Memorandum 1961 (November 2011). ISSN 1874-4850. Available from: http://www.math.utwente.nl/publications
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Appointments for Care Pathway Patients

Maartje E. Zonderland*· Richard J. Boucherie†· Ahmad Al Hanbali‡

November 14, 2011

Acknowledgments: The authors would like to thank Ivo Adan and Vidyadhar Kulkarni for their valuable comments.

1 Introduction

Care pathways have gained popularity in the healthcare sector the last two decades [2]. A care pathway is a management tool to organize multidisciplinary care for patients with identical characteristics (i.e., disease symptoms, diagnosis, age, etcetera). The care pathway specifies the steps in the care process [1] and routes patients along a pre-defined path of care providers and diagnostic facilities. Patients may complete a significant part of the path in one day. Given the vast number of hospital facilities incorporated in the path, the planning is usually involved and hospitals tend to prioritize these patients. It is therefore not uncommon that slots are reserved for care pathway patients in an otherwise walk-in clinic. Examples are for instance found at diagnostic services, such as Radiology outpatient clinics (X-ray, CT) and blood withdrawal facilities. When these facilities are highly utilized (>85%), reserving a few slots for care pathway may lead to a significant increase of the waiting time of walk-in patients.

In this paper we translate the above problem setting to a queuing model. The hospital facility decides on the number of slots that is reserved for care pathway patients. The model then enables a trade-off between the delay for walk-in patients and the probability that the number of slots reserved for the care pathway patients is not sufficient.

The service and hospitality industry is quite familiar with policies where a part of the (unscheduled) customer stream is diverted and scheduled on a later moment on the day. This concept is also known as virtual queuing (see e.g., [4, 9]). Probably the most famous organization that employs virtual queuing is Walt Disney, that uses for the most popular attractions in its theme parks the FastPass system [5]. Park guests decide upon arrival at an attraction whether they want to join the waiting line, or get a ticket (the 'FastPass'), that gives them a time-frame to return and enter the attraction without waiting. To avoid a large number of no-shows and long waiting time for the non-FastPass guests, it is only allowed to possess a FastPass ticket for one attraction at the same time. The queuing system behind FastPass is analyzed in [7]. However, in the FastPass system park guests are supported by information on the state of both the regular and FastPass queue (i.e., the waiting time in the regular queue and the come back time for the FastPass ticket) and decide upon arrival which queue they want to join. In this paper, the two patient types originate from separate arrival processes (walk-in or care pathway) that determine their type and thus

*Stochastic Operations Research & Center for Healthcare Operations Improvement and Research, University of Twente, Postbox 217, 7500 AE Enschede, the Netherlands, and Division I, Leiden University Medical Center, Postbox 9600, 2300 RC Leiden, the Netherlands. E: m.e.zonderland@lumc.nl

†Stochastic Operations Research & Center for Healthcare Operations Improvement and Research, University of Twente. E: r.j.boucherie@utwente.nl

‡Operational Methods for Production and Logistics, University of Twente. E: a.alhanbali@utwente.nl

the queuing discipline. We have found no evidence that the particular reservation discipline we consider has been studied before.

The remainder of this paper is organized as follows. In the next section we describe our queuing model, followed by the analysis in Section 3. In Section 4 we provide a couple of numeric examples, and we conclude with the discussion in Section 5.

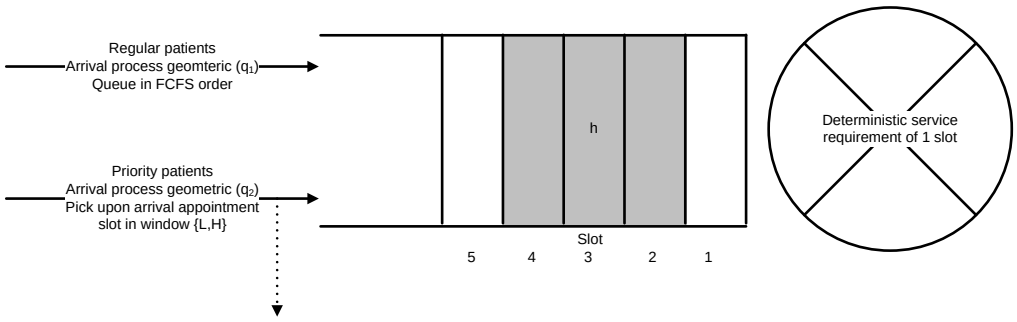
2 Model

For the ease of notation we refer to the walk-in patients as regular patients, and to the care pathway patients as priority patients.

