Queueing Systems 31 (1999) 239–251

Modeling the transplant waiting list: A queueing model with reneging

Stefanos A. Zenios
Graduate School of Business, Stanford University, Stanford, CA 94305, USA

Received 19 May 1998; revised 14 September 1998

Motivated by the problem of organ allocation, we develop a queueing model with reneging that provides a stylistic representation of the transplant waiting list. The model assumes that there are several classes of patients, several classes of organs, and patient reneging due to death. We focus on randomized organ allocation policies and develop closed-form asymptotic expressions for the stationary waiting time, stationary waiting time until transplantation, and fraction of patients who receive transplantation for each patient class. Analysis of these expressions identifies the main factors that underlie the performance of the transplant waiting list and demonstrates that queueing models can prove useful when evaluating the organ allocation system.

Keywords: reneging, asymptotic analysis, Laplace's method, transplant waiting list

1. Introduction

Kidney transplantation is the best treatment for thousands of patients suffering from chronic kidney failure. However, cadaveric organs for transplantation remain scarce and the waiting list for transplantation increases rapidly. At the end of 1996, there were 34 550 registered candidates for transplantation, but only 7 833 transplantations were performed during that year. This critical shortage of organs exacerbates the difficulties surrounding the development of a rational allocation policy. In this paper, we pursue a queueing-based analysis of the transplant waiting list and illustrate the main queueing phenomena that underlie its performance.

In the United States, organs are allocated using a system that prioritizes transplant candidates according to a combination of waiting time and medical criteria; see Zenios et al. [14] for a detailed background discussion. However, because this priority system was not the consequence of analysis, its impact on specific performance outcomes is poorly understood. This is demonstrated by the ongoing controversy surrounding the observed differences in waiting times between different groups, and the debate on the type of policies that can best eliminate these differences. In an article that highlights some of the major policy dilemmas that underlie the organ allocation problem, Sanfilippo et al. [11] show that "blacks have longer waiting times than whites" and implore the scientific community to analyze the causes for this phenomenon. However, the absence of appropriate analytical tools makes such analysis difficult and prevents the prospective evaluation of different allocation policies that can potentially eliminate unacceptable inequalities in the waiting times.

In this paper, we make a first attempt to develop the analytical tools that can bridge this gap. Specifically, we propose a simple queueing model that provides a stylistic representation of the transplant waiting list. The model assumes that there are several classes of patients who queue up for transplantation, and several classes of organs which, borrowing from queueing terminology, serve the patients. Because the transplant patients may die while waiting for a transplant, the model also assumes that patients renege. In this setting, the model takes the form of a multiclass, multiserver queueing system with reneging, and the allocation policies take the form of scheduling rules for the servers. Furthermore, for simplicity the allocation policies considered here are open-loop randomized policies.

Our analysis is motivated by the observations in Sanfilippo et al. [11] and focuses on two main objectives. The first one is to identify the major causes for the observed differences in the waiting time between various demographic groups, and the second one is to identify policies that can eliminate such differences. In order to achieve the first objective, we pursue a queueing-based analysis of our model and develop closed-form expressions for three performance outcomes: average waiting time, average waiting time for patients who receive transplantation, and fraction of patients who receive transplantation. These expressions identify the main factors that can generate differences in the mean waiting times between various patient classes. To achieve the second objective, we analyze the closed-form expressions to identify simple randomized policies that eliminate between-class differences in the waiting times and analyze the impact of such policies on the other two outcomes.

A theme that is absent from this paper is the impact of organ allocation policies on clinical outcomes such as patient survival, and the derivation of dynamic organ allocation policies. This theme is undertaken in Zenios et al. [14] where the transplant waiting list is modeled using a fluid model and an optimal control analysis leads to efficient dynamic allocation policies.

Queueing models with reneging phenomena have been used successfully to analyze telecommunication systems (Coffman et al. [5]), emergency rooms (Gnedenko and Kovalenko [7]), public housing systems (Kaplan [8]), and inventory systems with perishable products (Nahmias [10]). Analytical results for the performance of such systems are derived in Anker and Gafarian [1,2], Coffman et al. [5] and Stanford [12]. The main analytical results presented in this paper complement the existing analytical results. Furthermore, the interpretation in the context of the kidney transplantation problem and the focus on policies that equalize the mean waiting time between the various patient classes are, to the best of our knowledge, novel.

