A Model for Deceased-Donor Transplant Queue Waiting Times

Steve Drekic[*], David A. Stanford[†], Douglas G. Woolford[‡], Vivian C. McAlister M.D.[§]

Abstract

In many jurisdictions, organ allocation is done on the basis of the health status of the patient, either explicitly or implicitly. This paper presents a self-promoting priority queueing model for patient waiting times which takes into account changes in health status over time. In this model, most patients arrive as "regular" customers to the queue, but as the health of a patient degrades, their status is promoted to "priority" to reflect the increased urgency of the transplant. We model the queueing system as a level-dependent quasi-birth-and-death process, and the steady-state joint queue length distribution as well as the marginal delay distributions for each queue are computed via the use of matrix analytic techniques. The model is calibrated using liver transplantation wait-list data, provided by a regional health centre in Canada, that tracked approximately 1,100 patients over nearly 13 years. Blood-type specific models are fit and performance measures, such as the mean and distribution of the time until transplant, are obtained and compared to empirical estimates calculated using the raw data.

Keywords: Abandonments, Competing risks, Estimation, Liver transplantation, Phase-type distribution, Quasi-birth-and-death process, Reneging.

Mathematics Subject Classification: 60K25   90B22

---

[*]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada, sdrekic@uwaterloo.ca

[†]Department of Statistical & Actuarial Sciences, Western University, London, ON, N6A 5B7, Canada, stanford@stats.uwo.ca

[‡]Department of Mathematics, Wilfrid Laurier University, Waterloo, ON, N2L 3C5, Canada, dwoolford@wlu.ca

[§]Department of Surgery, Western University, London, ON, N6A 5B7, Canada, vmcalist@uwo.ca

1 Introduction

Solid organ transplantation is a therapy that is used widely around the world, to treat patients whose life expectancy can be meaningfully increased by the replacement of one or more organs that are failing. During the half century or so that transplantation has been a viable option in the treatment of patients with a failing organ, wait lists have been established for patients awaiting a suitable organ. In many jurisdictions, these wait lists have operated under a variety of modifications to the first come, first transplanted (FCFT) discipline, to account for the patient's health status. In recent years, however, patterns have developed which specifies this modification in a formal way. For example, Wiesner et al. [27] developed the "Model for End-stage Liver Disease (MELD) Score" as a means for ranking patients periodically so that those whose health status was degrading more quickly would gain priority access to a deceased-donor liver. The resulting service discipline used for the liver transplant wait lists does not fit into the realm of existing FCFT or priority queueing models, since a wait-listed patient's priority depends upon their health status, which in turn is influenced by the amount of time they have spent waiting.

Delays while patients await transplantation are frequently lengthy, often on the order of years, as has been observed by numerous authors for multiple organ types in a variety of jurisdictions. A sample of these studies follows which gives an indication of the variety of contexts in which this problem has been recognized: kidneys in Germany (Glander et al. [7] and Liefeldt et al. [18]), hearts in the UK (Hussey et al. [9]), kidneys in the US (Zenios et al. [29]), and livers in the US (Barone et al. [2]) and in Canada (Stanford et al. [21]).

Despite this fact, the use of analytical queueing models specifically to address transplant waiting times on the wait lists has been somewhat limited. The goal of this paper is to present a full analysis and application of a queueing model which we develop for liver-transplant patients of each ABO blood type that reflects the sickest patient first aspect and allows for abandonments. We derive the steady-state queue length and marginal delay distributions. We derive estimators for the parameters, and we calibrate and assess the fit

of the model using real wait-list data. Our work demonstrates how queueing theory can produce a model which can be used to provide a reasonable indication of key performance measures, such as the likelihood of successful transplantation (either prior to or after a degradation in health status), and the likelihood of abandonment or of death while waiting.

Zenios [28] appears to have been the first to present a queueing model for transplant waiting times. This queueing model was used in turn by Zenios et al. [29] in their study of kidney transplant wait lists in the US. The kidney allocation problem was revisited by Su and Zenios [23], from the perspective of customer choice. Further details on the specific assumptions of these papers follows later on in this section.

Stanford et al. [21] combined a statistical analysis of patient placements and deceased-donor organ availability with a section presenting simple queueing models to assess the consequences of three strategies to try to close a well-documented gap between the demand for donor livers and their availability in Canada. The statistical analysis revealed that in most of the Canadian regional liver transplant wait lists between 2000 and 2004, deceased-donor organ availability was well-approximated by a Poisson process. At the same time, patient placement differed significantly from a Poisson process in all cases. Working with a GI/M/1 model, the authors investigated the relative merits of increased donor card signing, greater reliance upon living donors, and a lottery system to allocate the limited number of organs available.

Recently, Stanford et al. [22] used a queueing model to respond to a recent development in addressing the so-called "Blood Type O Problem" (see [7]), in which too many type O organs are cross-transplanted to compatible blood groups, causing notably longer waits for blood type O recipients due to the reduced supply of O organs. Glander et al. [7] noted that this frequently has led to situations in which blood type O recipients experience worse health outcomes in terms of statistically-significant higher mortality rates during their longer waiting times. The authors also observed that there are "poorer graft outcomes" for type O recipients as a result. The analysis in [7] and [18] led to a change in the kidney transplant policy employed by the the Eurotransplant zone effective late 2011, insisting

3

upon ABO-identical transplantation. Aware of this development, Stanford et al. [22] developed a construct called the "Array of Idealized Transplant Queues (AITQ)", which was used to show that an ABO-identical transplant policy would be incapable in the long run of delivering comparable waiting times for all blood groups in the Canadian context (and by implication, any jurisdiction with a close match to the Canadian blood mix of 46% O, 42% A, 9% B, and 3% AB – see [5]). While ABO-identical transplantation would indeed resolve the blood type O problem, it would replace it with an even worse problem for patients of blood types B and AB, due to the relative rarity of these organs in the population at large. The analysis in [22] established that ABO-identical transplantation would lead to delays typically five (respectively, fifteen) times longer for blood group B (respectively, AB) patients in Canada and other countries with a close mix of blood types in the population.

We note as an aside that there are also studies involving simulations from queueing models which have been used to numerically examine the behaviour of kidney transplant wait lists; we note in particular the work of Abellán et al. [1] in this regard. However, we will not be considering simulation models in the remainder of this paper.

The goal of Stanford et al. [22] was to establish the long-run inability of an ABO-identical transplantation policy to deliver comparable waiting times for all blood groups. The question being addressed was the long-run system waiting time experience of the four blood groups, rather than individual patient experience. As such, its purpose was to address transplant policy, whereas the present paper has as its goal the development of a queueing model to assess patient waiting times on the wait lists. In our view, all of the previous models have either ignored or inadequately described one key aspect of deceased-donor transplant wait lists or another, as we now review. The interested reader is advised to also review the text in [22] where the distinguishing factors which arise in transplant queue settings are discussed at some length.

The first such factor is queue abandonment: the tendency for patients to leave the wait list due to death, a degradation in health status, personal reasons, or even in some cases,

improvement in patient health to the point where transplantation is no longer the preferred therapy. Stanford et al. [22] commented upon, but did not incorporate a mechanism for abandonments, as it affects all blood groups, and as such had secondary impact upon the blood type issue they addressed. Zenios [28] did account for abandonments, but did so at an exponential rate independent of the queue length, as opposed to a rate that is proportional to the number on the wait lists, which is a feature of the model we present herein.

The second factor is the service discipline: the fact that many transplant queues do not at present typically follow a FCFT discipline, but instead treat wait-listed patients who experience a degradation in health status on a priority basis (so long as they are still candidates for transplantation). None of the foregoing transplant papers reflected this priority aspect, which we attempt to do by invoking the idea of patients who "self promote" from regular status to priority status at an exponential rate, while waiting. The "self-promoting" literature is fairly extensive, comprising Krishnamoorthy and Narayanan [11], Wang [26], Gómez-Corral et al. [8], and Krishnamoorthy et al. [12, 13, 14, 15]. In telecommunication systems, the idea of promoting packets with low priority status to the high priority queue is referred to as "priority jumps" and has been studied in the recent past as well (e.g., see Maertens et al. [19], and references therein).

It is our view that a third factor, not adequately discussed in the queueing models apart from [22], is the fundamental role that patient and donor ABO status plays in determining individual patient waiting time experience. We feel it needs to be explicitly factored into the wait time modelling, as we do, for most jurisdictions in the world. A particular exception might be the US case, where the transplant system comprises a large network of centres arranged nationally, regionally, and locally, in which organs are procured under the Organ Procurement and Transportation Network (OPTN) and allocated under the rules of the United Network of Organ Sharing (UNOS). The UNOS rules include aspects which indicate under what circumstance compatible transplantation is allowed or not alongside the decisions about where the organ will be allocated. The complicated nature of these interactions of the various levels tends to obscure the role that blood type plays in the

US environment. In most jurisdictions that we are aware of, such as Canada and the Eurotransplant zone, organs are routinely allocated on an ABO-identical basis. Rules for access to ABO-compatible organs on the basis of urgent need vary depending on the organ to be transplanted and the jurisdiction. The numerical examples we present, which considers the experience of one particular Canadian liver transplantation centre, reflects the fact that there are interlinked wait lists. The principal way that this will manifest itself herein is that, in the data used for model calibration, we will aggregate both ABO-identical and ABO-compatible organs that were transplanted into patients of each of the four blood types.

We remark that the model presented herein is one for deceased-donor transplant wait lists. It is true that transplanted kidneys and livers can come from deceased donors or living ones. Widely known in the case of kidneys, this is also true for the liver, one lobe of which can be grafted into the recipient, while the donor retains the other. While living-donor transplantation could be viewed as another type of abandonment from the queue, such abandonments are planned, not random. We do not directly address the issue of such planned transplants in our model.

In summary, the pursuit of a suitable analytical queueing model to infer a likely waiting time for patients awaiting a deceased-donor organ transplantation is both timely and warranted, given the importance of the problem. The present paper is the first we are aware of to address the sickest patient first aspect, and to adequately address the question of abandonments. Such a model would be useful for decision makers to explore qualitatively the impact of various changes that could arise on the waiting times experienced by individual patients, in advance of such changes taking place. An example of potential decision makers, but by far not the only ones, would be teams of medical professionals and performance modellers interested in exploring the impact of the types of changes we list below.

Examples of policy changes that would likely impact waiting time include such aspects as: (i) the merging of formerly separate wait lists, (ii) changes in policies regarding ABO-identical versus ABO-compatible cross-transplantation, or the controlled policies in this

regard in light of Stanford et al. [22], (iii) a decrease in deceased donor organ rates due to improvements in accident prevention, and (iv) in the case of kidneys, a change in demand for deceased donor organs (and possibly the blood mix of the remaining recipients on the list) due to the increased use of "transplant chains" involving living donors. Other like examples can be thought of readily.