2.1 Assumptions

We consider a hospital facility which serves regular and priority patients. Both patient types have a deterministic service time requirement of 1 slot and arrive according to a geometric arrival process with arrival probability q_1 and q_2 respectively. The regular patients queue in FCFS order, while a priority patient picks upon arrival an appointment h slots later, $L \leq h \leq H$, where $1 \leq L \leq H < \infty$. When the desired slot is already taken by another priority patient, the newly

Figure 1: The $G/D/1$ queue with appointments, appointment window $(L, \ldots, H) = (2, 3, 4)$ and $h = 3$.



arrived priority patient proceeds to slot $h-1,\ldots,L$, until a slot is found that has not yet been taken by a priority patient. When all slots in the window that precede h are taken, the priority patient is blocked and lost. If the slot taken by the priority patient is occupied by a regular patient, then the regular patient is shifted to the first higher slot that is not taken by a priority patient. If this slot is non-empty as well, the regular patient that was occupying this slot is shifted upwards to the first slot not taken by a priority patient, and so on. Note that h equals the maximum number of slots the new priority patient has to wait until his service commences. It can readily be observed that the service facility can be modeled as a discrete-time single server queue serving priority and regular patients. Regular patients join the back of the queue. Priority patients select the last slot in the interval $(L, \ldots, h)$. Regular patients are shifted to higher queue positions when a priority patient takes their position (see Figure 1). The slot pick probability p_h can follow any discrete probability distribution. While the priority patients do not 'see' regular patients, the regular patients may experience significant delay when a priority patient joins the queue. If there is a priority patient on the first queue position at the moment of a service completion, this patient is served. Otherwise, a regular patient will be served. If there are no regular patients in the queue, the server is idle (even though there may be a priority patient on a slot position higher up in the queue).

2.2 Matrix Structure

The transitions in the appointment window at the end of each time slot are independent of the number of regular patients present. We therefore first define a submatrix with the 1-step transition probabilities for priority patients. Then we define submatrices for the 1-step transition probabilities of regular patients, which do depend on the state of the priority patient appointment window. Finally, we combine these matrices into one transition probability matrix.

2.2.1 Priority Patient Transition Probability Submatrix D

We define an appointment vector $\mathbf{v}$ of length H, specifying which slots contain priority patient appointments. At most one priority patient can claim an appointment slot, so $\mathbf{v} = (v_1 \dots v_H)$, where v_h is a binary variable, equal to 1 when slot h is reserved by a priority patient and 0 otherwise. Note that the appointment vector $\mathbf{v}$ is of length H, while the appointment window is of length $H - L + 1$. Even though the slots $(1, \dots, L-1)$ in the appointment window cannot be chosen anymore by priority patients, they possibly contain appointments and thus should be taken into account in the analysis. At the end of each time slot $\mathbf{v}$ is updated; new appointments are added and existing appointments are moved forward one slot. There are 2^H possible combinations for $\mathbf{v}$: when $H = 4$, $\mathbf{v}$ can for example be equal to $(0000), (0101), (1101)$, and so on. It follows immediately that the 1-step transition probability submatrix, D, has size $2^H \times 2^H$. Deriving D can be quite cumbersome for $H > 2$. We therefore present an algorithm to simplify this process.

2.2.2 Algorithm for computation of D

Step 1. Initialization
1a. Create the 2^H possible appointment combinations and order them lexicographically.
1b. Create an (empty) matrix of size $2^H \times 2^H$, where the rows and columns represent the 2^H lexicographically ordered possible combinations for $\mathbf{v}$ at time slot t and $t+1$ respectively.

