), \\ V^{k+1}(n) &= \sup_{b(n)\in[\underline{x},\bar{x}]} \frac{1}{\beta+1+\lambda+\gamma n} \\ &\quad \cdot [hn + \bar{F}(b(n))(V^k(n-1)+g(b(n))) \\ &\quad + (1-\bar{F}(b(n)))V^k(n) + \lambda V^k(n+1) \\ &\quad + \gamma n(V^k(n-1)+d)]. \end{aligned} \tag{18}$$

Next, define for $k \geq 0$ and $n \geq 1$,

$$\Delta^k(n) \equiv V^k(n) - V^k(n-1),$$

which, by the convergence of the value iteration algorithm, converges to $\Delta(n)$ as $k \to \infty$ for every $n \geq 1$.

Now, to establish that $\{\Delta(n)\}$ is nonincreasing in n, it suffices to show that $\{\Delta^k(n)\}$ is nonincreasing in n for every $k \geq 0$. We shall show this by induction. Notice that this holds trivially for $k = 0$. For $k \geq 0$ and $n > 1$, elementary algebra shows that

$$\begin{aligned} \Delta^{k+1}(n+1) &= V^{k+1}(n+1) - V^{k+1}(n) \\ &\leq \frac{1}{\beta+1+\lambda+(n+1)\gamma} \\ &\quad \cdot [h + \bar{F}(b^k(n+1))\Delta^k(n) \\ &\quad + (1-\bar{F}(b^k(n+1)))\Delta^k(n+1) \\ &\quad + \lambda\Delta^k(n+2) + \gamma n\Delta^k(n) + \gamma d], \end{aligned} \tag{19}$$

$$\begin{aligned} \Delta^{k+1}(n) &= V^{k+1}(n) - V^{k+1}(n-1) \\ &\geq \frac{1}{\beta+1+\lambda+(n+1)\gamma} \\ &\quad \cdot [h + \bar{F}(b^k(n-1))\Delta^k(n-1) \\ &\quad + (1-\bar{F}(b^k(n-1)))\Delta^k(n) \\ &\quad + \lambda\Delta^k(n+1) + \gamma(n-1)\Delta^k(n-1) \\ &\quad + \gamma\Delta^k(n) + \gamma d]. \end{aligned} \tag{20}$$

To obtain the inequalities in (19)–(20), we use the fact that $b^k(n+1)$ and $b^k(n-1)$ are suboptimal in state n, and we uniformize the transition rates. Inequalities (19)–(20), together with the inductive hypothesis, imply

$$\begin{aligned} \Delta^{k+1}(n+1) - \Delta^{k+1}(n) &\leq \frac{1}{\beta+1+\lambda+(n+1)\gamma} \\ &\quad \cdot [(1-\bar{F}(b^k(n+1)))(\Delta^k(n+1)-\Delta^k(n)) \\ &\quad + \bar{F}(b^k(n-1))(\Delta^k(n)-\Delta^k(n-1)) \\ &\quad + \lambda(\Delta^k(n+2)-\Delta^k(n+1)) \\ &\quad + \gamma(n-1)(\Delta^k(n)-\Delta^k(n-1))] \\ &\leq 0. \end{aligned} \tag{21}$$

Similarly, for the boundary terms, we have, for $k \geq 0$,

$$\begin{aligned} \Delta^{k+1}(2) &\leq \frac{1}{\beta+1+\lambda+2\gamma} \\ &\quad \cdot [h + \bar{F}(b^k(2))\Delta^k(1) + (1-\bar{F}(b^k(2)))\Delta^k(2) \\ &\quad + \lambda\Delta^k(3) + \gamma\Delta^k(1) + \gamma d], \end{aligned} \tag{22}$$

$$\Delta^{k+1}(1) \geq \frac{1}{\beta+1+\lambda+2\gamma}[h + \Delta^k(1) + \lambda\Delta^k(2) + \gamma\Delta^k(1) + \gamma d], \tag{23}$$

where (22) follows from taking $n = 1$ in (19), and transition rates are appropriately uniformized. Inequalities (22)–(23),

together with the inductive hypothesis, imply

$$\begin{aligned}\Delta^{k+1}(2)-\Delta^{k+1}(1) &\le \frac{1}{\beta+1+\lambda+2\gamma}\\ &\quad\cdot[(1-F(b^k(2)))(\Delta^k(2)-\Delta^k(1))\\ &\quad+\lambda(\Delta^k(3)-\Delta^k(2))]\\ &\le 0. \end{aligned} \tag{24}$$

Combining (21) and (24), our inductive proof is complete, and we have shown that $\Delta(n)$ is nonincreasing in n.