The present work presents a model for patients of each ABO blood type, reflecting the factors stated above, that can be used to provide a reasonable indication of the relative likelihood of the possible outcomes that individual patients can experience: successful transplantation prior to perceived health degradation, successful transplantation as a priority patient due to health degradation, the likelihood of abandonment or death while waiting, etc., as well as estimates of the time spent waiting.

The rest of the paper is organized as follows. In the next section, the details of our queueing model are specified. In Section 3, the underlying steady-state distribution for the model is obtained. Delay distributions for high-priority (HP) and low-priority (LP) wait-listed patients are investigated in the subsequent two sections. Section 6 develops estimators for the model parameters using a maximum-likelihood-based approach under a competing risks framework. Section 7 presents the results from an initial case study of real liver transplantation wait-list data: blood-type specific models are fit to this data and model-based performance measures are compared to qualitatively assess goodness of fit. The paper ends with some observations and concluding remarks in Section 8.

2   Description of the Queueing Model

Consider the queueing model depicted in Figure 1 in which a single server provides service to two classes (class 1 and class 2) of transplant requests, each having its own respective line. Wait-listed patients are served on a FCFT basis within their own line. As organ availability is always the limiting factor, the service time constitutes the interval from when a wait-listed patient reaches the head of their queue until an organ becomes available

to him/her. Furthermore, class 1 has preemptive priority over class 2, implying that a class-2 patient in service (i.e., next in line to receive an organ) would be preempted by an arriving class-1 patient to the system. Let $m < \infty$ and $n < \infty$ be the buffer sizes of the LP and HP queues, respectively.
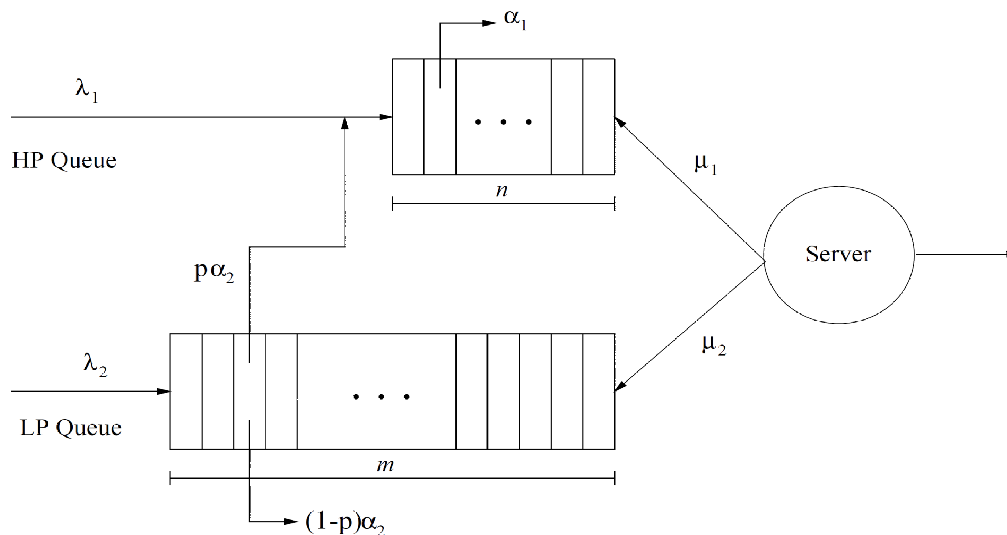

Figure 1: Proposed Queueing Model

We define λ_1 and λ_2 to be the respective (independent) Poisson arrival rates of HP and LP patients to the system. We allow for different service rates for HP and LP patients, both for reasons of generality, and to reflect situations in which patients with an urgent need for transplant have wider access to deceased-donor organs. Let μ_1 and μ_2 be the (individual) independent and exponentially distributed HP and LP service rates, respectively.

We distinguish between the reneging/abandonment behaviours of the two patient classes. At the class-1 level, reneging patients leave the system at rate α_1 and are unrecoverable (reflecting deaths and "coming off-list"). In contrast, at the class-2 level, patients renege at rate α_2, and either leave the system with probability q (for patients who die or come off-list) or are immediately promoted to the end of the HP queue with probability $p = 1 - q$ (reflecting degraded health status still suitable for transplantation). Once in the

8

HP queue, further reneging for such promoted patients occurs at the class-1 rate α_1. Note that a patient who reaches the service stage is still subject to potential reneging (since service constitutes waiting for an organ to become available).

The concept of self-generation of priorities was proposed by Krishnamoorthy and Narayanan [11] and subsequently analyzed in a series of papers by Wang [26], Gómez-Corral et al. [8], and Krishnamoorthy et al. [12, 13, 14, 15], most of which essentially concern a model where regular customers in a multi-server, single class queue would "self-promote" at a constant rate while waiting, independently of other customers in the queue. A self-promoted customer would displace any non-priority item in service. For the foregoing models, in the event that no regular customers were found in service, the promoted customer leaves the queue to obtain urgent service elsewhere. In contrast, our self-promoted customers join the HP queue and only immediately depart the system if no waiting space is available in the HP queue.

In this paper, we determine the following main performance measures associated with our queueing model:

(i) The steady-state joint queue length probability $\pi_{i,j} = Pr(X_L = i, X_H = j)$ for the number of HP patients, X_H, and LP patients, X_L, present in the system, as well as the respective LP and HP blocking probabilities given by

$$LPBlock = \sum_{j=0}^{n} \pi_{m,j}$$

and

$$HPBlock = \sum_{i=0}^{m} \pi_{i,n} \, .$$

The joint queue length distribution $\{\pi_{i,j}; \ i = 0, 1, \ldots, m, \ j = 0, 1, \ldots, n\}$ is obtained in Section 3.

(ii) The cumulative distribution function (CDF) and moments related to the stationary waiting time W_H^* of an originally-arriving HP patient who successfully completes service prior to reneging, obtained in Section 4.

(iii) The CDF and moments related to the stationary waiting time $W_{L,1}^*$ of an arriving LP patient who successfully completes service within the LP queue prior to reneging, obtained in Section 5.

(iv) The CDF and moments related to the stationary waiting time $W_{L,2}^*$ of a promoted LP patient from the point of transfer to the HP queue until that patient successfully completes service prior to reneging, obtained also in Section 5.

(v) The probabilities $HPRenege$, $LPRenege$, and $PromLPRenege$ representing the various likelihoods that a patient reneges prior to receipt of an organ. Specifically, $HPRenege$ is the probability that a patient who arrives as HP will renege, $LPRenege$ is the reneging probability for a LP patient, and $PromLPRenege$ is the reneging probability for a HP patient who initially arrived as LP. These probabilities are determined in Sections 4 and 5.

3 Determination of the Steady-state Probabilities

Our first objective is to determine the steady-state joint probabilities $\{\pi_{i,j}; \ i = 0, 1, \ldots, m, \ j = 0, 1, \ldots, n\}$, where $\pi_{i,j} = Pr(X_L = i, X_H = j)$. We say that the process is "at level i" whenever $X_L = i$. For $k \geq 0$, we define the kth steady-state probability row vector (of dimension $n + 1$) to be $\underline{\pi}_k = (\pi_{k,0}, \pi_{k,1}, \ldots, \pi_{k,n})$. Let $\underline{\pi} = (\underline{\pi}_0, \underline{\pi}_1, \ldots, \underline{\pi}_m)$ be the concatenated steady-state probability row vector having a total of $m + 1$ levels. To determine $\underline{\pi}$, we need to solve $\underline{\tilde{0}} = \underline{\pi} Q$ where Q is the infinitesimal generator of the process and $\underline{\tilde{0}} = (\underline{0}, \underline{0}, \ldots, \underline{0})$ is the concatenated row vector (having a total of $m+1$ levels) in which $\underline{0}$ denotes a $1 \times (n + 1)$ row vector of zeros.

With X_L serving as the level of the process (and X_H as the sub-level), we note that Q is block-structured with blocks $Q_{i,j}$ (of size $n + 1$) containing all transitions where X_L changes from i to j. Due to the presence of reneging in the model and the fact that X_L can only change by ± 1, we end up with a level-dependent quasi-birth-and-death (QBD)

process having infinitesimal generator of the form

$$
Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ m-2 \\ m-1 \\ m \end{array}
\begin{array}{ccccccc}
0 & 1 & 2 & \cdots & m-2 & m-1 & m \\
\end{array}
\left(\begin{array}{ccccccc}
Q_{0,0} & Q_{0,1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
Q_{1,0} & Q_{1,1} & Q_{1,2} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & Q_{2,1} & Q_{2,2} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & Q_{m-2,m-2} & Q_{m-2,m-1} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & Q_{m-1,m-2} & Q_{m-1,m-1} & Q_{m-1,m} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & Q_{m,m-1} & Q_{m,m}
\end{array}\right). \tag{1}
$$

In (1) above, $\mathbf{0}$ denotes an appropriately dimensioned square matrix of zeros (which, in this case, is of dimension $n+1$). The overall dimension of Q is $(m+1)(n+1) \times (m+1)(n+1)$.

Note that $Q_{0,1} = Q_{1,2} = \cdots = Q_{m-1,m} = \lambda_2 I_{n+1}$ where I_k, in general, denotes the $k \times k$ identity matrix. Moreover, it readily follows that

$$
Q_{i,i-1} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ n-1 \\ n \end{array}
\begin{array}{cccccc}
0 & 1 & 2 & \cdots & n-1 & n \\
\end{array}
\left(\begin{array}{cccccc}
\mu_2 + iq\alpha_2 & ip\alpha_2 & 0 & \cdots & 0 & 0 \\
0 & iq\alpha_2 & ip\alpha_2 & \ddots & 0 & 0 \\
0 & 0 & iq\alpha_2 & \ddots & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & iq\alpha_2 & ip\alpha_2 \\
0 & 0 & 0 & \cdots & 0 & i\alpha_2
\end{array}\right), \quad i = 1, 2, \ldots, m.
$$