Step 2. Creating the Block Structure
The possible shifts in $\mathbf{v}$ at the end of each time slot lead to a unique submatrix structure. Since at the end of each time slot the appointments are advanced one slot, all vectors with a 1-entry (an appointment) on position x, $x > 1$, will not have a possible transition to a vector with a 0-entry (no appointment) one position to the left, i.e., on position $x - 1$. Also, since appointments on the first position will be removed from $\mathbf{v}$ in the next shift, the submatrix' structure is identical for the first and second 2^{H-1} rows. Figure 2 shows the repetition in the structure of D for $H = \{1, ..., 4\}$. In fact, for $H > 3$ the upper-left block of four rows and eight columns is repeated each four rows down and eight columns to the right.

Step 3. Calculating the Required Number of Arrivals N
For each possible transition a certain number of priority patient arrivals, N, is required. It follows that for $H > 3$ the upper-left 4×8 building block is filled with the number of required arrivals, as given in Figure 3, and each repetition to the right, the required number of arrivals is raised by one. When the first entry of $\mathbf{v}$ in the column of D equals 1, a minimum number of arrivals is required to make this transition (denoted in Figure 3 with $N = n+$). When the first entry of $\mathbf{v}$ equals 0, an exact number of arrivals is required to make this transition ($N = n$). For example, see Figure 3. For the transition from (1000) to (0111) exactly 3 arrivals are required, but for the transition from (0001) to (1011) at least 2 (2+) arrivals are required. Not only the structure of the upper-left building block is identical for $H > 3$, but also the required number of arrivals (as given in Figure 3) remains the same.

Step 4. Adapting the Blocks for $L > 1$
If $L > 1$, the slots $(1, ..., L-1)$ cannot be claimed by priority patients. This changes the structure of D: the blocks are halved $L - 1$ times. In the left half of the remaining part of the block n arrivals are required, while in the right half n or more arrivals are required (see Figure 4 for an example with $H = 3$ and $L = \{1, 2, 3\}$).