The remainder of this paper is organized as follows: section 2 introduces the queueing model and the notation. Section 3 develops an expression for the invariant queue-length distribution of each patient class, and Laplace transforms for the stationary waiting time and the stationary waiting time until transplantation for patients of each class. These transforms lead to integral expressions for the three queueing outcomes described above. Because these expressions are cumbersome, section 4 develops closed-form expressions which hold in an asymptotic regime. The asymptotic regime is representative of the US transplantation system and the interpretation of the asymptotic expressions achieves the two objectives of this article. Concluding remarks appear in section 5.

2. A multiclass queueing model with reneging

In this section, we introduce the queueing model for the transplant waiting list. The model assumes that there are K classes of transplant candidates who arrive according to independent Poisson processes with rate λ_k^+, $k = 1, \ldots, K$, and J classes of organs that arrive according to independent Poisson processes with rate $\hat{\lambda}_j^-$, $j = 1, \ldots, J$; the class definitions are based on the demographic, immunological and physiological characteristics of the patients and donors. The organs are allocated to transplant candidates and force their departure from the waiting list. In addition, patients of each class $k = 1, \ldots, K$ renege from the system due to death after an exponentially distributed amount of time with rate μ_k. The allocation policy takes the form of a static randomized policy. In particular, $v = (v_{jk})$ is the fraction of class j organs that are allocated to patients of class k, and candidates of the same class are allocated organs on a first-come first-transplanted (FCFT) basis.

It is convenient to define several parameters for future reference. Let $\lambda^+ = \sum_{k=1}^{K} \lambda_k^+$ be the total patient arrival rate, $\lambda^- = \sum_{j=1}^{J} \hat{\lambda}_j^-$ be the total organ arrival rate and

$$\lambda_k^- = \sum_{j=1}^{K} \hat{\lambda}_j^- v_{jk}, \quad k = 1, \ldots, K,$$

be the total organ allocation rate to patients of class k. In addition, let

$$\bar{\mu} = \sum_{k=1}^{K} \frac{\lambda_k^+ \mu_k}{\lambda^+}$$

denote the mean patient death rate. The following dimensionless parameters will also be useful: $\rho = \lambda^+/\lambda^-$, which gives the traffic intensity and also represents the organ demand-to-supply ratio, $\rho_k = \lambda_k^+/\lambda_k^-$, $k = 1, \ldots, K$, which give the demand-to-supply ratio for class k patients, and $\nu_k = \lambda_k^-/\mu_k$, $k = 1, \ldots, K$, which give the ratio of the organ allocation rate for patients of class k over their respective death rate. Although the parameters λ_k^-, ρ_k and ν_k depend on the allocation policy v, this is not explicitly reflected in the notation for clarity. We will also assume that the demand for organs exceeds the supply, $\rho > 1$, and that the organ arrival rate is much larger than the individual patient death rates, $\lambda^- \gg \mu_k$, $k = 1, \ldots, K$. Both conditions currently hold in the United States.

Our multiclass queueing model is closely related to the classical multiple-class, multiple-server queueing system with reneging. In this setting, the transplant patients correspond to customers, the organs correspond to servers, the departure of patients from the waiting list due to death corresponds to customer reneging, and the allocation policy corresponds to the server scheduling rule. The only difference between our model and the classical queueing model is that in our model the patient-customers depart from the system when service begins and not when service ends; that is, the service initiation epochs correspond to the arrival of an organ. Despite this difference between the two models, the queue-length process for our model is equivalent to the queue-length process for the traditional queueing model with reneging.

Our queueing model is also related to the queueing models with negative customers (or signals) which were introduced by Gelenbe [6]. The distinguishing feature of these models is their inclusion of positive customers (in our setting transplant candidates), which behave as customers in traditional queueing models, and negative customers (organs), which play the role of a secondary service. In this context, our queueing model is equivalent to a multiclass $M/M/\infty$ queue with several classes of negative customers. The connection between our model and queueing models with negative arrivals justifies the notation $\lambda_k^-,\ k = 1, \ldots, K$.

The assumption to model the allocation policies by simple randomization rules was made for analytical tractability. A convenient consequence of this assumption is that the queueing system decomposes into K independent queueing systems, one for each patient class. Therefore, to do performance analysis for this queueing system, one can start by analyzing the simpler system that has one class of patients and one class of organs. This is the approach taken in this paper.

3. Major performance results

In this section, we derive integral expressions for the following three performance outcomes for each patient class: expected stationary waiting time, expected stationary waiting time for patients who receive transplantation, and the stationary fraction of patients who receive transplantation. The derivation of these integral expressions proceeds in three steps. First, an expression is derived for the equilibrium distribution of the queue-length process. Second, using the equilibrium distribution we derive expressions for the Laplace transform of the stationary waiting time for each patient class, and the stationary waiting time for patients of each class who receive transplantation. Third, using simple properties of Laplace transforms, we derive integral expressions for the three performance outcomes.