Step 2. Next, after removing terms that do not depend on $b(n)$, observe that for each $n \ge 1$, $b^*(n)$ is the maximizer of:

$$V(b(n),\Delta(n)) \equiv \bar{F}(b(n))(g(b(n))-\Delta(n)). \tag{25}$$

This function satisfies the following increasing differences (ID) property:

$$V(b',\Delta)-V(b,\Delta) \quad \text{is strictly increasing in } \Delta\in\mathbb{R} \text{ for any } b,b'\in(\underline{x},\bar{x}) \text{ s.t. } b<b'.$$

We can verify this property by observing the relation $\partial^2 V(b,\Delta)/\partial b\partial\Delta = f(b)$ and integrating from b to b'.

Step 3. Finally, consider some arbitrary n, n' such that $n < n'$. The result from Step 1 allows us to conclude that $\Delta(n) \ge \Delta(n')$. We consider the following two cases.

Case 1. $\Delta(n) = \Delta(n')$

Here, $b^*(n)$ and $b^*(n')$ are maximizing the same objective function (25). If the solution is unique, then $b^*(n) = b^*(n')$; otherwise, we can choose the same solution for both and we still have $b^*(n) = b^*(n')$.

Case 2. $\Delta(n) > \Delta(n')$

Suppose for sake of contradiction that $b^*(n) < b^*(n')$. The ID property gives

$$\begin{aligned}&V(b^*(n'),\Delta(n))-V(b^*(n),\Delta(n))\\ &\quad> V(b^*(n'),\Delta(n'))-V(b^*(n),\Delta(n')),\end{aligned}$$

but by the optimality of $b^*(n)$ and $b^*(n')$ in (25), we know that

$$\begin{aligned}&V(b^*(n'),\Delta(n))-V(b^*(n),\Delta(n))\\ &\quad\le 0 \le V(b^*(n'),\Delta(n'))-V(b^*(n),\Delta(n')),\end{aligned}$$

which gives the desired contradiction. Therefore, $b^*(n) \ge b^*(n')$.

Therefore, we can always find an optimal policy $\{b^*(n)\}$ that is nonincreasing in n. The proof is complete. □

Proof of Proposition 2. Consider an arbitrary control policy $\{b(n)\}$, let $\{V_k(n)\}$ and $\{a_k(n)\}$ denote the equilibrium value function and patient decision thresholds. We shall begin by showing that $\{b(n)\}$ cannot be optimal if there is some n for which

$$a_n(n) > V_n(n) \quad \text{and} \quad a_n(n) \ne \underline{x}. \tag{26}$$

The proof proceeds by contradiction. Consider the value of n for which (26) holds. In this case, we must have $a_n(n) = b(n)$; otherwise, the equilibrium cannot be sustained because a smaller $a_n(n)$ is feasible and will be chosen instead since $a_n(n)$ denotes the lowest acceptable quality, and $V_n(n)$ is the continuation payoff from waiting. This similarly implies that if $b(n)$ is reduced to $b(n)-\epsilon$, the equilibrium value of $a_n(n)$ decrease and $V_n(n)$ will increase. Hence, the policy $\{b(n)\}$ is therefore suboptimal.