Figure 2: Structure of D for $H = \{1, \ldots, 4\}$

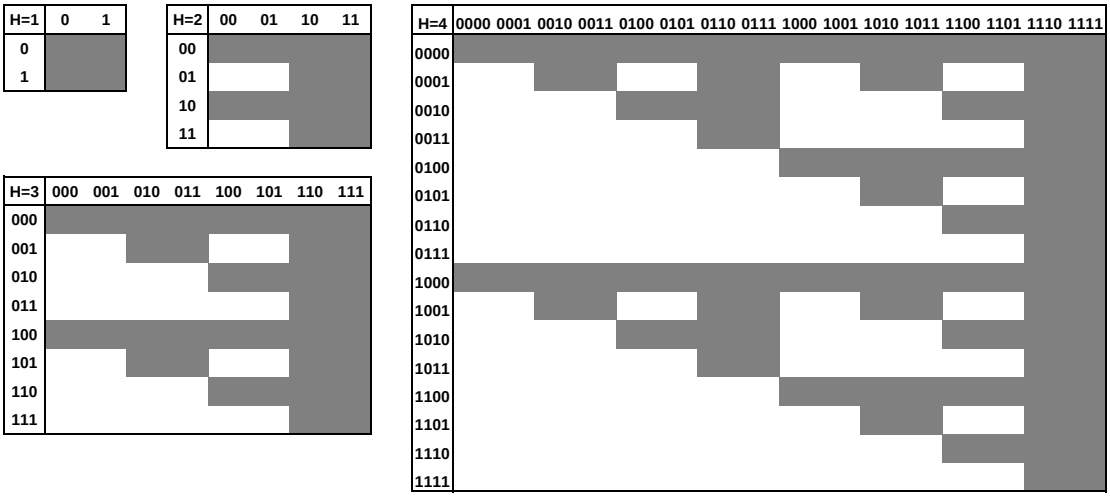


Figure 3: Required number of arrivals in D for $H = 4$

H=4	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
0000	0	1	1	2	1	2	2	3	1+	2+	2+	3+	2+	3+	3+	4+
0001			0	1			1	2			1+	2+			2+	3+
0010					0	1	1	2					1+	2+	2+	3+
0011							0	1							1+	2+
0100									0+	1+	1+	2+	1+	2+	2+	3+
0101											0+	1+			1+	2+
0110													0+	1+	1+	2+
0111															0+	1+
1000	0	1	1	2	1	2	2	3	1+	2+	2+	3+	2+	3+	3+	4+
1001			0	1			1	2			1+	2+			2+	3+
1010					0	1	1	2					1+	2+	2+	3+
1011							0	1							1+	2+
1100									0+	1+	1+	2+	1+	2+	2+	3+
1101											0+	1+			1+	2+
1110													0+	1+	1+	2+
1111															0+	1+

Step 5. Calculating the Transition Probabilities
In the last step of the algorithm we need to calculate the transition probabilities $\mathbb{P}(\mathbf{v}^t \to \mathbf{v}^{t+1})$ that fill the gray cells in D (in all white cells, no transition is possible and $\mathbb{P}(\mathbf{v}^t \to \mathbf{v}^{t+1}) = 0$). Recall that we use N to denote the number of required arrivals as given in D. The transition probabilities are multinomial distributed and given by:

$$\mathbb{P}(\mathbf{v}^t \to \mathbf{v}^{t+1}) = \begin{cases} 0 & \text{if} \quad \mathbf{v}^t \not\to \mathbf{v}^{t+1}, \\ \sum_{j=N}^{J} b_j \sum_{\substack{k_L,\ldots,k_H \\ \sum_{h=L}^{H} k_h = j}} \binom{j}{k_L, \ldots, k_H} p_H^{k_H} \cdots p_L^{k_L} & \text{otherwise,} \end{cases} \tag{1}$$

where $J = N$ if $N = n$ and ∞ if $N = n+$, b_j is the geometric probability that j priority patients arrive in a time slot, given by:

$$b_j = (1 - q_2) q_2^j, \tag{2}$$

Figure 4: Structure and required number of arrivals in D for $H = 3$ and $L = \{1, 2, 3\}$

L=1	000	001	010	011	100	101	110	111
000	0	1	1	2	1+	2+	2+	3+
001			0	1			1+	2+
010					0+	1+	1+	2+
011							0+	1+
100	0	1	1	2	1+	2+	2+	3+
101			0	1			1+	2+
110					0+	1+	1+	2+
111							0+	1+

L=2	000	001	010	011	100	101	110	111
000	0	1	1+	2+				
001			0+	1+				
010					0	1	1+	2+
011							0+	1+
100	0	1	1+	2+				
101			0+	1+				
110					0	1	1+	2+
111							0+	1+

L=3	000	001	010	011	100	101	110	111
000	0	1+						
001			0	1+				
010					0	1+		
011							0	1+
100	0	1+						
101			0	1+				
110					0	1+		
111							0	1+

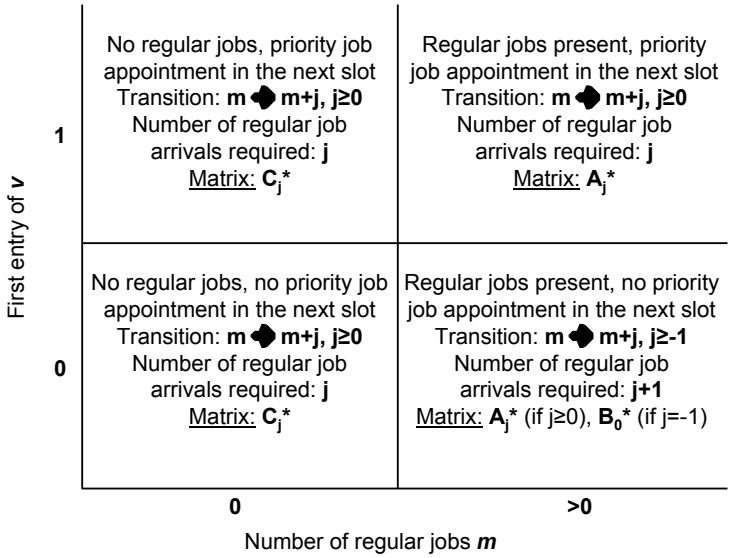
and p_h is the slot pick probability. The distribution of the j arrivals over the slots is denoted by $k_L, \ldots, k_H$, and for each slot $h = (L, \ldots, H)$ the following should hold to ensure the j arrivals are distributed over the slots such that $\mathbf{v}^{t+1}$ is obtained:

$$
\begin{aligned}
&\text{If} \quad (v_h^{t+1} - v_{h+1}^t) = 0 \quad \text{for} \quad h = (L, \ldots, H-1), \quad \text{or} \quad v_H^{t+1} = 0, \\
&\text{then} \quad k_h = 0, \quad \text{and} \quad \sum_{i=h+1}^{H} k_i = \sum_{i=h+1}^{H-1} (v_i^{t+1} - v_{i+1}^t) + v_H^{t+1} \quad \text{for} \quad h = (L, \ldots, H-1). \\
&\text{If} \quad (v_h^{t+1} - v_{h+1}^t) = 1 \quad \text{for} \quad h = (L, \ldots, H-1), \quad \text{or} \quad v_H^{t+1} = 1, \\
&\text{then} \quad \sum_{i=h}^{H} k_i \geq \sum_{i=h}^{H-1} (v_h^{t+1} - v_{h+1}^t) + v_H^{t+1} \quad \text{for} \quad h = (L, \ldots, H-1), \quad \text{and} \quad k_H \geq 1.
\end{aligned}
\tag{3}
$$

2.2.3 Regular Patient Transition Probability Submatrices A^*, B^*, and C^*

While D is the same for all possible priority patient transitions, the regular patient transition probability submatrices, which contain the probabilities for transitions in the number of regular patients present, m, depend on the appointment vector $\mathbf{v}$. Since we consider 1-step transitions, only the first entry of $\mathbf{v}$ is of interest. Three submatrices, A^*, B^*, and C^*, can be identified, which one to apply depends on m and $\mathbf{v}$ (see Figure 5). The submatrices given all have size $2^H \times 2^H$ and are constructed as follows. Define $\mathbf{u}$ and $\mathbf{w}$ as vectors of length 2^H. The first 2^{H-1} entries of $\mathbf{u}$ are equal to q_1, and the second 2^{H-1} entries of $\mathbf{u}$ are equal to 1. The first 2^{H-1} entries of $\mathbf{w}$ are equal to 1, and the second 2^{H-1} entries of $\mathbf{w}$ are equal to 0. Furthermore, define $\mathbf{e}$ as the vector

Figure 5: Applicability of regular patient submatrices



of ones, also of length 2^H. Then we obtain:

$$\begin{aligned} A_j^* &= a_j A^* \quad \text{where} \quad A^* = \mathbf{u}^T \times \mathbf{e}, \\ B_0^* &= a_0 B^* \quad \text{where} \quad B^* = \mathbf{w}^T \times \mathbf{e}, \\ C_j^* &= a_j C^* \quad \text{where} \quad C^* = \mathbf{e}^T \times \mathbf{e}. \end{aligned} \tag{4}$$

Since the arrival process of regular patients is geometrically distributed, the probability a_m that m regular patients arrive in a time slot is given by:

$$a_m = (1 - q_1)q_1^m, \quad m \geq 0. \tag{5}$$

2.2.4 The Combined Transition Probability Matrix P

The priority and regular patient arrival processes are independent, and therefore we can multiply D element wise with A^*, B^*, and C^*, i.e., every (m, n)-entry of D is multiplied with the (m, n)-entry of A^*, B^*, and C^*, in order to obtain the transition probability matrix P with elements A_j, B_0, and C_j, $j \geq 0$. Each entry of P is a matrix in itself of size $2^H \times 2^H$, and represents the state transition $(m_t, \mathbf{v}^t) \to (m_{t+1}, \mathbf{v}^{t+1})$.