Assuming that the allocation policy is v and that there exists a stationary distribution for the queue-length process, let $L_k,\ D_k$ and W_k denote, respectively, the stationary number of class k patients in the system, the stationary waiting time for patients of class k, and the stationary waiting time for patients of class k who receive transplantation. In addition, let ϕ_k denote the stationary fraction of class k patients who receive transplantation. Finally, let $\pi_k(n) = P(L_k = n),\ k = 1, \ldots, K,\ n \in \mathbb{N}$, denote the stationary distribution for the queue-length process, $G_{D_k}(s) = E[\mathrm{e}^{-sD_k}]$ and $G_{W_k}(s) = E[\mathrm{e}^{-sW_k}]$. It is also convenient to define the stationary *virtual waiting time* for patients of class $k = 1, \dots, K$, denoted by V_k, which gives the stationary amount of time that elapses between the arrival of a class k patient and the instance that this patient would receive an organ (assuming that the patient will not have died by then); in the traditional queueing context, this variable gives the stationary *workload* from class k patients.

The first result is elementary and specifies the stationary probability distribution.

Proposition 1. The stationary probability distribution for the queue-length process under a randomized allocation policy v is

$$\pi_k(n) = \frac{1}{n!} \cdot \frac{\int_0^1 (\rho_k \nu_k z)^n (1-z)^{\nu_k - 1}\,\mathrm{d}z}{\int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k - 1}\,\mathrm{d}z}, \quad k = 1, \dots, K, \tag{1}$$

and the expected stationary queue length is

$$E[L_k] = \rho_k \nu_k \frac{\int_0^1 z(1-z)^{\nu_k - 1} \exp(\rho_k \nu_k z)\,\mathrm{d}z}{\int_0^1 (1-z)^{\nu_k - 1} \exp(\rho_k \nu_k z)\,\mathrm{d}z}, \quad k = 1, \dots, K. \tag{2}$$

Proof. These expressions follow from the flow balance equations and the elementary properties of the gamma and beta functions. The details of the proof are almost identical to the proof of Coffman et al. [5, proposition 1].

The flow balance equations state

$$\pi_k(n)\big(\lambda_k^- + n\mu_k\big) = \pi_k(n-1)\lambda_k^+, \quad n = 1, 2, \dots,$$

and this implies that

$$\pi_k(n) = \frac{C}{\rho_k \nu_k} \prod_{j=0}^{n} \frac{\lambda_k^+}{\lambda_k^- + j\mu_k} = \frac{C}{\rho_k \nu_k} \prod_{j=0}^{n} \frac{\lambda_k^+/\mu_k}{\lambda_k^-/\mu_k + j} = C(\rho_k \nu_k)^n \prod_{j=0}^{n} \frac{1}{\nu_k + j}, \tag{3}$$

where C is a normalization constant.

Expression (1) can now be obtained by expressing the right-hand side of (3) in terms of the gamma and beta functions (see Arfken [3])

$$\begin{aligned}
\pi_k(n) &= C(\rho_k \nu_k)^n \prod_{j=0}^{n} \frac{1}{\nu_k + j} = C(\rho_k \nu_k)^n \frac{\Gamma(\nu_k)}{\Gamma(\nu_k + n + 1)} \\
&= \frac{C(\rho_k \nu_k)^n}{n!} \cdot \frac{\Gamma(\nu_k)\Gamma(n+1)}{\Gamma(\nu_k + n + 1)} = \frac{C(\rho_k \nu_k)^n}{n!} B(\nu_k, n+1) \\
&= \frac{C}{n!} \int_0^1 (\rho_k \nu_k z)^n (1-z)^{\nu_k - 1}\,\mathrm{d}z. \qquad (4)
\end{aligned}$$

Substitution of (4) into $\sum_{n=0}^{\infty} \pi_k(n) = 1$ shows that

$$C = 1 \Big/ \int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k - 1}\,\mathrm{d}z,$$

and thus

$$\pi_k(n) = \frac{1}{n!} \cdot \frac{\int_0^1 (\rho_k \nu_k z)^n (1-z)^{\nu_k - 1}\,\mathrm{d}z}{\int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k - 1}\,\mathrm{d}z}.$$

The derivation of the expected stationary queue length involves straightforward calculations and is omitted. □

The next result gives expressions for the Laplace transform of the stationary virtual waiting time, the stationary waiting time and the stationary waiting time for patients who receive transplantation.