This establishes that if $\{b(n)\}$ is optimal, then $a_n(n) \le V_n(n)$ or $a_n(n) = \underline{x}$ for every n. Since $a_n(n) < V_n(n)$ is not sustainable in equilibrium, we are left to consider policies with either $a_n(n) = V_n(n)$ or $a_n(n) = \underline{x}$. For any such policy, feasibility implies that $b(n) = \underline{x}$ for any n with $a_n(n) = \underline{x}$. For all other values of n, decreasing the control $b(n)$ to $\underline{x}$ does not affect the equilibrium decision thresholds, because $a_n(n)$ already satisfies the first-order condition in (11).

This implies that all these policies yield the same equilibrium outcome as the no-control policy $b(n) \equiv \underline{x}$, which is therefore optimal. □

Proof of Proposition 3. This follows as a corollary to Proposition 5 by using $p = 0$. □

Proof of Proposition 4. This follows as a corollary to Proposition 5 by using $p = 1$. □

Proof of Proposition 5. We shall begin by writing down the optimality equations under randomized absolute priority with parameter p. Let $a^p(k)$ denote the decision threshold that attains the supremum and $V^p(k)$ the optimal value function for a patient in position k, we can express the optimality equations as

$$\begin{aligned}V^p(1) &= \frac{1}{\beta+1+\lambda+\gamma}\\ &\quad\cdot[h+\bar{F}(a^p(1))g(a^p(1))+(1-\bar{F}(a^p(1)))V^p(1)\\ &\quad+\lambda(pV^p(2)+(1-p)V^p(1))+\gamma d],\end{aligned} \tag{27}$$

$$\begin{aligned}V^p(k) &= \frac{1}{\beta+1+\lambda+\gamma k}\\ &\quad\cdot[h+\bar{F}(a^p(k-1))V^p(k-1)+\bar{F}(a^p(k))g(a^p(k))\\ &\quad-\bar{F}(a^p(k-1))g(a^p(k-1))+(1-\bar{F}(a^p(k)))V^p(k)\\ &\quad+\lambda(pV^p(k+1)+(1-p)V^p(k))\\ &\quad+\gamma d+(k-1)\gamma V^p(k-1)].\end{aligned} \tag{28}$$

Now, let $\bar{V}^p(k) = \sum_{i=1}^k V^p(i)$. Then, retaining the supremum only for the last term in each summation and substituting $a^p(k)$ into the other terms, we have

$$\begin{aligned}\bar{V}^p(1) &= \sup_{a(1)\in[\underline{x},\bar{x}]}\frac{1}{\beta+1+\lambda+\gamma}\\ &\quad\cdot[h+\bar{F}(a(1))g(a(1))+(1-\bar{F}(a(1)))\bar{V}^p(1)\\ &\quad+\lambda(p(\bar{V}^p(2)-V^p(1))+(1-p)\bar{V}^p(1))+\gamma d],\end{aligned} \tag{29}$$