$$P = \begin{pmatrix} C_0 & C_1 & C_2 & \cdots & C_m & \cdots \\ B_0 & A_0 & A_1 & \cdots & A_{m-1} & \cdots \\ 0 & B_0 & A_0 & \cdots & A_{m-2} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \cdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

Note that A_j can also be written as $a_j \bar{A} D$, where $\bar{A}$ is the diagonal matrix with the elements of $\mathbf{u}$ on the diagonal. The same holds for B_0, which can be written as $b_0 \bar{B} D$, where $\bar{B}$ is the diagonal matrix with the elements of $\mathbf{w}$ on the diagonal, and for C_j, which can be written as $c_j \bar{C} D$, where $\bar{C}$ is the diagonal matrix with the elements of $\mathbf{e}$ on the diagonal.

3 Analysis

The matrix P shows similarities with the transition probability matrix for the $M/G/1$ queue embedded at departure moments (see [8] for further reference). An overview of discrete time queuing systems can be found in [3]. Several priority disciplines have been studied for discrete time queuing models, but these are usually related to the non-preemptive [12] or preemptive resume priority disciplines [10]. In [11] a different, but related, service discipline is considered, where a slot is reserved for regular patients at the end of the queue. In the case of high load traffic from priority patients, it is then guaranteed that regular patients receive service as well.

3.1 Stability of the Queue

In order for the queue to be stable, the mean load, ρ, should be less than one. Since the service time is 1 slot, ρ equals the sum of the mean number of regular patient arrivals per slot and the accepted priority patients per slot:

$$\rho = \frac{q_1}{1-q_1} + (1 - \mathbb{P}_{B_2})\frac{q_2}{1-q_2} < 1, \tag{6}$$

It follows immediately that $q_1, q_2 < \frac{1}{2}$. The blocking probability for priority patients, $\mathbb{P}_{B_2}$, is calculated as follows. A priority patient is accepted when the slot h, picked with probability p_h, is still available, or if not, when one of the slots $(L, \ldots, h-1)$ is still available. The blocking probability for priority patients is therefore given by:

$$\mathbb{P}_{B_2} = 1 - p_h \cdot \sum_{\substack{v^t \to v^{t+1}: \\ \sum_{i=L}^{h} v_i^{t+1} < h}} \mathbb{P}(\mathbf{v}^t \to \mathbf{v}^{t+1}) \ . \tag{7}$$

3.2 Vector Generating Function of Equilibrium Probabilities $\pi(m, \mathbf{v})$

We derive the vector generating function of the equilibrium probability $\pi(m, \mathbf{v})$ for the number of regular patients present, m, and the realization of the appointment vector, $\mathbf{v}$. For notation purposes, denote $\pi(m, \mathbf{v})$ by the vector π_s, where $s = (0, 1, \ldots)$. Using the property $\Pi P = \Pi$, we obtain:

$$\pi_s = \pi_0 C_s + \sum_{i=1}^{s} \pi_i A_{s-i} + \pi_{s+1} B_0 \quad \text{for} \quad s \geq 1, \quad \text{and} \tag{8}$$

$$\pi_0 = \pi_0 C_0 + \pi_1 B_0, \quad \text{where} \quad \sum_{s=0}^{\infty} \pi_s \mathbf{e}^T = 1. \tag{9}$$

Define the vector generating function for π_s, $P_\Pi(z)$, as

$$P_\Pi(z) = \sum_{s=0}^{\infty} \pi_s z^s. \tag{10}$$

Furthermore, define

$$A(z) = \sum_{s=0}^{\infty} A_s z^s, \quad \text{and} \quad C(z) = \sum_{s=0}^{\infty} C_s z^s. \tag{11}$$