Proposition 2. Under a randomized allocation policy v, the Laplace transforms for the stationary virtual waiting time, the stationary waiting time and the stationary waiting time for patients who receive transplantation for each patient class $k = 1, \ldots, K$ are as follows:

$$G_{V_k}(s) = \frac{\int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k + s/\mu_k - 1}\,\mathrm{d}z}{\int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k - 1}\,\mathrm{d}z}, \tag{5}$$

$$G_{D_k}(s) = \frac{\mu_k}{\mu_k + s}\bigl(1 - G_{V_k}(\mu_k + s)\bigr) + G_{V_k}(\mu_k + s), \tag{6}$$

$$G_{W_k}(s) = \frac{G_{V_k}(\mu_k + s)}{G_{V_k}(\mu_k)}. \tag{7}$$

Proof. The proof examines the system from the perspective of a newly arrived "tagged" patient of class k (see Kleinrock [9]) and proceeds in three steps. In the first step, a simple conditioning argument leads to the expression for $G_{V_k}(s)$. In the second step, the waiting time for the tagged patient is expressed as the minimum of the virtual waiting time and the reneging time, and from this the expression for $G_{D_k}(s)$ is derived. In the third step, an expression is derived for the conditional distribution of the patient waiting time *given* that the patient receives transplantation and this leads to the expression for $G_{W_k}(s)$.

Step 1. Because Poisson arrivals see time average, the Laplace transform for the stationary virtual waiting time V_k satisfies

$$E\bigl[\mathrm{e}^{-sV_k}\bigr] = \sum_{n=0}^{\infty} \pi_k(n) E\bigl[\mathrm{e}^{-sV_k} \mid L_k = n\bigr]. \tag{8}$$

Because the service discipline within each patient class is FCFT, the conditional virtual waiting time V_k *given* that the queue length seen by the "tagged" patient is $L_k = n$ can be expressed as $X_0 + \cdots + X_n$, where $X_j,\ j = 0, \ldots, n$, are independent exponential random variables with rates $\lambda_k^- + (n-j)\mu_k$. It follows that

$$E\bigl[\mathrm{e}^{-sV_k} \mid L_k = n\bigr] = \prod_{j=0}^{n} \frac{\lambda_k^- + (n-j)\mu_k}{\lambda_k^- + s + (n-j)\mu_k}. \tag{9}$$

Substituting (9) into (8), and using simple properties of the beta and gamma functions, we derive the expression for $G_{V_k}(s)$.

Step 2. Let $g_{V_k}(y)$ denote the probability density function for V_k. Because the waiting time for a tagged patient, D_k, is the minimum of the virtual waiting time V_k and the reneging time X, and because the reneging time is exponential with rate μ_k, it follows that

$$G_{D_k}(s) = \int_{y=0}^{\infty}\int_{x=0}^{y} \mathrm{e}^{-sx}\mu_k \mathrm{e}^{-\mu_k x} g_{V_k}(y)\,\mathrm{d}x\,\mathrm{d}y + \int_{y=0}^{\infty}\int_{x=y}^{\infty} \mathrm{e}^{-sy}\mu_k \mathrm{e}^{-\mu_k x} g_{V_k}(y)\,\mathrm{d}x\,\mathrm{d}y, \tag{10}$$

which yields (6).

Step 3. To derive equation (7), we let T_k denote the event that the tagged patient undergoes transplantation (which is equivalent to the event that the virtual waiting time V_k is less than the reneging time, X) and let $f_{D_k|T_k}(w)$ denote the conditional probability density for the waiting time of the tagged patient *given* that the patient receives transplantation. It follows that

$$G_{W_k}(s) = \int_0^{\infty} \mathrm{e}^{-sw} f_{D_k|T_k}(w)\,\mathrm{d}w. \tag{11}$$

Furthermore, assuming that δw is sufficiently small,

$$\begin{aligned}
&P\bigl(D_k \in [w, w+\delta w) \mid T_k\bigr) \\
&\quad = \frac{P(D_k \in [w, w+\delta w), T_k)}{P(T_k)} = \frac{P(D_k \in [w, w+\delta w), X > V_k)}{P(X > V_k)} \qquad (12)\\
&\quad = \frac{P(V_k \in [w, w+\delta w), X > V_k)}{P(X > V_k)} = \delta w \frac{g_{V_k}(w)\int_{x=w}^{\infty}\mu_k \mathrm{e}^{-\mu_k x}\,\mathrm{d}x}{\int_{y=0}^{\infty} P(X > y) g_{V_k}(y)\,\mathrm{d}y} + \mathrm{o}(\delta w) \\
&\quad = \delta w \frac{g_{V_k}(w)\,\mathrm{e}^{-\mu_k w}}{\int_{y=0}^{\infty}\mathrm{e}^{-\mu_k y} g_{V_k}(y)\,\mathrm{d}y} + \mathrm{o}(\delta w) = \delta w \frac{g_{V_k}(w)\,\mathrm{e}^{-\mu_k w}}{G_{V_k}(\mu_k)} + \mathrm{o}(\delta w). \qquad (13)
\end{aligned}$$