For notational convenience, we define $\lambda = \lambda_1 + \lambda_2$, $\beta_i = \mu_1 + i\alpha_1$, and $\gamma_i = \mu_2 + i\alpha_2$. Also, let $\underline{e}_k$ be a $1 \times k$ row vector with 1 as the first entry and zeros everywhere else. Based on

this notation, the diagonal components of Q can be expressed as

$$
Q_{0,0} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ n-1 \\ n \end{array}
\begin{array}{cccccc}
0 & 1 & 2 & \cdots & n-1 & n \\
\left(\begin{array}{cccccc}
-\lambda & \lambda_1 & 0 & \cdots & 0 & 0 \\
\beta_1 & -(\lambda+\beta_1) & \lambda_1 & \ddots & 0 & 0 \\
0 & \beta_2 & -(\lambda+\beta_2) & \ddots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -(\lambda+\beta_{n-1}) & \lambda_1 \\
0 & 0 & 0 & \cdots & \beta_n & -(\lambda_2+\beta_n)
\end{array}\right)
\end{array},
$$

$$
Q_{m,m} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ n-1 \\ n \end{array}
\begin{array}{cccccc}
0 & 1 & 2 & \cdots & n-1 & n \\
\left(\begin{array}{cccccc}
-(\lambda_1+\gamma_m) & \lambda_1 & 0 & \cdots & 0 & 0 \\
\beta_1 & -(\lambda_1+\beta_1+m\alpha_2) & \lambda_1 & \ddots & 0 & 0 \\
0 & \beta_2 & -(\lambda_1+\beta_2+m\alpha_2) & \ddots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -(\lambda_1+\beta_{n-1}+m\alpha_2) & \lambda_1 \\
0 & 0 & 0 & \cdots & \beta_n & -(\beta_n+m\alpha_2)
\end{array}\right)
\end{array},
$$

and

$$
Q_{i,i} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ n-1 \\ n \end{array}
\begin{array}{cccccc}
0 & 1 & 2 & \cdots & n-1 & n \\
\left(\begin{array}{cccccc}
-(\lambda+\gamma_i) & \lambda_1 & 0 & \cdots & 0 & 0 \\
\beta_1 & -(\lambda+\beta_1+i\alpha_2) & \lambda_1 & \ddots & 0 & 0 \\
0 & \beta_2 & -(\lambda+\beta_2+i\alpha_2) & \ddots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -(\lambda+\beta_{n-1}+i\alpha_2) & \lambda_1 \\
0 & 0 & 0 & \cdots & \beta_n & -(\lambda_2+\beta_n+i\alpha_2)
\end{array}\right)
\end{array}
$$

for $i = 1, 2, \ldots, m-1$.

Level-dependent QBDs have been well-studied in the literature (e.g., see Bright and Taylor [4]), and it is possible to adapt a computational procedure proposed by Gaver et

al. [6] to calculate the steady-state probabilities associated with our model, which we now summarize below. The equilibrium equations in block form are as follows:

$$\underline{0} = \underline{\pi}_0 Q_{0,0} + \underline{\pi}_1 Q_{1,0}, \tag{2}$$

$$\underline{0} = \lambda_2 \underline{\pi}_{i-1} + \underline{\pi}_i Q_{i,i} + \underline{\pi}_{i+1} Q_{i+1,i}, \ \ i = 1, 2, \ldots, m-1, \tag{3}$$

$$\underline{0} = \lambda_2 \underline{\pi}_{m-1} + \underline{\pi}_m Q_{m,m}. \tag{4}$$

Solving (3) and (4) in a backward fashion ultimately yields

$$\underline{\pi}_i = \underline{\pi}_0 \prod_{j=1}^{i} R_j, \ \ i = 1, 2, \ldots, m, \tag{5}$$

where the set of matrices $\{R_j\}_{j=1}^{m}$ satisfy the recursive relation

$$R_j = -\lambda_2 (Q_{j,j} + R_{j+1} Q_{j+1,j})^{-1}, \ \ j = 1, 2, \ldots, m-1,$$

with

$$R_m = -\lambda_2 Q_{m,m}^{-1}.$$

Defining $R_0 = Q_{0,0} + R_1 Q_{1,0}$, (2) becomes

$$\underline{\pi}_0 R_0 = \underline{0}. \tag{6}$$

The summation of all the probabilities can be expressed as

$$\underline{\pi}_0 \underline{1}'_{n+1} + \underline{\pi}_0 R_1 \underline{1}'_{n+1} + \underline{\pi}_0 R_1 R_2 \underline{1}'_{n+1} + \cdots + \underline{\pi}_0 R_1 R_2 \cdots R_m \underline{1}'_{n+1} = 1, \tag{7}$$

where $\underline{1}'_k$, in general, denotes a $k \times 1$ column vector of ones. Factoring out $\underline{\pi}_0$ from (7) and defining the column vector

$$\underline{u}' = \left(I_{n+1} + \sum_{i=1}^{m} \prod_{j=1}^{i} R_j \right) \underline{1}'_{n+1},$$

(6) and (7) give rise to the following system of linear equations which must be solved to determine $\underline{\pi}_0$, namely:

$$\underline{\pi}_0 [R_0, \underline{u}'] = (\underline{0}, 1). \tag{8}$$

In (8) above, $[R_0, \underline{u}']$ and $(\underline{0}, 1)$ represent concatenated matrices of sizes $(n+1) \times (n+2)$ and $1 \times (n+2)$, respectively. Once $\underline{\pi}_0$ is determined, we obtain $\underline{\pi}_i$, $i \geq 1$, via (5).

13

4   Delay Distributions for HP Transplant Requests

We focus initially on the stationary distribution of a random variable W_H, representing the duration of time from the arrival of an external HP patient to the system until this patient successfully completes service. We refer to W_H as the "nominal" HP patient delay, as we implicitly assume that there is room in the HP queue for the arriving patient to enter the system, and that the arriving HP patient is not subject to reneging. (Later on, we will incorporate the reneging behaviour of this patient in our analysis.)

Define the following conditional steady-state probabilities, given that a random HP arrival is able to enter the HP queue:

$$\theta_{i,j} = \frac{\pi_{i,j}}{1 - HPBlock}, \ i = 0, 1, \ldots, m, \ j = 0, 1, \ldots, n - 1.$$

Let $F_H(\omega)$, $\omega \geq 0$, denote the CDF of W_H. Assuming successful entry into the HP queue, the PASTA property (e.g., see Tijms [25], Theorem 2.4.1) ensures that our tagged Poisson-arriving HP patient finds the system in state (i, j) with probability $\theta_{i,j}$. Therefore, conditioning on the state of the system encountered upon entry, we obtain

$$F_H(\omega) = \sum_{i=0}^{m} \sum_{j=0}^{n-1} \theta_{i,j} \Pr(W_H \leq \omega \mid X_L = i, X_H = j). \tag{9}$$

Due to the nature of the preemptive service rule in place, the quantity $\Pr(W_H \leq \omega \mid X_L = i, X_H = j)$ does not depend upon the number of LP transplant requests present upon arrival of our tagged HP patient, so that (9) reduces to

$$F_H(\omega) = \sum_{j=0}^{n-1} \theta_{\bullet,j} \Pr(W_H \leq \omega \mid X_H = j), \tag{10}$$

where $\theta_{\bullet,j} = \sum_{i=0}^{m} \theta_{i,j}$ denotes the marginal distribution of the HP queue length.

With $X_H = j$, we observe that W_H initially consists of the total time required to clear the j class-1 patients ahead of the newly-arriving HP patient. Since all class-1 patients ahead of this HP patient might renege (including the one receiving service at rate μ_1), the total delay (including the service time of our tagged HP patient) can be represented by

$\sum_{k=0}^{j} Y_k^H$, a sum of $j+1$ independent random variables in which each Y_k^H is exponentially distributed at rate $\beta_k = \mu_1 + k\alpha_1$. The resulting distribution is a member of the phase-type family of distributions (e.g., see Latouche and Ramaswami [16], Chapter 2), so that

$$\Pr(W_H \leq \omega \mid X_H = j) = 1 - \underline{e}_{j+1} \exp\{T_{j+1}\omega\}\underline{1}'_{j+1}, \tag{11}$$

where T_{j+1} is a $(j+1) \times (j+1)$ matrix of the form

$$T_{j+1} = \begin{bmatrix} -\beta_j & \beta_j & 0 & \cdots & 0 & 0 \\ 0 & -\beta_{j-1} & \beta_{j-1} & \ddots & 0 & 0 \\ 0 & 0 & -\beta_{j-2} & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\beta_1 & \beta_1 \\ 0 & 0 & 0 & \cdots & 0 & -\beta_0 \end{bmatrix}. \tag{12}$$

Substituting (11) into (10), it is possible to express the distribution of W_H in terms of a single, succinct phase-type representation of dimension n, namely

$$F_H(\omega) = 1 - \underline{\Theta}_n \exp\{T_n\omega\}\underline{1}'_n,$$

where $\underline{\Theta}_n$ is the $1 \times n$ row vector given by $\underline{\Theta}_n = (\theta_{\bullet,n-1}, \theta_{\bullet,n-2}, \ldots, \theta_{\bullet,1}, \theta_{\bullet,0})$.

We now incorporate the reneging behaviour that our tagged HP patient can exhibit while residing in the system. Let W_H^* denote the "actual" HP patient delay, representing the arriving HP patient's total time spent in system (which incorporates successfully completing service prior to reneging). For $\omega \geq 0$, $G_H(\omega) = \Pr(W_H^* \leq \omega) = \Pr(W_H \leq \omega \mid W_H \leq R_H)$ where R_H denotes an exponentially distributed random variable, independent of W_H, with mean $1/\alpha_1$. Making use of fundamental matrix algebraic techniques, the following expressions for $G_H(\omega)$ and the moments of W_H^* are ultimately derived:

$$\begin{aligned} G_H(\omega) &= 1 - \Pr(W_H > \omega \mid W_H \leq R_H) \\ &= 1 - \frac{\Pr(\omega < W_H \leq R_H)}{\Pr(W_H \leq R_H)} \end{aligned}$$

15

$$= 1 - \frac{\int_\omega^\infty \Pr(W_H > \omega)\alpha_1 e^{-\alpha_1 x}dx - \int_\omega^\infty \Pr(W_H > x)\alpha_1 e^{-\alpha_1 x}dx}{1 - \int_0^\infty \Pr(W_H > x)\alpha_1 e^{-\alpha_1 x}dx}$$

$$= 1 - \frac{\underline{\Theta}_n \exp\{T_n\omega\}[I_n - \alpha_1(\alpha_1 I_n - T_n)^{-1}]\underline{1}'_n e^{-\alpha_1 \omega}}{1 - \alpha_1 \underline{\Theta}_n(\alpha_1 I_n - T_n)^{-1}\underline{1}'_n} \tag{13}$$

and

$$E(W_H^{*r}) = \frac{r!\underline{\Theta}_n[I_n - \alpha_1(\alpha_1 I_n - T_n)^{-1}](\alpha_1 I_n - T_n)^{-r}\underline{1}'_n}{1 - \alpha_1 \underline{\Theta}_n(\alpha_1 I_n - T_n)^{-1}\underline{1}'_n}, \; r = 1, 2, \dots \; . \tag{14}$$

Note that the denominators of (13) and (14) involve a quantity which we denote by $HPRenege = \alpha_1 \underline{\Theta}_n(\alpha_1 I_n - T_n)^{-1}\underline{1}'_n$, representing the probability our tagged HP patient abandons the HP queue prior to an organ becoming available to him/her.