$$\begin{aligned}\bar{V}^p(k) = {} & \frac{1}{\beta+1+\lambda+\gamma} \\ & \cdot[h+\bar{F}(a^p(1))g(a^p(1))+(1-\bar{F}(a^p(1)))V^p(1) \\ & +\lambda(pV^p(2)+(1-p)V^p(1))+\gamma d], \\ & +\sum_{i=2}^{k-1}\left\{\frac{1}{\beta+1+\lambda+i\gamma}[h+\bar{F}(a^p(i-1))V^p(i-1)\right. \\ & \quad +\bar{F}(a^p(i))g(a^p(i))-\bar{F}(a^p(i-1))g(a^p(i-1)) \\ & \quad +(1-\bar{F}(a^p(i)))V^p(i)+\lambda(pV^p(i+1) \\ & \quad \left.+(1-p)V^p(i))+\gamma d+(i-1)\gamma V^p(i-1)]\right\} \\ & + \sup_{a(k)\in[\underline{x},a^p(k-1)]}\left\{\frac{1}{\beta+1+\lambda+k\gamma}[h+\bar{F}(a^p(k-1))\right. \\ & \quad \cdot V^p(k-1)+\bar{F}(a(k))g(a(k)) \\ & \quad -\bar{F}(a^p(k-1))g(a^p(k-1)) \\ & \quad +(1-\bar{F}(a(k)))V^p(k)+\lambda(pV^p(k+1) \\ & \quad +(1-p)V^p(k))+\gamma d \\ & \quad \left.+(k-1)\gamma V^p(k-1)]\right\} \qquad (30) \\ = {} & \frac{1}{\beta+1+\lambda+k\gamma} \\ & \cdot[h+\bar{F}(a^p(1))g(a^p(1))+(1-\bar{F}(a^p(1)))V^p(1) \\ & +\lambda(pV^p(2)+(1-p)V^p(1))+\gamma d+(k-1)\gamma V^p(1)] \\ & +\sum_{i=2}^{k-1}\left\{\frac{1}{\beta+1+\lambda+k\gamma}[h+\bar{F}(a^p(i-1))V^p(i-1)\right. \\ & \quad +\bar{F}(a^p(i))g(a^p(i))-\bar{F}(a^p(i-1))g(a^p(i-1)) \\ & \quad +(1-\bar{F}(a^p(i)))V^p(i)+\lambda(pV^p(i+1) \\ & \quad +(1-p)V^p(i))+\gamma d+(i-1)\gamma V^p(i-1) \\ & \quad \left.+(k-i)\gamma V^p(i)]\right\} \\ & + \sup_{a(k)\in[\underline{x},a^p(k-1)]}\left\{\frac{1}{\beta+1+\lambda+k\gamma}[h+\bar{F}(a^p(k-1))\right. \\ & \quad \cdot V^p(k-1)+\bar{F}(a(k))g(a(k)) \\ & \quad -\bar{F}(a^p(k-1))g(a^p(k-1)) \\ & \quad +(1-\bar{F}(a(k)))V^p(k)+\lambda(pV^p(k+1) \\ & \quad +(1-p)V^p(k))+\gamma d \\ & \quad \left.+(k-1)\gamma V^p(k-1)]\right\} \\ = {} & \sup_{a(k)\in[\underline{x},a^p(k-1)]}\frac{1}{\beta+1+\lambda+\gamma k} \\ & \cdot[hq+\bar{F}(a(k))g(a(k))+\bar{V}^p(k-1)+(1-\bar{F}(a(k)))V^p(k) \\ & +\lambda(p(\bar{V}^p(k+1)-V^p(1))+(1-p)\bar{V}^p(k))+\gamma kd \\ & +\sum_{i=1}^{q}(k-i)\gamma V^p(i)+\sum_{i=1}^{q}(i-1)\gamma V^p(i-1)] \qquad (31) \\ = {} & \sup_{a(k)\in[\underline{x},a(k-1)]}\frac{1}{\beta+1+\lambda+\gamma k} \\ & \cdot[hk+\bar{F}(a(k))g(a(k))+\bar{F}(a(k))\bar{V}^p(k-1) \\ & +(1-\bar{F}(a(k)))\bar{V}^p(k)+\lambda(p(\bar{V}^p(k+1)-V^p(1)) \\ & +(1-p)\bar{V}^p(k))+\gamma kd+\gamma k\bar{V}^p(k-1)]. \qquad (32)\end{aligned}$$

Next, let $J(k)=z+\bar{V}^p(k)$, where $z=(\lambda p/\beta)V^p(1)$. Then, from the expression for $\bar{V}^p(1)$ in (29), and after some algebra we have

$$\begin{aligned}J(1) = {} & \sup_{a(1)\in[\underline{x},\bar{x}]}\frac{1}{\beta+1+\lambda+\gamma} \\ & \cdot\left[h+\bar{F}(a(1))\left(g(a(1))+\frac{\lambda p}{\beta+\lambda p}J(1)\right)+(1-\bar{F}(a(1)))J(1)\right. \\ & \quad \left.+\lambda(pJ(2)+(1-p)J(1))+\gamma\left(d+\frac{\lambda p}{\beta+\lambda p}J(1)\right)\right]. \qquad (33)\end{aligned}$$