This implies that

$$f_{D_k|T_k}(w) = \lim_{\delta w \to 0} \frac{P(D_k \in [w, w+\delta w) \mid T_k)}{\delta w} = \frac{g_{V_k}(w)\,\mathrm{e}^{-\mu_k w}}{G_{V_k}(\mu_k)}. \tag{14}$$

Substituting (14) into (11) gives (7). □

With the expressions for the Laplace transforms at hand, we can now proceed to derive expressions for the three main performance outcomes of interest. These are given in the following proposition:

Proposition 3. Under the randomized policy v,

$$E[D_k] = \frac{1}{\mu_k} \cdot \frac{\int_0^1 z(1-z)^{\nu_k - 1} \exp(\rho_k \nu_k z)\,\mathrm{d}z}{\int_0^1 (1-z)^{\nu_k - 1} \exp(\rho_k \nu_k z)\,\mathrm{d}z}, \tag{15}$$

$$E[W_k] = \frac{1}{\mu_k} \cdot \frac{\int_0^1 \exp(\rho_k \nu_k z) \ln(1/(1-z))(1-z)^{\nu_k}\,\mathrm{d}z}{\int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k}\,\mathrm{d}z}, \tag{16}$$

$$\phi_k = \frac{\int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k}\,\mathrm{d}z}{\int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k - 1}\,\mathrm{d}z}. \tag{17}$$

Proof. Expressions (15) and (16) follow from $E[D_k] = -G'_{D_k}(0)$ and $E[W_k] = -G'_{W_k}(0)$, respectively.

To derive (17), notice that the elementary renewal theorem implies that $\phi_k = P(V_k < X)$, where X is the reneging time for the tagged customer defined in the proof of proposition 2. Furthermore, because X is exponential with rate μ_k it follows that

$$\phi_k = G_{V_k}(\mu_k) = \frac{\int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k}\,\mathrm{d}z}{\int_0^1 \exp(\rho_k \nu_k z)(1-z)^{\nu_k - 1}\,\mathrm{d}z}. \qquad \square$$

4. Asymptotic expressions

In this section, we derive asymptotic expressions for the expected stationary waiting time, expected stationary waiting time for patients who receive transplantation and fraction of patients who receive transplantation for each patient class. The asymptotic regime for these expressions assumes that the patient and organ arrival rates become infinitely large, the reneging rates are maintained fixed, and the patient arrival rate for each class exceeds the corresponding organ allocation rate. This asymptotic regime provides a realistic approximation for the US transplantation system where the patient and organ arrival rates are considerably larger than the individual patient mortality rates, and the demand for organs far exceeds the supply.

First, we must formally define the asymptotic regime. Assume that the candidate arrival rates are $n\lambda_k^+$, $k = 1, \ldots, K$, the organ arrival rates are $n\hat{\lambda}_j^-$, $j = 1, \ldots, J$, and the reneging rates are μ_k, $k = 1, \ldots, K$. Also, assume that the allocation policy v is independent of the scaling factor n, and define $n\lambda_k^-$, $k = 1, \ldots, K$, to be the organ allocation rates to each patient class. The asymptotic regime assumes that $\lambda_k^- < \lambda_k^+$, $k = 1, \ldots, K$, and lets $n \to \infty$. The following proposition gives the main asymptotic results.

Proposition 4. The following asymptotics hold as $n \to \infty$:

$$E[D_k] \sim \frac{1}{\mu_k}\left(1 - \frac{\lambda_k^-}{\lambda_k^+}\right), \tag{18}$$

$$E[W_k] \sim \frac{1}{\mu_k} \ln\left(\frac{\lambda_k^+}{\lambda_k^-}\right), \tag{19}$$

$$\phi_k \sim \frac{\lambda_k^-}{\lambda_k^+}. \tag{20}$$

In order to prove this proposition, we need the following lemma which describes the typical application of Laplace's method; see Carrier et al. [4, p. 254].