5 Delay Distributions for LP Transplant Requests

We next determine stationary delay distributions associated with the arrival of an arbitrary LP patient to the system. First of all, we focus our attention on a random variable $W_{L,1}$, which we define as the duration of time from the arrival of a LP patient to the system until this patient successfully completes service within the LP queue. Therefore, as in the definition of W_H from the previous section, we refer to $W_{L,1}$ as the nominal LP patient delay and likewise assume that there is room in the LP queue for the arriving patient to enter the system, and that the arriving LP patient is not subject to reneging. (We will incorporate the reneging behaviour of this patient in our analysis later.)

We begin by defining similar conditional steady-state probabilities, namely

$$\phi_{i,j} = \frac{\pi_{i,j}}{1 - LPBlock}, \; i = 0, 1, \dots, m-1, \; j = 0, 1, \dots, n.$$

We next introduce several row vectors required in the subsequent analysis. First of all, we define $\underline{\delta}_i$ to be a $1 \times (n+1)$ row vector with 1 in position i, $i = 1, 2, \dots, n+1$, and zeros everywhere else. Next, let $\underline{\phi}_i = (\phi_{i,n}, \phi_{i,n-1}, \phi_{i,n-2}, \dots, \phi_{i,0})$, $i = 0, 1, \dots, m-1$. Furthermore, let $\underline{\tilde{\phi}}_i = (\underline{\phi}_i, \underline{0}, \underline{0}, \dots, \underline{0})$ be the concatenated row vector having a total of $m - i$ levels. $\underline{\tilde{\phi}}_i$ has an overall dimension of $(m-i)(n+1)$, $i = 0, 1, \dots, m-1$. If we now

define $\underline{\Phi} = (\underline{\tilde{\phi}}_{m-1}, \underline{\tilde{\phi}}_{m-2}, \underline{\tilde{\phi}}_{m-3}, \ldots, \underline{\tilde{\phi}}_0)$ to be the concatenated row vector of dimension $\ell = \sum_{i=0}^{m-1}(m-i)(n+1) = m(m+1)(n+1)/2$, then by construction $\underline{\Phi}\,\underline{1}'_\ell = 1$.

Upon entry to the system, our tagged Poisson-arriving LP patient must not only wait for all LP patients in front of it to clear, but for **all** HP patients, including those present upon arrival as well as those arriving later to be cleared from the system. (This potentially includes promoted LP patients who queued behind the tagged LP patient.) As a result, $W_{L,1}$ can be modelled as the time to absorption in a Markov chain with infinitesimal generator of the form

$$\mathcal{P} = \begin{bmatrix} \mathcal{R} & -\mathcal{R}\underline{1}'_\ell \\ \underline{\hat{0}}_\ell & 0 \end{bmatrix},$$

where $\underline{\hat{0}}_r$, in general, denotes a $1 \times r$ row vector of zeros and the rate matrix $\mathcal{R}$ corresponds to the following state space partitioning:

(i) Level i – the number i of LP patients in front of our tagged LP patient, $i = 0, 1, \ldots, m-1$;

(ii) Phase j – the number j of LP patients who queue behind our tagged LP patient, $j = 0, 1, \ldots, m-i-1$;

(iii) Sub-phase k – the number k of HP patients present in the system, $k = 0, 1, \ldots, n$.

Based on the permissible values that i, j, and k can assume, one can readily verify that the cardinality of this state space equals ℓ, the total number of elements comprising the row vector $\underline{\Phi}$.

In regard to this state space partitioning, the form of the matrix $\mathcal{R}$ is given by

$$
\mathcal{R} =
\begin{array}{c}
 \\
m-1 \\
m-2 \\
m-3 \\
\vdots \\
1 \\
0
\end{array}
\begin{array}{c}
\begin{array}{cccccc}
m-1 & m-2 & m-3 & \cdots & 1 & 0
\end{array} \\
\left(
\begin{array}{cccccc}
\mathcal{A}_{m-1} & \mathcal{B}_{m-1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathcal{A}_{m-2} & \mathcal{B}_{m-2} & \ddots & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathcal{A}_{m-3} & \ddots & \mathbf{0} & \mathbf{0} \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathcal{A}_1 & \mathcal{B}_1 \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathcal{A}_0
\end{array}
\right)
\end{array}.
\tag{15}
$$

In (15) above, the diagonal blocks $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_{m-1}$, corresponding to state transitions in which the number of LP patients in front of our tagged LP patient does not change, are such that $\mathcal{A}_i$ is a square matrix of block dimension $m-i$ in which

$$
\mathcal{A}_i =
\begin{array}{c}
 \\
0 \\
1 \\
2 \\
\vdots \\
m-i-2 \\
m-i-1
\end{array}
\begin{array}{c}
\begin{array}{cccccc}
0 & 1 & 2 & \cdots & m-i-2 & m-i-1
\end{array} \\
\left(
\begin{array}{cccccc}
\mathcal{F}_0^{(i)} & \lambda_2 I_{n+1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\
\alpha_2 \mathcal{D} & \mathcal{F}_1^{(i)} & \lambda_2 I_{n+1} & \ddots & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & 2\alpha_2 \mathcal{D} & \mathcal{F}_2^{(i)} & \ddots & \mathbf{0} & \mathbf{0} \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathcal{F}_{m-i-2}^{(i)} & \lambda_2 I_{n+1} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & (m-i-1)\alpha_2 \mathcal{D} & \mathcal{F}_{m-i-1}^{(i)}
\end{array}
\right)
\end{array},
$$

where $\mathcal{D}$, the $(n+1) \times (n+1)$ matrix governing the state transitions in which the number of queued LP patients behind our tagged LP patient decreases by one, is given by

$$
\mathcal{D} =
\begin{array}{c}
 \\
n \\
n-1 \\
n-2 \\
\vdots \\
1 \\
0
\end{array}
\begin{array}{c}
\begin{array}{cccccc}
n & n-1 & n-2 & \cdots & 1 & 0
\end{array} \\
\left(
\begin{array}{cccccc}
1 & 0 & 0 & \cdots & 0 & 0 \\
p & q & 0 & \ddots & 0 & 0 \\
0 & p & q & \ddots & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & q & 0 \\
0 & 0 & 0 & \cdots & p & q
\end{array}
\right)
\end{array}.
$$

18

The off-diagonal blocks $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{m-1}$ in (15), corresponding to state transitions which reduce the LP queue in front of our tagged LP patient by one, are such that $\mathcal{B}_i$ is a $(m-i) \times (m-i+1)$ block matrix given by

$$
\mathcal{B}_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ m-i-1 \end{array}
\begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & m-i-1 & m-i \end{array} \\
\left(\begin{array}{cccccc}
\mathcal{C}_i & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathcal{C}_i & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathcal{C}_i & \ddots & \mathbf{0} & \mathbf{0} \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathcal{C}_i & \mathbf{0}
\end{array} \right),
\end{array}
$$

where $\mathcal{C}_i = \mu_2(\underline{\delta}'_{n+1} \otimes \underline{\delta}_{n+1}) + i\alpha_2 \mathcal{D}$ and "$\otimes$" denotes the Kronecker product operator. Finally, to obtain the diagonal blocks of $\mathcal{A}_i$ (i.e., the rate matrices $\mathcal{F}_0^{(i)}, \mathcal{F}_1^{(i)}, \ldots, \mathcal{F}_{m-i-1}^{(i)}$ in which the only state transitions occur within the HP queue) for $i = 0, 1, \ldots, m-1$, we make use of our earlier assumption that the tagged LP patient does not renege, and the fact that the row sums of $\mathcal{P}$ must be equal to 0. This enables us to determine

$$
\mathcal{F}_{m-i-1}^{(i)} = \begin{array}{c} \\ n \\ n-1 \\ n-2 \\ \vdots \\ 1 \\ 0 \end{array}
\begin{array}{c} \begin{array}{cccccc} n & n-1 & n-2 & \cdots & 1 & 0 \end{array} \\
\left(\begin{array}{cccccc}
-[\beta_n + (m-1)\alpha_2] & \beta_n & 0 & \cdots & 0 & 0 \\
\lambda_1 & -[\lambda_1 + \beta_{n-1} + (m-1)\alpha_2] & \beta_{n-1} & \ddots & 0 & 0 \\
0 & \lambda_1 & -[\lambda_1 + \beta_{n-2} + (m-1)\alpha_2] & \ddots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -[\lambda_1 + \beta_1 + (m-1)\alpha_2] & \beta_1 \\
0 & 0 & 0 & \cdots & \lambda_1 & -(\lambda_1 + \gamma_{m-1})
\end{array} \right)
\end{array}
$$

and

$$
\mathcal{F}_j^{(i)} = \begin{array}{c} \\ n \\ n-1 \\ n-2 \\ \vdots \\ 1 \\ 0 \end{array}
\begin{array}{c} \begin{array}{cccccc} n & n-1 & n-2 & \cdots & 1 & 0 \end{array} \\
\left(\begin{array}{cccccc}
-[\lambda_2 + \beta_n + (i+j)\alpha_2] & \beta_n & 0 & \cdots & 0 & 0 \\
\lambda_1 & -[\lambda + \beta_{n-1} + (i+j)\alpha_2] & \beta_{n-1} & \ddots & 0 & 0 \\
0 & \lambda_1 & -[\lambda + \beta_{n-2} + (i+j)\alpha_2] & \ddots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -[\lambda + \beta_1 + (i+j)\alpha_2] & \beta_1 \\
0 & 0 & 0 & \cdots & \lambda_1 & -(\lambda + \gamma_{i+j})
\end{array} \right)
\end{array}
$$

19

for $j = 0, 1, \ldots, m - i - 2$.

With these pieces in place, the time to absorption in such a Markov chain is once again phase-type distributed, and so the CDF of $W_{L,1}$ is given by

$$F_{L,1}(\omega) = 1 - \underline{\Phi} \exp\{\mathcal{R}\omega\} \underline{1}'_\ell .$$

However, we now incorporate the reneging behaviour of the tagged LP patient by defining $W^*_{L,1}$ to be the actual LP delay (i.e., the arriving LP patient's total time spent in system to achieve successful service within the LP queue). Clearly, $G_{L,1}(\omega) = \Pr(W^*_{L,1} \leq \omega) = \Pr(W_{L,1} \leq \omega | W_{L,1} \leq R_L)$ where R_L denotes an exponentially distributed random variable, independent of $W_{L,1}$, with mean $1/\alpha_2$. Following the same approach which led to the derivation of (13) and (14), we ultimately obtain

$$G_{L,1}(\omega) = 1 - \frac{\underline{\Phi} \left[I_\ell - \alpha_2(\alpha_2 I_\ell - \mathcal{R})^{-1} \right] \exp\{\mathcal{R}\omega\} \underline{1}'_\ell \, e^{-\alpha_2 \omega}}{1 - \alpha_2 \underline{\Phi}(\alpha_2 I_\ell - \mathcal{R})^{-1} \underline{1}'_\ell} \tag{16}$$

as well as

$$E(W^{*r}_{L,1}) = \frac{r! \underline{\Phi} \left[I_\ell - \alpha_2(\alpha_2 I_\ell - \mathcal{R})^{-1} \right] (\alpha_2 I_\ell - \mathcal{R})^{-r} \underline{1}'_\ell}{1 - \alpha_2 \underline{\Phi}(\alpha_2 I_\ell - \mathcal{R})^{-1} \underline{1}'_\ell}, \quad r = 1, 2, \ldots . \tag{17}$$

We remark that the denominators of (16) and (17) involve the LP reneging probability $LPRenege = \alpha_2 \underline{\Phi}(\alpha_2 I_\ell - \mathcal{R})^{-1} \underline{1}'_\ell$.