Multiplying both sides of (8) with the scalar z^s, where $|z| \leq 1$, and summing the result for $s = (0, \ldots, \infty)$, we obtain:

$$\sum_{s=0}^{\infty} \pi_s z^s = \sum_{s=0}^{\infty} \pi_0 C_s z^s + \sum_{s=1}^{\infty} \sum_{i=1}^{s} \pi_i A_{s-i} z^s + \sum_{s=0}^{\infty} \pi_{s+1} B_0 z^s, \tag{12}$$

and it follows that

$$P_\Pi(z) = \pi_0 C(z) + P_\Pi(z) A(z) - \pi_0 A(z) + B_0 z^{-1} P_\Pi(z) - \pi_0 B_0 z^{-1} \; . \tag{13}$$

Multiplication of (13) with z and rearranging terms gives:

$$P_\Pi(z) \left[zI - zA(z) - B_0 \right] = \pi_0 \left[zC(z) - zA(z) - B_0 \right]. \tag{14}$$

3.3 Mean Number of Regular Patients Present

We derive the mean number of regular patients in the queue, $\mathbb{E}[L_R]$, by following the analysis from [8], pp. 143-148. Let $z = 1$. First we list the relations we already have.

$$\begin{aligned} \mathbb{E}[L_R] &= P'_\Pi(1)\mathbf{e}^T \\ \mathbb{E}[L_R]\pi^\infty &= P'_\Pi(1)\mathbf{e}^T\pi^\infty \\ P_\Pi(1) &= \pi^\infty \\ P_\Pi(1)\mathbf{e}^T &= 1 \\ A(1) + B_0 &= C(1) = D \\ D\mathbf{e}^T &= \mathbf{e}^T \; , \end{aligned} \tag{15}$$

where π^∞ is the vector with the equilibrium probabilities of the number of priority patients in the queue, which can be obtained from $\pi^\infty D = \pi^\infty$. The first derivative of (14) with respect to z is

$$\begin{aligned} P'_\Pi(z) & \left[zI - zA(z) - B_0 \right] + P_\Pi(z) \left[I - A(z) - zA'(z) \right] \\ & = \pi_0 \left[C(z) + zC'(z) - A(z) - zA'(z) \right]. \end{aligned} \tag{16}$$

For $z = 1$, it follows that:

$$P'_\Pi(1) \left[I - D \right] + \pi^\infty \left[I - A(1) - A'(1) \right] = \pi_0 \left[C(1) + C'(1) - A(1) - A'(1) \right]. \tag{17}$$

Denote $\left[I - D + \mathbf{e}^T\pi^\infty \right]$ by U and $\left[I - \frac{1}{1-q_1}\bar{A}D \right]$ by K. Furthermore, note that $\left[\frac{1}{1-q_1}D - \frac{1}{1-q_1}\bar{A}D \right]$ is equal to $\bar{B}D$. By adding $P'_\Pi(1)\mathbf{e}^T\pi^\infty = \mathbb{E}[L_R]\pi^\infty$ we obtain:

$$\begin{aligned} P'_\Pi(1) & \left[I - D + \mathbf{e}^T\pi^\infty \right] + \pi^\infty \left[I - \bar{A}D - \frac{q_1}{1-q_1}\bar{A}D \right] \\ & = \mathbb{E}[L_R]\pi^\infty + \pi_0 \left[D + \frac{q_1}{1-q_1}D - \bar{A}D - \frac{q_1}{1-q_1}\bar{A}D \right] \\ \Rightarrow & \\ P'_\Pi(1) & \left[I - D + \mathbf{e}^T\pi^\infty \right] + \pi^\infty \left[I - \frac{1}{1-q_1}\bar{A}D \right] \\ & = \mathbb{E}[L_R]\pi^\infty + \pi_0 \left[\frac{1}{1-q_1}D - \frac{1}{1-q_1}\bar{A}D \right]. \end{aligned} \tag{18}$$