Lemma 1. Consider the integral $f(x) = \int_a^b \mathrm{e}^{xh(t)} q(t)\,\mathrm{d}t$, where all quantities are real. Suppose that $h(t)$ attains its maximum at the point t_0 which is in the interior of $[a, b]$, and $h(t_0)$ exists and is twice differentiable. Then, the first term in the asymptotic expansion of $f(x)$ as $x \to \infty$ is given by

$$f(x) \sim q(t_0)\,\mathrm{e}^{xh(t_0)} \sqrt{-\frac{2\pi}{xh''(t_0)}}.$$

We are now in a position to present the proof of proposition 4. This proof is motivated by the analysis in Coffman et al. [5].

Proof of proposition 4. Proposition 3 implies that

$$E[D_k] = \frac{1}{\mu_k} \cdot \frac{\int_0^1 z(1-z)^{(n\lambda_k^-/\mu_k)-1} \exp(n\lambda_k^+ z/\mu_k)\,\mathrm{d}z}{\int_0^1 (1-z)^{(n\lambda_k^-/\mu_k)-1} \exp(n\lambda_k^+ z/\mu_k)\,\mathrm{d}z}. \tag{21}$$

We can now obtain asymptotic expansions for the two integrals in (21). First, one can show that

$$\int_0^1 z(1-z)^{(n\lambda_k^-/\mu_k)-1} \exp\left(\frac{n\lambda_k^+ z}{\mu_k}\right)\mathrm{d}z \tag{22}$$

$$= \int_0^1 \frac{z}{1-z} \exp\left(\frac{n\lambda_k^-}{\mu_k}\left(\frac{\lambda_k^+ z}{\lambda_k^-} + \ln(1-z)\right)\right)\mathrm{d}z. \tag{23}$$

Because the maximum of $\ln(1-z) + z\lambda_k^+/\lambda_k^-$ is attained at $z_0 = 1 - \lambda_k^-/\lambda_k^+$ which is in the interior of $[0, 1]$, it follows from lemma 1 that

$$\int_0^1 z(1-z)^{(n\lambda_k^-/\mu_k)-1} \exp\left(\frac{n\lambda_k^+ z}{\mu_k}\right)\mathrm{d}z$$

$$\sim \frac{\lambda_k^+ - \lambda_k^-}{\lambda_k^-} \exp\left(\frac{n\lambda_k^-}{\mu_k}\left(\frac{\lambda_k^+}{\lambda_k^-} - 1 + \ln\left(\frac{\lambda_k^-}{\lambda_k^+}\right)\right)\right) \sqrt{\frac{2\pi}{(n\lambda_k^-/\mu_k)(\lambda_k^+/\lambda_k^-)^2}}. \tag{24}$$

Similarly, one can obtain the following expression for the integral in the denominator of (21):

$$\int_0^1 (1-z)^{(n\lambda_k^-/\mu_k)-1} \exp\left(\frac{n\lambda_k^+ z}{\mu_k}\right) \mathrm{d}z$$
$$\sim \frac{\lambda_k^+}{\lambda_k^-} \exp\left(\frac{n\lambda_k^-}{\mu_k}\left(\frac{\lambda_k^+}{\lambda_k^-} - 1 + \ln\left(\frac{\lambda_k^-}{\lambda_k^+}\right)\right)\right)\sqrt{\frac{2\pi}{(n\lambda_k^-/\mu_k)(\lambda_k^+/\lambda_k^-)^2}}. \qquad (25)$$

Substituting equations (24) and (25) into equation (21) gives (18).

Similar calculations lead to the asymptotic expressions for $E[W_k]$ and ϕ_k. $\square$

Having derived closed-form asymptotic expressions for the three performance outcomes, we are now in a position to address the two objectives of this paper.

With respect to the first objective, expression (18) implies that differences in the expected stationary waiting times between various classes can be caused either by differences in the allocation rates or differences in the mortality rates between these classes. Although the former is recognized explicitly in the medical literature the latter is not. Furthermore, the latter provides an additional factor that may contribute to the longer waiting time for African-Americans. Specifically, Wolfe et al. [13] shows that the mortality rates for African-Americans on the waiting list are lower than Caucasians. Therefore, expression (18) implies that African-Americans would have longer mean waiting times than Caucasians, even if they had the same allocation rates. An empirical evaluation to the extent to which the lower mortality rate of African-Americans contributes to their longer waiting times is beyond the scope of this paper.

Let us now turn to the second objective. The following proposition identifies the policy that can equalize the expected stationary waiting times for all patient classes and examines the impact of that policy on the mean waiting time of patients of each class who receive transplantation.