We next turn our attention to characterizing the distribution of $W_{L,2}$, which we define as the nominal delay experienced from the point of transfer to the HP queue of a promoted LP patient until that patient successfully completes service. We remark that this distribution will look different than that which is perceived by the externally-arriving class-1 stream, as the instants at which promotions occur is clearly dependent upon the present length of the LP queue. However, conditional on encountering j class-1 patients ahead of the promoted LP patient, the results of the previous section can still be applied to ascertain that the nominal delay the promoted patient experiences in order to complete service is phase-type distributed with rate matrix given by (12). The difference will lie in the initial probability vector to use in connection with this phase-type distribution, as $\underline{\Theta}_n$ from the previous section is no longer appropriate.

20

As a means of identifying the proper initial probability vector to use, we track the potential path our tagged LP patient can take with respect to three distinct **final** outcomes (or dispositions) within the LP queue. We label these final outcomes as follows:

$F_1 \equiv$ tagged LP patient completes successful service in the LP queue,

$F_2 \equiv$ tagged LP patient reneges and exits the system from the LP queue, and

$F_3 \equiv$ tagged LP patient receives promotion to the HP queue.

The event F_3 can be further broken down into elemental final outcomes $F_3 = \{f_n, f_{n-1}, \ldots, f_0\}$, where f_i represents the outcome in which our tagged LP patient finds i class-1 patients ahead of it upon promotion. Note that f_n represents the outcome that the HP queue is full at the instant of promotion. As a matter of completeness, we include f_n in the set F_3 (although its associated probability will be singled out later on as a blocking probability of interest). Let $\Delta_F = \{F_1, F_2, f_n, f_{n-1}, \ldots, f_0\}$ be the set of states corresponding to these final outcomes. Also, let Δ_L represent the set of states corresponding to our tagged LP patient residing in the LP queue, having the same state space partitioning as the rate matrix $\mathcal{R}$ above. In order to track which of these final outcomes the tagged LP patient ultimately experiences, we introduce the following Markov chain with underlying infinitesimal generator

$$
\mathcal{Q} = \begin{array}{c} \\ \Delta_L \\ \Delta_F \end{array}
\begin{pmatrix}
\overset{\Delta_L}{\mathcal{S}} & \overset{\Delta_F}{\mathcal{T}} \\
\mathbf{0} & \mathbf{0}
\end{pmatrix},
$$

where $\mathcal{S} = \mathcal{R} - \alpha_2 I_\ell$. Recall that the matrix $\mathcal{R}$ given by (15) did not take into account the reneging behaviour of our tagged LP patient. However, our current treatment necessitates the incorporation of this reneging behaviour, and as such, the definition of $\mathcal{S}$ includes the

additional term $-\alpha_2 I_\ell$. Furthermore, the block-structured matrix $\mathcal{T}$ is given by

$$
\mathcal{T} =
\begin{array}{c}
\\
m-1\\
m-2\\
m-3\\
\vdots\\
1\\
0
\end{array}
\begin{pmatrix}
F_1 & F_2 & f_n & f_{n-1} & \cdots & f_0\\
\underline{\hat{0}}'_{n+1} & q\alpha_2\underline{1}'_{n+1} & p\alpha_2(\underline{1}'_1 \otimes \underline{\delta}'_1) & p\alpha_2(\underline{1}'_1 \otimes \underline{\delta}'_2) & \cdots & p\alpha_2(\underline{1}'_1 \otimes \underline{\delta}'_{n+1})\\
\underline{\hat{0}}'_{2(n+1)} & q\alpha_2\underline{1}'_{2(n+1)} & p\alpha_2(\underline{1}'_2 \otimes \underline{\delta}'_1) & p\alpha_2(\underline{1}'_2 \otimes \underline{\delta}'_2) & \cdots & p\alpha_2(\underline{1}'_2 \otimes \underline{\delta}'_{n+1})\\
\underline{\hat{0}}'_{3(n+1)} & q\alpha_2\underline{1}'_{3(n+1)} & p\alpha_2(\underline{1}'_3 \otimes \underline{\delta}'_1) & p\alpha_2(\underline{1}'_3 \otimes \underline{\delta}'_2) & \cdots & p\alpha_2(\underline{1}'_3 \otimes \underline{\delta}'_{n+1})\\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
\underline{\hat{0}}'_{(m-1)(n+1)} & q\alpha_2\underline{1}'_{(m-1)(n+1)} & p\alpha_2(\underline{1}'_{m-1} \otimes \underline{\delta}'_1) & p\alpha_2(\underline{1}'_{m-1} \otimes \underline{\delta}'_2) & \cdots & p\alpha_2(\underline{1}'_{m-1} \otimes \underline{\delta}'_{n+1})\\
\mu_2(\underline{1}'_m \otimes \underline{\delta}'_{n+1}) & q\alpha_2\underline{1}'_{m(n+1)} & p\alpha_2(\underline{1}'_m \otimes \underline{\delta}'_1) & p\alpha_2(\underline{1}'_m \otimes \underline{\delta}'_2) & \cdots & p\alpha_2(\underline{1}'_m \otimes \underline{\delta}'_{n+1})
\end{pmatrix}.
$$

We wish to determine the absorption probabilities into each of the $n+1$ states of F_3, which we denote by $\psi(f_i)$, $i = 0, 1, \ldots, n$. This is accomplished by analyzing the embedded Markov chain (or jump process) associated with the above continuous-time Markov chain. Specifically, the (one-step) transition probability matrix of the associated jump process is given by (e.g., see Syski [24], p. 14)

$$
P =
\begin{array}{c}
\\
\Delta_L\\
\Delta_F
\end{array}
\begin{pmatrix}
\Delta_L & \Delta_F\\
I_\ell - \text{diag}(\mathcal{S})^{-1}\mathcal{S} & -\text{diag}(\mathcal{S})^{-1}\mathcal{T}\\
\mathbf{0} & I_{n+3}
\end{pmatrix},
\tag{18}
$$

where $\text{diag}(\mathcal{S})$ denotes the matrix containing only $\text{diag}(\mathcal{S})$'s diagonal entries. The fundamental matrix (e.g., see Pinsky and Karlin [20], p. 142) associated with (18) works out to be

$$
\left[I_\ell - \left(I_\ell - \text{diag}(\mathcal{S})^{-1}\mathcal{S}\right)\right]^{-1} = \mathcal{S}^{-1}\text{diag}(\mathcal{S}),
$$

from which it readily follows that

$$
\psi(f_i) = \underline{\Phi}\,\mathcal{S}^{-1}\text{diag}(\mathcal{S})\left(-\text{diag}(\mathcal{S})^{-1}\mathcal{T}'_{n+3-i}\right) = \underline{\Phi}\left(\alpha_2 I_\ell - \mathcal{R}\right)^{-1}\mathcal{T}'_{n+3-i},
$$

where $\mathcal{T}'_{n+3-i}$ denotes the $(n+3-i)^{th}$ column of $\mathcal{T}$, $i = 0, 1, \ldots, n$. Of particular relevance is the quantity $\psi(f_n)$, which represents the probability that the tagged LP patient encounters a full HP queue at the instant of its promotion (and is thus denied entry and subsequently

exits the system). In other words, $\psi(f_n)$ is another blocking probability of interest. In keeping with our earlier choice of notation, we also let $PromLPBlock = \psi(f_n)$.

Clearly, $\sum_{i=0}^{n} \psi(f_i) = p \cdot LPRenege$, and so the $1 \times n$ row vector

$$\underline{\Psi}_n = \frac{1}{p \cdot LPRenege - \psi(f_n)} \Big(\psi(f_{n-1}), \psi(f_{n-2}), \ldots, \psi(f_0) \Big)$$

contains the (normalized) probabilities for the number of potential HP patients present in the HP queue at the instant that the LP patient is successfully promoted to the HP queue. With this analogue of $\underline{\Theta}_n$ in hand, we make use of the phase-type distributional results of the previous section to immediately obtain

$$F_{L,2}(\omega) = \Pr(W_{L,2} \leq \omega) = 1 - \underline{\Psi}_n \exp\{T_n \omega\} \underline{1}'_n.$$

In addition, the associated delay distribution defined by the conditional random variable $W^*_{L,2} = W_{L,2} \,|\, (W_{L,2} \leq R_H)$ has CDF

$$G_{L,2}(\omega) = 1 - \frac{\underline{\Psi}_n \exp\{T_n \omega\}[I_n - \alpha_1 (\alpha_1 I_n - T_n)^{-1}] \underline{1}'_n e^{-\alpha_1 \omega}}{1 - PromLPRenege}$$

and moments of the form

$$E(W^{*r}_{L,2}) = \frac{r! \underline{\Psi}_n \big[I_n - \alpha_1 (\alpha_1 I_n - T_n)^{-1} \big] (\alpha_1 I_n - T_n)^{-r} \underline{1}'_n}{1 - PromLPRenege} \,, \ r = 1, 2, \ldots \,,$$

where $PromLPRenege = \alpha_1 \, \underline{\Psi}_n (\alpha_1 I_n - T_n)^{-1} \underline{1}'_n$ represents the probability that a self-promoted LP patient leaves the system prior to an organ becoming available to him/her.

6 Model Calibration Via a Parametric Competing Risks Framework

Solid organ transplantation wait-list data differs from the usual queueing perspective in several aspects. A patient's arrival time, waiting time, and service time are still tracked, but there are additional complications due to both abandonments (e.g., deaths and coming off-list) and self-promotion in priority due to health degradation. In this context, the cause

23

for a patient to leave the queue needs to be tracked as well as their time in the system; such data can be modelled using a "competing risks" framework (e.g., see Kalbfleisch and Prentice [10] or Lawless [17]).

To be specific, we are interested in simultaneously modelling more than one event time variable. Consider the arrival of a LP patient, who can exit the LP queue either by obtaining a transplanted organ, or by reneging (i.e., promotion to the HP queue, death, or recovery to off-list status). If such a patient left the system as a result of obtaining a transplanted organ, we know not only the exact time of this event, but we also know that their time to renege would have been longer (i.e., their reneging time is right-censored). Conversely, had the patient reneged, then their time to transplant would have been right-censored.