Similarly, using the expression for $\bar{V}^p(k)$ in (32), we have

$$\begin{aligned}J(k) = {} & \sup_{a(k)\in[\underline{x},a(k-1)]}\frac{1}{\beta+1+\lambda+\gamma k} \\ & \cdot[hk+\bar{F}(a(k))g(a(k))+\bar{F}(a(k))J(k-1) \\ & +(1-\bar{F}(a(k)))J(k)+\lambda[pJ(k+1)+(1-p)J(k)] \\ & +\gamma kd+\gamma kJ(k-1)]. \qquad (34)\end{aligned}$$

After removing "dummy" transitions, which were introduced purely for purposes of uniformization, (33)–(34) become

$$\begin{aligned}J(1) = {} & \sup_{a(1)\in[\underline{x},\bar{x}]}\frac{1}{\beta+1+\lambda p+\gamma} \\ & \cdot\left[h+\bar{F}(a(1))\left(g(a(1))+\frac{\lambda p}{\beta+\lambda p}J(1)\right)\right. \\ & \quad \left.+(1-\bar{F}(a(1)))J(1)+\lambda pJ(2)+\gamma\left(d+\frac{\lambda p}{\beta+\lambda p}J(1)\right)\right], \qquad (35)\end{aligned}$$

$$\begin{aligned}J(k) = {} & \sup_{a(k)\in[\underline{x},a(k-1)]}\frac{1}{\beta+1+\lambda p+\gamma k} \\ & \cdot[hk+\bar{F}(a(k))g(a(k))+\bar{F}(a(k))J(k-1) \\ & +(1-\bar{F}(a(k)))J(k)+\lambda pJ(k+1)+\gamma kd \\ & +\gamma kJ(k-1)]. \qquad (36)\end{aligned}$$

The monotonicity result of Proposition 1 implies that imposing a constraint that $a(k)$ is decreasing in k does not change the solution of optimality Equations (35)–(36). Therefore,

$$J(0) = \frac{\lambda p}{\beta + \lambda p} J(1), \tag{37}$$

$$J(1) = \sup_{a(1)\in[\underline{x},\,\bar{x}]} \frac{1}{\beta + 1 + \lambda p + \gamma} \cdot \big[h + \bar{F}(a(1))(J(0) + g(a(1))) + (1 - \bar{F}(a(1)))J(1) + \lambda p J(2) + \gamma(J(0) + d)\big], \tag{38}$$

$$J(k) = \sup_{a(k)\in[\underline{x},\,a(k-1)]} \frac{1}{\beta + 1 + \lambda p + \gamma k} \cdot [hk + \bar{F}(a(k))(J(k-1) + g(a(k))) + (1 - \bar{F}(a(k)))J(k) + \lambda p J(k+1) + \gamma k(J(k-1) + d)]. \tag{39}$$

Therefore, the optimal decision thresholds $\{a^p(k)\}$ must solve (1)–(2) with the arrival rate modified at λp. □

Proof of Proposition 6. The proof shall proceed in three steps. In the first step, we consider the relative value function for the optimality Equations (1)–(2), defined by $\Delta(n) \equiv V(n) - V(n-1)$ and $\Delta(0) \equiv V(0)$, and show that it solves a linear program. In the second step, we consider the relative value function when the arrival rate is λ' and show that it is feasible for the linear program corresponding to $\lambda < \lambda'$. This, in turn, implies that the relative value function is decreasing in λ. The claim is then established in Step 3 based on the argument presented in the proof of Proposition 1.

Step 1. Let us express the optimality Equations (1)–(2), in terms of the relative value functions.

$$\beta\Delta(0) = \lambda\Delta(1) \tag{40}$$

$$\beta V(n) = \sup_{b(n)\in[\underline{x},\,\bar{x}]} [hn + \bar{F}(b(n))(g(b(n)) - \Delta(n)) + \lambda\Delta(n+1) + \gamma n(d - \Delta(n))], \quad \forall\, n \geq 1 \tag{41}$$