Proposition 5. As $n \to \infty$, the expected stationary waiting times for all patient classes can be asymptotically equalized if and only if $\max_k\{\mu_k\} < \rho\bar{\mu}/(\rho-1)$. If this condition holds, then the static policy that equalizes the expected stationary waiting times for all classes is given by

$$\lambda_k^- = \left(\frac{(1-\rho)\mu_k + \rho\bar{\mu}}{\bar{\mu}}\frac{\lambda_k^+}{\lambda^+}\right)\lambda^-, \quad k = 1,\ldots,K. \qquad (26)$$

Furthermore, under such a policy, the expected stationary waiting time for patients of class $k = 1,\ldots,K$ who receive transplantation is

$$E[W_k] = -\frac{1}{\mu_k}\ln\left(1 - \frac{(\rho-1)\mu_k}{\rho\bar{\mu}}\right). \qquad (27)$$

Proof. The proof proceeds in two steps. In the first step, asymptotic expression (18) is used to derive a system of linear equalities and inequalities which characterize the allocation rates that equalize the asymptotic expected stationary waiting time across all patient classes. In the second step, this system is analyzed. The analysis leads to necessary and sufficient conditions which guarantee existence of a solution, and establishes equations (26) and (27).

Step 1. The asymptotic expression (18) implies that the allocation rates λ_k^-, $k = 1, \ldots, K$, which equalize the asymptotic expected stationary waiting time across all patient classes must satisfy the following equalities:

$$\frac{1}{\mu_1}\left(1 - \frac{\lambda_1^-}{\lambda_1^+}\right) = \cdots = \frac{1}{\mu_K}\left(1 - \frac{\lambda_K^-}{\lambda_K^+}\right). \tag{28}$$

Moreover, they must satisfy the following consistency constraint, which states that they must be consistent with the asymptotic regime described earlier in this section:

$$\lambda_k^- < \lambda_k^+, \quad k = 1, \ldots, K, \tag{29}$$

and a second constraint which states that the allocation policy should allocate all available organs

$$\sum_{k=1}^{K} \lambda_k^- = \lambda^-. \tag{30}$$

Step 2. Next, we derive an expression for the solution to equations (28)–(30). Assuming that $w > 0$ denotes the common value of the asymptotic expected stationary waiting time for all patient classes, then equation (28) implies that the allocation rates for each class must satisfy the following expression:

$$\lambda_k^- = (1 - w\mu_k)\lambda_k^+, \quad k = 1, \ldots, K. \tag{31}$$

Substituting expression (31) into (29) and (30), respectively, we conclude that

$$\mu_k < \frac{1}{w}, \quad k = 1, \ldots, K, \tag{32}$$

and

$$w = \frac{\rho - 1}{\rho\bar{\mu}}. \tag{33}$$

Finally, combining (31)–(33) we conclude that equations (28)–(30) have a solution if and only if $\mu_k < \rho\bar{\mu}/(\rho - 1)$ for all $k = 1, \ldots, K$. If that condition holds, then the allocation policy that equalizes the stationary waiting times across all patient classes is given by

$$\lambda_k^- = \left(1 - \frac{\rho - 1}{\rho\bar{\mu}}\mu_k\right)\lambda_k^+ = \left(\frac{(1-\rho)\mu_k + \rho\bar{\mu}}{\bar{\mu}} \cdot \frac{\lambda_k^+}{\lambda^+}\right)\lambda^-, \quad k = 1, \ldots, K.$$

This establishes equation (26), and that the expected stationary waiting times for all patient classes can be asymptotically equalized if and only if $\max_k\{\mu_k\} < \rho\bar{\mu}/(\rho - 1)$.

Finally, substituting (26) into the asymptotic expressions for $E[W_k]$ we derive expression (27). □

This proposition implies that to equalize the mean stationary waiting times for all patient classes, one must allocate a higher percentage of organs to the patients with the lowest mortality rate. However, this is done at the expense of differences in the mean waiting times for patients of each class who receive transplantation and the fraction of patients of each class who receive transplantation. This implies that one should not evaluate an allocation policy based only on its impact on waiting times. Instead, one must consider at least two of the three outcomes considered in this report and thus obtain a more concise picture about the impact of a policy on the various classes of patients considered. In particular, a reasonable approach would be to consider the mean waiting times for each patient class, which include the confounding effect of mortality, and the fraction of patients of each class who receive transplantation, which isolate the confounding effect of mortality when organs are allocated using a randomized rule.

5. Concluding remarks

The organ allocation problem remains one of the most perplexing problems of modern medicine and several attempts have been made over the years to develop effective policies for organ allocation. However, to date these attempts were met with partial success for several reasons. One of them is the absence of sophisticated enough analytical tools that can provide the framework for evaluating and comparing different allocation policies.