We have assumed that external arrivals to both queues follow independent Poisson processes and that the transplant and reneging times are exponentially distributed. In this parametric modelling context, it is possible to construct a likelihood function that incorporates the notion of a competition between the class-specific transplantation and reneging rates. Given data, the model can then be calibrated by estimating the parameters using a maximum-likelihood-based approach. The suitability of assuming exponential inter-availability times was discussed at length in Stanford et al. [21], and that result was employed in [22] to show why it is reasonable to make that assumption in the presence of random ABO-compatible transplantation.

Consider the case of two generic competing risks represented by the event time random variables T_1 and T_2. When the observed times $\{t_i;\ i \in \mathbb{Z}^+\}$ can be viewed as independent, our likelihood function is given by $L = \prod_i L_i$, where

$$L_i = \left(f_{T_1}(t_i)\ [1 - F_{T_2}(t_i)] \right)^{\delta_i} \cdot \left([1 - F_{T_1}(t_i)]\ f_{T_2}(t_i) \right)^{1-\delta_i}. \tag{19}$$

In (19) above, δ_i is an indicator function taking the value 1 if the ith event was due to the type 1 event, and 0 if it was due to the type 2 event, while $f_{T_j}(\cdot)$ and $F_{T_j}(\cdot)$, $j = 1, 2$, are used to denote the respective probability density and cumulative distribution functions for the two competing event times.

24

It is not possible to directly employ the likelihood approach outlined above using T_1 as the system time of a LP patient and T_2 as the corresponding time to renege, because the system times of successive patients are not independent. Rather, they are highly correlated since successive system times in a heavily-loaded queue will have a great degree of overlap. Instead, we consider the sequence of observed "inter-exit times" within each priority stream, which we assume are at least approximately independent.

Calibration of the LP stream using maximum likelihood based on the sequence of observed inter-exit times from the LP queue proceeds as follows. Suppose n_2 such exit events are observed in the LP queue over the duration of the study period, and that these are indexed by i. Let the random variable T_i denote the ith inter-exit time, and let t_i be its observed value. Here, $T_i = \min\{T_{1i}, T_{2i}\}$, where T_{1i} is the random time until the next LP transplantation and T_{2i} is the random time until the next abandonment. The ith inter-exit time then contributes the factor L_i from (19) to the likelihood function L, where δ_i is an indicator of successful service.

Due to the memoryless property of the exponential distribution, each residual time to transplant remains exponentially distributed at rate μ_2. Each residual time until the first abandonment from the LP queue is also still exponentially distributed (because it is the minimum of a finite number of independent exponential random variables), but at an accelerated rate of $\alpha_2(t) = N_2(t)\alpha_2$, where $N_2(t)$ represents the random number of LP patients waiting at time t. In this context, the likelihood function for the LP stream is approximately proportional to

$$L(\mu_2, \alpha_2; t_1, t_2, \ldots, t_{n_2}) = \prod_{i=1}^{n_2} \left(\mu_2 e^{-\mu_2 t_i} \cdot e^{-\alpha_2(t_i)t_i} \right)^{\delta_i} \cdot \left(e^{-\mu_2 t_i} \cdot \alpha_2(t_i) e^{-\alpha_2(t_i)t_i} \right)^{1-\delta_i}. \quad (20)$$

The likelihood function given by (20) is an approximation, because we do not allow for $N_2(t)$ to vary during an inter-exit time due to new arrivals to the LP queue. To further simplify our estimation scheme, we assume a fixed value for the LP queue size, setting $N_2(t) = N_2 \in \mathbb{Z}^+$. Our rationale for this restriction is that, although there is fluctuation in the LP queue size, it appeared to remain fairly constant in the liver transplant data we

consider in the next section.

Given such data, the maximum likelihood estimates (MLEs), obtained by maximizing the log-likelihood (with respect to the parameters), are as follows:

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n_2} \delta_i}{\sum_{i=1}^{n_2} t_i} \quad \text{and} \quad \hat{\alpha}_2 = \frac{\sum_{i=1}^{n_2} (1 - \delta_i)}{N_2 \sum_{i=1}^{n_2} t_i} \; .$$

We observe that if there were no abandonments and no censoring, the estimated service rate would be the reciprocal of the average service time, as one would expect from a random sample of an exponential distribution.

In our framework, the probability p of being promoted to HP status occurs at random. LP patients who renege are considered to either immediately "self-promote" to the HP queue with probability p, or to exit the system with probability $q = 1 - p$, independently of the others. The total number of reneging LP patients who become promoted is thus a binomial random variable, whose MLE $\hat{p}$ is well known to be the empirical proportion of "successes". In our context, $\hat{p}$ is the observed proportion of reneging LP patients who receive promotion to the HP queue.

The calibration of the HP stream follows by analogy to the LP case described above. The LP and HP streams are analyzed separately due to the assumption that HP patients are served on a FCFT basis, regardless as to how they entered that queue. Suppose that there were n_1 customers who entered the HP queue, either directly as a new arrival or through reneging from the LP queue, and that these observations are indexed by j. The resulting MLEs for the HP parameters are

$$\hat{\mu}_1 = \frac{\sum_{j=1}^{n_1} \delta_j^*}{\sum_{j=1}^{n_1} t_j} \quad \text{and} \quad \hat{\alpha}_1 = \frac{\sum_{j=1}^{n_1} (1 - \delta_j^*)}{N_1 \sum_{j=1}^{n_1} t_j} \; ,$$

where δ_j^* is an indicator function taking the value 1 if the jth HP patient successfully achieved service, $\{t_j; \; j \in \mathbb{Z}^+\}$ are the observed inter-exit times in the HP queue, and N_1 is the (assumed) number of HP patients present in the queue at any given instant.

Finally, the class-specific arrivals are assumed to follow separate, homogeneous Poisson processes and we assume patients arrive independently to the queueing system. In this

context, the MLEs are clearly $\hat{\lambda}_1 = n_1/S$ and $\hat{\lambda}_2 = n_2/S$, where S denotes the length of the study period.

7   Case Study: Analysis of Liver Transplantation Data

In this section, we present an analysis of liver transplantation wait-list data provided by a regional health centre in Canada, under its Research Ethics Board guidelines for secondary use of anonymous information. Anonymous parameters were derived from a data set of nearly 1,100 patients who were on that region's liver wait list from January 2000 through December 2012, inclusive. A status and date profile was provided for each patient. This information tracked changes to each patient's CanWAIT (Canadian wait-listing algorithm in transplantation) status for human livers over the duration of their stay on the wait list. The ordinal set of CanWAIT status codes for a human liver are described in Table 1 (adapted from Bazarah et al. [3], Table 2).

As noted in [3], donor livers are typically allocated and transplanted regionally. However, patients with a CanWAIT status of 3F, 4, or 4F are considered "high-status patients", and are placed on a national wait list. Donor livers are directed to these patients as they become available from anywhere in the country. We categorized such patients as our HP patients, while the remaining patients with CanWAIT scores of 1, 1T, 2, or 3 were considered to be LP patients. Patients with initial scores in this range, whose health subsequently degraded to a 3F, 4, or 4F status, were considered to be LP patients who self-promoted to the HP queue. If a patient was placed "on hold" (a CanWAIT score of 0), we assumed their priority in the queue did not change during the period they were in that status. In addition, although some short-duration oscillations between the LP and HP states were observed in the records of some patients, these were ignored and those patients were treated as if they stayed in the originating priority stream for the entire period in question.

Estimates for each of the model's parameters were obtained via the methodology outlined in Section 6. Estimates for N_1 and N_2, the number of patients in each respective

CanWAIT Status	Patient Criteria
1	At home
1T	At home with a liver tumour
2	In hospital in stable condition
3	In intensive or equivalent care facility but not requiring mechanical support, with either: Creatinine > 200 mmol/L or rising by > 50 mmol/L/day; or, Grade 3 or 4 encephalopathy
3F	In intensive or equivalent care facility for fulminant liver failure but not on mechanical support, who fulfills the King's College criteria for high risk of mortality without liver transplantation
4	In intensive care requiring mechanical ventilation support; without liver transplantation, death is considered imminent
4F	In intensive care requiring mechanical ventilation for fulminant liver failure, including nonfunction of a primary graft; without liver transplantation, death is considered imminent
0	On hold

Table 1: Canadian wait-listing algorithm in transplantation (CanWAIT) status codes for liver transplantation (adapted from Bazarah et al. [3], Table 2).

priority stream at any given instant of time, were obtained by viewing time-series plots of the moving average of each of the two priority queue sizes. A set of ABO blood-type specific estimates appears in Table 2. Due to the disparate relative frequencies of blood types in Canada (e.g., see [5]), no results for blood type AB are presented. There were only about 60 type AB patients and of these, less than a dozen patients entered the HP queue (either

directly or through self-promotion). Consequently, the corresponding parameter estimates were deemed not to be reliable. Blood types A and O each had over 400 patients in the wait-list records and approximately 130 patients had type B blood.

We remark that the LP placement rate for type O of 0.08214 per day is about 10% higher than type A's placement rate of 0.07665 per day, which reflects the Canadian blood mix of 46% type O versus 42% type A quite nicely (e.g., see [5]). In contrast to this, the type A transplantation rate of 0.05830 per day is itself about 10% higher than the type O value of 0.05354 per day. In the case of ABO-identical transplantation, we would anticipate the reverse situation, so this seems to suggest that recipients of type A must be getting a non-trivial number of type O livers. This is in keeping, qualitatively at least, with what has been reported in other jurisdictions (e.g., see Glander et al. [7], Liefeldt et al. [18], and other references listed in Stanford et al. [22]). The estimates also suggest that there is about one HP placement for every 5 LP placements for blood types O and B, and about one HP placement for every 6 LP placements for blood type A.

The model was run under three scenarios (i.e., for each of blood types O, A, and B), corresponding to the sets of parameter estimates given in Table 2. The buffer sizes used for our model, along with the corresponding set of blocking probabilities obtained, are displayed in Table 3 for each blood-type specific case. We remark that the chosen values for m and n yield negligible blocking probabilities in all cases.

Figures 2 through 4 display the estimated cumulative distribution functions of the system times for three different categories of successfully-transplanted patients. Each figure presents model-based and empirically-estimated results for the three blood-type specific models. We note that Figures 2 and 3 are measured in days, while Figure 4 is measured in months, reflecting the fact that LP patients who ultimately get transplanted in the LP queue wait much longer than individuals who are transplanted as HP patients.