Using (40) and the fact that $V(n) = \sum_{i=0}^{n} \Delta(i)$, we can write this as $\forall\, n \geq 1$,

$$0 = \sup_{b(n)\in[\underline{x},\,\bar{x}]} \left[hq + \bar{F}(b(n))(g(b(n)) - \Delta(n)) + \gamma n(d - \Delta(n)) + \lambda(\Delta(n+1) - \Delta(1)) - \beta \sum_{i=1}^{n} \Delta(i) \right]. \tag{42}$$

It then follows that the optimal relative value functions must solve the following linear program (see Puterman 1994), where $\{w(n)\}$ are arbitrary positive constants:

$$\min_{\{\Delta(n)\}} \sum_{n=1}^{\infty} w(n)\Delta(n) \tag{43}$$

$$\text{s.t.} \quad 0 \geq \left[hn + \bar{F}(b(n))(g(b(n)) - \Delta(n)) + \gamma n(d - \Delta(n)) + \lambda(\Delta(n+1) - \Delta(1)) - \beta \sum_{i=1}^{n} \Delta(i) \right], \quad \forall\, b(n) \in [\underline{x}, \bar{x}], \forall\, n \geq 1. \tag{44}$$

We shall call this linear program $(LP; \lambda)$.

Step 2. Consider now the relative value function $\Delta'(n)$ that corresponds to an arrival rate $\lambda' > \lambda$. We shall show that $\Delta'(n) \leq \Delta(n)$ for every n.

The optimal relative value functions $\{\Delta(n)\}$ and $\{\Delta'(n)\}$ must solve $(LP; \lambda)$ and $(LP; \lambda')$, respectively. Because we have previously shown that $\Delta(n)$ is nonincreasing in n, it is not difficult to see that $\{\Delta(n)\}$ is feasible in $(LP; \lambda')$. This is because the right-hand side of (44) remains negative when λ is replaced by $\lambda' \geq \lambda$.

Now, suppose that there is some n for which $\Delta'(n) > \Delta(n)$. Then, we can choose positive weights $\{w(n)\}$ such that $\sum_{i=1}^{n} w(i)\Delta'(i) > \sum_{i=1}^{n} w(i)\Delta(i)$, which contradicts the optimality of $\{\Delta'(n)\}$ in $(LP; \lambda')$. Therefore, we must have $\Delta'(n) \leq \Delta(n)$ for every n.

Step 3. Finally, we can show that $b'(n) \leq b(n)$ for every n. This follows directly from Step 2 using the same argument provided in the proof of Proposition 1. □

References

Altman, E., N. Shimkin. 1998. Individual equilibrium and learning in processor sharing systems. *Oper. Res.* **46**(6) 776–784.

Bell, C., S. Stidham. 1983. Individual versus social optimization in the allocation of customers to alternate servers. *Management Sci.* **29**(7) 831–839.

Bertsekas, D. M. 1995. *Dynamic Programming and Optimal Control*, Vol. 2. Athena Scientific, Belmont, MA.

Cullis, J. G., P. R. Jones, C. Propper. 2000. Waiting lists and medical treatment. A. J. Culyer, J. P. Newhouse, eds. *Handbook of Health Economics*, Vol. 1B. Elsevier Science, Amsterdam, The Netherlands.

Feldman, R., F. Lobo. 1997. Global budgets and excess demand for hospital care. *Health Econom.* **6**(2) 187–196.

Fudenberg, D., J. Tirole. 1990. *Game Theory.* MIT Press, Cambridge, MA.

George, J. M., J. M. Harrison. 2001. Dynamic control of a queue with adjustable service rates. *Oper. Res.* **49**(5) 720–731.

Gibbons, R. 1992. *Game Theory for Applied Economists.* Princeton University Press, Princeton, NJ.

Goddard, J., M. Tavakoli. 1994. Rationing and waiting list management—Some efficiency and equity considerations. M. Malek, ed. *Setting Priorities in Health Care*. John Wiley and Sons, Chichester, U.K., 71–92.

Goddard, J., M. Malek, M. Tavakoli. 1995. An economic model of the market for hospital treatment for non-urgent conditions. *Health Econom.* **4**(1) 41–55.

Hassin, R. 1985. Notes and comments on the optimality of first come last served queues. *Econometrica* **53**(1) 201–202.