In this paper we have made a first attempt to develop a queueing model that bridges this gap. Although the model makes several simplifying assumptions for analytical tractability, it captures some of the major effects that underlie the organ allocation problem. Namely, it contains several classes of patients and organs, it allows patients to withdraw from the waiting list due to death, and it also assumes that the demand for organs far exceeds the supply. Furthermore, an extensive simulation study reported in Zenios et al. [14] demonstrates that the qualitative insights derived for this simple queueing model hold for a more general queueing model with non-exponential reneging and time inhomogeneous Poisson arrivals.

Asymptotic analysis of this model leads to closed-form expressions for three queueing-based outcomes: waiting time, waiting time for patients who receive transplantation, and fraction of patients who receive transplantation. These expressions also reveal two driving forces behind the performance of any allocation policy: the allocation rates and the mortality rates for the various patient classes. The identification of the mortality rate as one of the main driving forces behind the performance of the allocation policies motivates a new hypothesis that can explain a paradox that has defied the transplantation community for several years: what are the main reasons that cause African-Americans to have longer waiting times than Caucasians? In addition, the results show that in order to evaluate the performance of an allocation policy, one should consider both the impact of the policy on waiting times and on the percentage of candidates who receive transplantation. Focusing on waiting time alone can be misleading because the patient mortality serves as a confounding factor.

In addition, our analysis demonstrates that queueing theory can prove useful in the ongoing debate about the development of a rational organ allocation policy and can provide the conceptual framework for understanding the performance of the transplant waiting list. A policy-oriented companion paper, Zenios et al. [15], exploits the insights derived from the queueing analysis and provides a comprehensive evaluation of alternative organ allocation policies.

Acknowledgements

I would like to thank Larry Wein for suggesting this research problem. I have also benefited from numerous discussions with Glenn Chertow, Dick Larson, Arnie Barnett, David Paltiel, Ed. Kaplan and David Markowitz. This research was supported in part by a dissertation fellowship 30-P-90673/1-01 from HCFA.

References

[1] C.J. Ancker and A.V. Gafarian, Some queueing problems with balking and reneging I, Oper. Res. 11 (1963) 88–100.
[2] C.J. Ancker and A.V. Gafarian, Some queueing problems with balking and reneging II, Oper. Res. 11 (1963) 928–937.
[3] G. Arfken, *Mathematical Methods for Physicists* (Academic Press, New York, 1985).
[4] G.F. Carrier, M. Krook and C.E. Pearson, *Functions of Complex Variables* (Hod Books, New York, 1983).
[5] E.G. Coffman, A.A. Puhalskii, M.I. Reiman and P.E. Wrigth, Processor-shared buffers with reneging, Performance Evaluation 19 (1994) 25–46.
[6] E. Gelenbe, Product form queueing networks with negative and positive customers, J. Appl. Probab. 28 (1991) 656–663.
[7] B.V. Gnedenko, and I.N. Kovalenko, *Introduction to Queueing Theory* (Israel Program for Scientific Translation, Israel, 1968).
[8] E.H. Kaplan, Tenant assignment models, Oper. Res. 34 (1986) 832–408.
[9] L. Kleinrock, *Queueing Systems*, Vol. 1: Theory (Wiley, New York, 1975).
[10] S. Nahmias, Perishable inventory theory: A review, Oper. Res. 30 (1982) 680–708.
[11] F.P. Sanfilippo, W.K. Vaughn, T.G. Peters, C.F. Shield, III, P.L. Adams et al., Factors affecting the waiting time of cadaveric kidney transplant candidates in the United States, J. Amer. Medical Assoc. 1267 (1992) 247–252.
[12] R.E. Stanford, Reneging phenomena in single channel queues, Math. Oper. Res. 4 (1979) 162–178.
[13] R.A. Wolfe, V.B. Ashby, E.L. Milford, A.O. Ojo, R.E. Ettenger, L.Y.C. Agodoa, P.J. Held and F.K. Port, Patient survival for wait-listed versus cadaveric renal transplant patients in the United States, J. Amer. Soc. Nephrology 8 (1997) 708A.
[14] S.A. Zenios, G.M. Chertow and L.M. Wein, Dynamic allocation of kidneys to candidates on the transplant waiting list, Oper. Res. (1999), to appear.
[15] S.A. Zenios, L.M. Wein and G.M. Chertow, Evidence-based organ allocation, Amer. J. Medicine (1999), to appear.