Figures 2 and 3 compare the empirical waiting times of HP patients with the results of our model for patients who arrived directly to the HP queue and for LP arrivals that subsequently self-promoted to the HP queue, respectively. For LP arrivals that subsequently

Parameter	Type O	Type A	Type B
λ_1	0.01605	0.01225	0.00465
λ_2	0.08214	0.07665	0.02196
μ_1	0.11888	0.37037	0.20968
μ_2	0.05354	0.05830	0.01449
α_1	0.05828	0.05926	0.14516
α_2	0.00096	0.00148	0.00120
p	0.23810	0.20430	0.23404
N_1	1	1	1
N_2	31	13	9

Table 2: Blood-type specific parameter estimates for the study data. (* Note that the blood type AB results are not presented due to small-sample effects, as this blood type is rare in Canada.)

Quantity	Type O	Type A	Type B
m	65	40	20
n	3	3	3
$LPBlock$	0.000735	0.000171	0.001226
$HPBlock$	0.001064	0.000039	0.000003
$PromLPBlock$	0.000116	0.000003	0.000000

Table 3: Blood-type specific buffer sizes used and blocking probabilities obtained for the study data.

self-promoted to the HP queue, the "system time" we display starts at the moment of promotion. Typically, the percentiles estimated from the models are similar to the corresponding empirical percentiles. The empirical results for blood type B are more variable, which is not suprising as this blood type is less common. Less than a dozen patients with

type B blood self-promoted from LP to HP, which explains why the bottom panel of Figure 3 is not as smooth as the empirical cumulative distribution functions for types O and A.

Whereas our HP results might be described as adequate, the same could not be said for the LP queue. In contrast to the consistency of the HP results, Figure 4 reveals that the LP system time forecast by the model can be overly conservative for short delays, and can fail to capture the extreme tail of the empirical waiting time distribution. Upon reflection and further analysis of the source data records, it appears that the empirical behaviour is consistent with the operation of a wait list involving three priority levels. Specifically, patient codes 0 through 3 are being further pooled into two distinct priority classes, and not a single class as we have considered here. This possibility calls for investigation of other centres, to see whether what we have observed at the health centre under study is common there as well. In parallel, it suggests that a three-class analogue of the model being presented here should be pursued, even if the present model may indeed prove to be adequate at a wide range of health centres.

We remark that the existence of a multi-level clinical distinction of patient acuity such as CanWAIT rarely translates operationally class for class, in the second author's experience, to a multi-priority queue. Typically, there are either two priority classes, or three priority classes as we have observed at play here. An iterative model building approach, aided by expert perception of what is occurring, medically and mathematically, is therefore needed to determine the number of priority classes in operation.

Table 4 compares the model-based reneging rates for each of the three separate models to the corresponding empirically-observed rates. The overall reneging rate from the LP queue as well as the abandonment rate of LP patients from the entire system are given. Separate reneging rates for the two types of HP patients, namely those who directly arrived to the HP queue and those who arrived as a LP patient and subsequently self-promoted to the HP queue, are also given. In general, the model-based reneging rates exceed those observed in the data for blood type O. A similar pattern is also seen for the blood type A results, although the model and empirical-based estimates are in relatively close agreement.

31

Blood Type & Rate Estimate	Overall LP	LP Abandon Only	HP	Promoted LP
Type O (model)	0.4348	0.3313	0.3522	0.3527
Type O (empirical)	0.3779	0.2879	0.3289	0.3429
Type A (model)	0.2749	0.2187	0.1420	0.1423
Type A (empirical)	0.2562	0.2039	0.1379	0.1579
Type B (model)	0.3730	0.2857	0.4123	0.4125
Type B (empirical)	0.4519	0.3462	0.4091	0.4545

Table 4: Model-based and empirical-based estimates of reneging rates, stratified by the blood type of the patient. Overall LP is the overall reneging rate from the LP queue, while LP Abandon Only is the abandonment rate. HP is the reneging rate of HP patients who arrived directly to that priority stream, while Promoted LP is the reneging rate of HP patients who arrived as LP patients and subsequently self-promoted to the HP queue.

For the type B model, the opposite is observed and there is close agreement for the reneging rates of patients who entered directly as HP.

Table 5 presents model-based and sample-based estimates for the mean system times, as well as the corresponding estimated standard deviations, for the different categories of patients who were successfully transplanted. The model means for direct-entry HP patients and self-promoted LP patients are similar for type O patients and so are the corresponding standard deviations. The empirical means for successful HP patients – regardless as to how they entered that stream – are about twice as large than what is estimated by the model for type A and type B patients and the corresponding empirical standard deviations are also larger than those based on the fitted models. Conversely, the model-based estimated mean system time for successful LP patients is much larger than the HP means, regardless of blood type. For type A and type B patients, the empirical mean time to transplant for LP patients is larger than the model means. The empirical standard deviations for the mean time to transplantation as a LP patient are also quite large, reflecting the fact that

	Model Mean	Empirical Mean	Model SD	Empirical SD
Type O Model				
Successful direct-entry HP	6.1	6.2	6.0	5.7
Successful self-promoted LP	6.1	6.1	6.0	6.4
Successful LP	581.5	474.9	156.5	532.5
Type A Model				
Successful direct-entry HP	2.4	4.8	2.4	4.1
Successful self-promoted LP	2.4	5.2	2.4	4.6
Successful LP	209.2	306.5	104.0	304.3
Type B Model				
Successful direct-entry HP	2.8	4.5	2.8	4.1
Successful self-promoted LP	2.8	5.7	2.8	5.8
Successful LP	360.2	528.7	211.7	508.2

Table 5: Model-based and empirical-based estimates of the mean system times (measured in days) for patients who were successfully transplanted, along with their corresponding standard deviations (SDs). Results are given for each of the blood-type specific models. The system times for three different types of patients are considered: direct-entry HP patients, self-promoted LP patients, and LP patients who never left the LP queue.

there are some patients who appear to spend a very long time in the LP queue, without abandoning that queue or self-promoting to the HP queue due to health degradation, who eventually get successfully transplanted.

Overall, these results seem to suggest that our model of LP system time fails to capture the high observed variability in waiting times, and the HP models are generally overly conservative. Future work will need to pursue all of the following avenues: revisiting the model assumptions regarding placement and abandonment rates, refining the parameter estimates, and testing the model against observed data from other wait lists. Nonetheless,

we remark that a queueing model would still be a useful decision-making tool if it could be used qualitatively to explore the impact of various policy changes. Its use is not primarily a matter of forecasting the likelihood of delays as experienced by individual patients placed on the wait list.

8 Concluding Remarks

We have presented a model for estimating patient waiting times in deceased-donor transplant queues, which reflects both the propensity for patients to renege from the queue, and to self-promote to urgent status. Performance measures of interest we obtain include the waiting time distributions and their moments, the queue length distributions, and the reneging probabilities. These results are obtained for patients who are urgent when placed on the wait list, for those who have a regular status and receive their organ in regular status, and for those regular status placements who become urgent prior to transplantation.

In addition to the probabilistic results, which take the form of matrix geometric solutions for the queue lengths and phase-type distributions for the waiting times, we also have presented a maximum-likelihood-based procedure for estimation of the model parameters. These results were then applied to study a single liver transplantation centre in Ontario. Whereas the fit of waiting time for patients transplanted with an urgent status appeared to be appropriate, the fit for the regular stream of patients was deemed inadequate, since it appeared that the wait list comprised three priority classes operationally. Future work will proceed along two directions. On the one hand, we hope to be able to apply the data to other transplantation centres, to see if our two-class priority model is appropriate. On the other, we plan to extend the existing results to allow for a third priority class. The approach would parallel the methodology used in the two-class case.

We remind readers that some of the estimates in the previous section were based on the analysis of relatively small samples of data, in particular for the case of blood type B in the overall population. Furthermore, we felt that our estimates of AB data were too imprecise

to be reliable, as they were based on too few data values. More precise estimates could be obtained through an analysis of longer records or through analysis of data from multiple health centres over the same study period.

Our model's assumption that all blood types have the same propensity to renege per unit time waiting is consistent with the medical view that blood type does not influence health degradation. Nonetheless, the longer a patient waits, the greater the chance for that patient to experience degraded health. The results obtained from our model suggest a substantial discrepancy in the HP reneging rate for blood type B than for blood types O and A; however, once again one needs to view this in the context of small numbers of urgent patients of all blood types. It would seem that further work to statistically estimate the abandonment parameters is called for.

Finally, it is clear that donor allocation in transplantation is a complex system depending on several interacting clinical and administrative events. We believe we have shown that this system is, despite its complexity, amenable to logical analysis using standard queueing theory. Further work is needed in terms of adequate parameter estimation, and other possible refinements to the modelling assumptions noted above may be needed after further numerical investigations.

Acknowledgements

References

[1] Abellán, J.J., Armero, C., Conesa, D., Pérez-Panadés, J., Martínez-Beneito, M.A., Zurriaga, O., García-Blasco, M.J., Vanaclocha, H.: Analysis of the renal transplant waiting list in the País Valencià (Spain). Statistics in Medicine **25**, 345-358 (2006)

[2] Barone, M., Avolio, A.W., Di Leo, A., Burra, P., Francavilla, A.: ABO blood group-related waiting list disparities in liver transplant candidates: Effect of the MELD adoption. Transplantation **85**, 844-849 (2008)

[3] Bazarah, S.M., Peltekian, K.M., McAlister, V.C., Bitter-Suermann, H., MacDonald, A.S.: Utility of MELD and Child-Turcotte-Pugh scores and the Canadian waitlisting algorithm in predicting short-term survival after liver transplant. Clinical and Investigative Medicine **27**, 162-167 (2004)

[4] Bright, L., Taylor, P.G.: Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. Stochastic Models **11**, 497-525 (1995)

[5] Canadian Blood Services. Blood Type Facts. Accessed April 27th, 2013 at `http://www.blood.ca/CentreApps/Internet/UW_V502_MainEngine.nsf/page/Blood+Type+Facts?OpenDocument&CloseMenu`

[6] Gaver, D.P., Jacobs, P.A., Latouche, G.: Finite birth-and-death models in randomly changing environments. Advances in Applied Probability **16**, 715-731 (1984)

[7] Glander, P., Budde, K., Schmidt, D., Fuller, T.F., Giessing, M., Neumayer, H.-H., Liefeldt, L.: The 'blood group O problem' in kidney transplantation – time to change? Nephrology Dialysis Transplantation **25**, 1998-2004 (2010)

[8] Gómez-Corral, A., Krishnamoorthy, A., Narayanan, V.C.: The impact of self-generation of priorities on multi-server queues with finite capacity. Stochastic Models **21**, 427-447 (2005)

[9] Hussey, J.C., Parameshwar, J., Banner, N.R.: Influence of blood group on mortality and waiting time before heart transplantation in the United Kingdom: Implications for equity of access. Journal of Heart and Lung Transplantation **26**, 30-33 (2007)

[10] Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data, 2nd edition. John Wiley & Sons Ltd, New Jersey (2002)