Howard, D. H. 2001. Dynamic analysis of liver allocation policies. *Medical Decision Making* **21**(4) 257–266.

Iversen, T. 1997. The effect of a private sector on the waiting time of a national health service. *J. Health Econom.* **16**(4) 381–396.

Kaplan, E. H. 1987. Analyzing tenant assignment policies. *Management. Sci.* **33**(3) 395–408.

Larson, R. C. 1987. Perspectives on queues: Social justice and the psychology of queueing. *Oper. Res.* **35**(6) 895–905.

Lippman, S. A., S. Stidham. 1977. Individual versus social optimization in exponential congestion systems. *Oper. Res.* **25**(2) 233–247.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Comm. ACM* **28**(3) 312–321.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* **38**(5) 870–883.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.

Organ Procurement and Transportation Network (*OPTN*). 2003. Retrieved May 5, 2003, http://www.optn.org.

Pastan, S., J. Bailey. 1998. Dialysis therapy. *New England J. Medicine* **338**(20) 1428–1437.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, New York.

Smith, P., A. Van Ackere. 2002. A note on the integration of system dynamics and economic models. *J. Econom. Dynam. Control* **26**(1) 1–10.

Stidham, S. 1978. Socially and individually optimal control of arrivals to a GI/M/1 queue. *Management Sci.* **24**(15) 1598–1610.

Su, X., 2004. Essays on patient choice. Doctoral dissertation, Graduate School of Business, Stanford University, Stanford, CA.

Su, X., S. A. Zenios. 2004a. Patient choice in kidney allocation: A sequential stochastic assignment model. *Oper. Res.* Forthcoming.

Su, X., S. A. Zenios. 2004b. Mechanism design for kidney allocation. Working paper, Graduate School of Business, Stanford University, Stanford, CA.

Su, X., S. A. Zenios, G. M. Chertow. 2004. Incorporating recipient choice in kidney transplantation. *J. Amer. Soc. Nephrology* **15**(6) 1656–1663.

United Network of Organ Sharing (*UNOS*). 2002. *Annual Report of the U.S. Organ Procurement and Transplantation Network and the Scientific Registry of Transplant Recipients: Transplant Data 1992–2001.* http://www.optn.org/data/annualreport.asp.

United States Renal Data System (*USRDS*). 2002. *Annual Data Report: Atlas of End-Stage Renal Disease in the United States.* National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

Van Ackere, A., P. Smith. 1999. Towards a macro model of National Health Service waiting lists. *System Dynam. Rev.* **15**(3) 225–252.

Van Mieghem, J. 2000. Price and service discrimination in queueing systems: Incentive-compatibility of $Gc\mu$ scheduling. *Management Sci.* **46**(9) 1249–1267.

Venables, W. N., B. D. Ripley. 1997. *Modern Applied Statistics with S-PLUS*, 2nd ed. Springer-Verlag, New York.

Votruba M. 2002. Efficiency-equity tradeoffs in the allocation of cadaveric kidneys. Working paper, Princeton University, Princeton, NJ.

Weinstein, M. C. 2001. Should physicians be gatekeepers of medical resources? *J. Medical Ethics* **27**(4) 268–274.

Wolfe, R. A., V. B. Ashby, E. L. Milford, A. O. Ojo, R. E. Ettenger, L. Y. C. Agodoa, P. J. Held, F. K. Port. 1999. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *New England J. Medicine* **341**(23) 1725–1730.

Yates, J. 1995. *Private Eye, Heart and Hip.* Churchill Livingstone, London, U.K.

Yechiali, U. 1971. On optimal balking rules and toll charges in the GI/M/1 queuing process. *Oper. Res.* **19**(2) 349–370.

Zenios, S. A. 1999. Modeling the transplant waiting list: A queueing model with reneging. *Queueing Systems* **31**(3–4) 239–251.

Zenios, S. A., G. M. Chertow, L. M. Wein. 2000. Dynamic allocation of kidneys to candidates on the transplant waiting list. *Oper. Res.* **48**(4) 549–569.