[11] Krishnamoorthy, A., Narayanan, V.C.: On a queueing system with self generation of priorities. In: Stochastic Point Processes (Eds. S.K. Srinivasan & A. Vijayakumar), Narosa Publishing House, Chennai, India, pp. 212-217 (2003)

[12] Krishnamoorthy, A., Narayanan, V.C., Deepak, T.G.: On a queueing system with self generation of priorities. Neural, Parallel & Scientific Computations **13**, 119-130 (2005)

[13] Krishnamoorthy, A., Babu, S., Narayanan, V.C.: MAP/(PH/PH)/c queue with self-generation of priorities and non-preemptive service. Stochastic Analysis and Applications **26**, 1250-1266 (2008)

[14] Krishnamoorthy, A., Babu, S., Narayanan, V.C.: The MAP/(PH/PH)/1 queue with self-generation of priorities and non-preemptive service. European Journal of Operational Research **195**, 174-185 (2009)

[15] Krishnamoorthy, A., Narayanan, Chakravarthy, S.R.: The impact of priority generations in a multi-priority queueing system – a simulation approach. In: Proceedings of the 2009 Winter Simulation Conference (Eds. M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin & R.G. Ingalls), IEEE Computer Society, New York, USA, pp. 1622-1633 (2009)

[16] Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. ASA SIAM, Philadelphia (1999)

[17] Lawless, J.F.: Statistical Models and Methods for Lifetime Data, 2nd edition. John Wiley & Sons Ltd, New Jersey (2003)

[18] Liefeldt, L., Budde, K., Glander, P.: Accumulation of elderly ESRD patients with blood group O on the waiting list. (Letter to editors). Transplant International **24**, e83-e84 (2011)

[19] Maertens, T., Walraevens, J., Bruneel, H.: Performance comparison of several priority schemes with priority jumps. Annals of Operations Research **162**, 109-125 (2008)

[20] Pinsky, M.A., Karlin. S.: An Introduction to Stochastic Modeling, 4th edition. Academic Press, Boston (2011)

[21] Stanford, D.A., Renouf, E.M., McAlister, V.C.: Waiting for liver transplantation in Canada: Waitlist history 2000-04 and sensitivity analysis for the future. Health Care Management Science **11**, 184-195 (2008)

[22] Stanford, D.A., Lee, J.M., Chandok, N., McAlister, V.C.: A queueing model to address wait time inconsistency in solid-organ transplantation. Operations Research for Health Care **3**, 40-45 (2014)

[23] Su, X., Zenios, S: Patient choice in kidney allocation: The role of the queueing discipline. Manufacturing and Service Operations Management **6**, 280-301 (2004)

[24] Syski, R.: Passage Times for Markov Chains. IOS Press, Amsterdam (1992)

[25] Tijms, H.C.: A First Course in Stochastic Models. John Wiley & Sons Ltd, Chichester (2003)

[26] Wang, Q.: Modeling and analysis of high risk patient queues. European Journal of Operational Research **155**, 502-515 (2004)

[27] Wiesner, R., Edwards, E., Freeman, R., Harper, A., Kim, R., Kamath, P., et al.: Model for end-stage liver disease (MELD) and allocation of donor livers. Gastroenterology **124**, 91-96 (2003)

[28] Zenios, S.A.: Modeling the transplant waiting list: A queueing model with reneging. Queueing Systems **31**, 239-251 (1999)

[29] Zenios, S.A., Chertow, G.M., Wein, L.M.: Dynamic allocation of kidneys to candidates on the transplant waiting list. Operations Research **48**, 549-569 (2000)
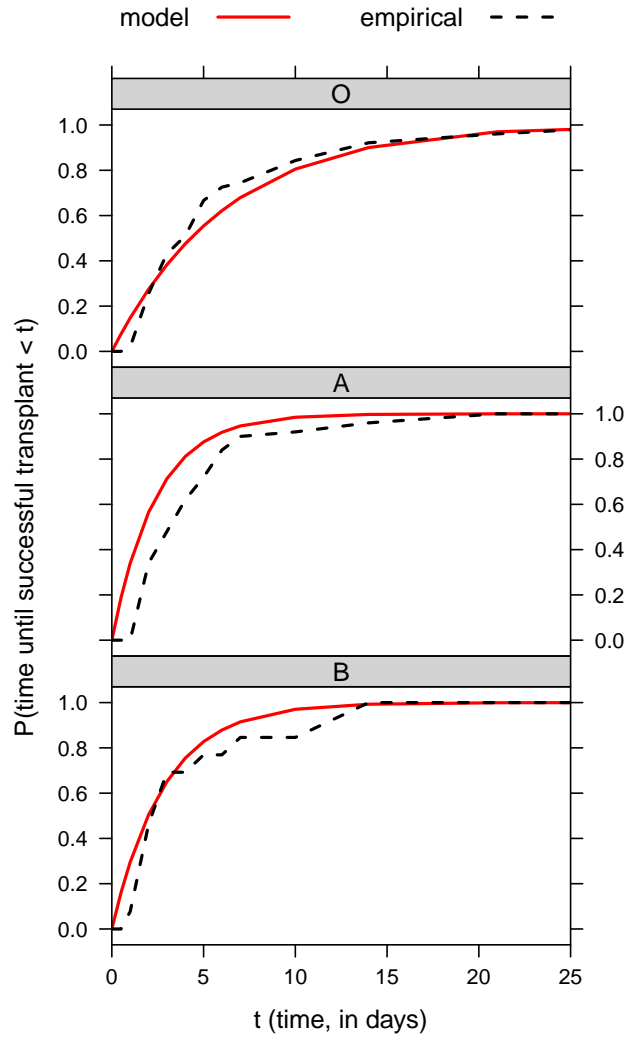
Figure 2: Model-based (solid red lines) and empirical-based (dashed black lines) estimates for the cumulative distribution functions of the system time of a successfully-transplanted HP patient who entered the system originally as an HP patient. Results for each of the blood-type specific models appear in separate panels.
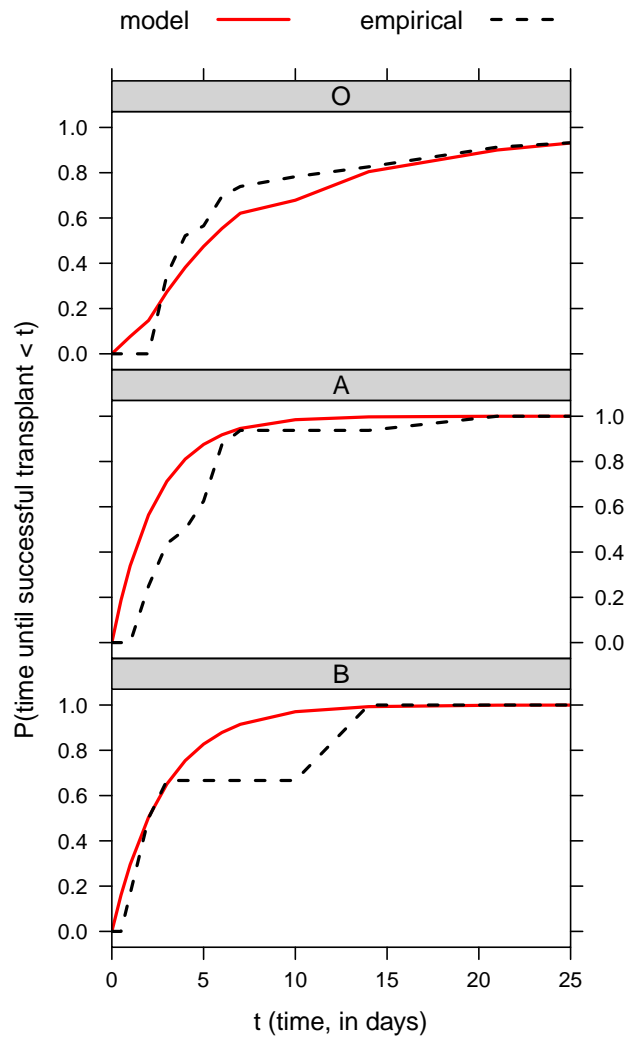
Figure 3: Model-based (solid red lines) and empirical-based (dashed black lines) estimates for the cumulative distribution functions of the system time of a successfully-transplanted HP patient who originally arrived as a LP patient and subsequently self-promoted to the HP queue. Here, system time is measured from the time of promotion to the HP queue. Results for each of the blood-type specific models appear in separate panels.
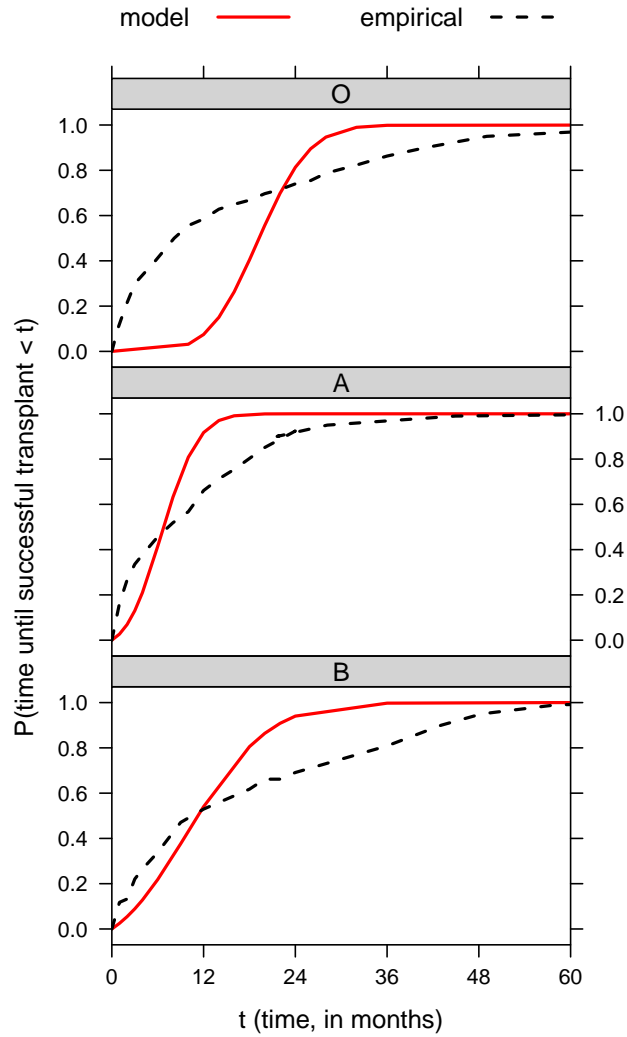
Figure 4: Model-based (solid red lines) and empirical-based (dashed black lines) estimates for the cumulative distribution functions of the system time of a LP patient who was successfully transplanted within the LP queue. Results for each of the blood-type specific models appear in separate panels.