From Theorem 5.1.3 in [6] it follows directly that the matrix U is invertible. We then have that $\pi^\infty U^{-1} = \pi^\infty$ and thus:

$$P'_\Pi(1) \quad = \mathbb{E}[L_R]\pi^\infty + \pi_0 \bar{B} D U^{-1} - \pi^\infty K U^{-1}. \tag{19}$$

Multiplying with ^{T} it follows that:

$$\pi_0 \bar{B} D \mathbf{e}^T \quad = \quad \pi^\infty K \mathbf{e}^T. \tag{20}$$

By taking the second derivative of (14) with respect to z, setting $z = 1$ and multiplying with $\mathbf{e}^T$ we obtain:

$$\begin{aligned} P''_\Pi(1) \quad & [I - D]\,\mathbf{e}^T + 2P'_\Pi(1) K \mathbf{e}^T \\ & = \pi^\infty \left[\frac{2q_1}{(1-q_1)^2} \bar{A} \mathbf{e}^T \right] + \pi_0 \left[\frac{2q_1}{1-q_1} \bar{B} D \mathbf{e}^T \right]. \end{aligned} \tag{21}$$

Since $P''_\Pi(1)\,[I - D]\,\mathbf{e}^T = 0$ we get:

$$P'_\Pi(1) K \mathbf{e}^T = \frac{q_1}{(1-q_1)^2} \pi^\infty \bar{A} \mathbf{e}^T + \frac{q_1}{1-q_1} \pi_0 \bar{B} D \mathbf{e}^T. \tag{22}$$

Now we combine (19) and (22) to obtain an expression for $\mathbb{E}[L_R]$:

$$\begin{aligned} & \mathbb{E}[L_R]\pi^\infty K \mathbf{e}^T \\ & = \pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K \right] \mathbf{e}^T + \pi_0 \bar{B} D \left[\frac{q_1}{1-q_1} I - U^{-1} K \right] \mathbf{e}^T \\ & = \pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K \right] \mathbf{e}^T + \pi_0 \bar{B} D U^{-1} \left[\frac{3q_1 - 1}{1-q_1} \mathbf{e}^T + (\mathbf{e}^T - \mathbf{w}^T) \right] \\ & = \pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K \right] \mathbf{e}^T + \pi_0 \bar{B} \left[\frac{3q_1 - 1}{1-q_1} \mathbf{e}^T + D U^{-1} (\mathbf{e}^T - \mathbf{w}^T) \right]. \end{aligned} \tag{23}$$

Using (20) this simplifies to:

$$\mathbb{E}[L_R]\pi^\infty K \mathbf{e}^T \quad = \quad \pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K + \frac{2q_1}{1-q_1} K \right] \mathbf{e}^T - \pi_0 \bar{B} D U^{-1} \mathbf{w}^T \ , \tag{24}$$

and

$$\mathbb{E}[L_R] \quad = \quad \left[\pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K \right] \mathbf{e}^T - \pi_0 \bar{B} D U^{-1} \mathbf{w}^T \right] \left[\pi^\infty K \mathbf{e}^T \right]^{-1} + \frac{2q_1}{1-q_1} \ . \tag{25}$$

The second and higher moments of $\mathbb{E}[L_R]$ can be computed using the same approach.

In expression (25) there is still an unknown, π_0. We suggest two approximations for π_0 and thus for $\mathbb{E}[L_R]$. Since the load for regular patients is high and therefore the probability that the server is idle while there are priority patients in the queue is low, the first approximation is obtained by $\pi_0 = (1 - \rho)\pi^\infty$. The second approximation is to set $\pi_0 \bar{B} D U^{-1} \mathbf{w}^T = \mathbf{0}$. We use simulation (see Table 1) to determine which of the two approximations is most accurate in terms of the parameter values of our problem setting, i.e., a high load for regular patients ($q_1 = 0.45$) and a

Table 1: Comparing the values of $\mathbb{E}[L_R]$ that follow from the simulation and approximations

Case	L	H	ρ_S	$\mathbb{E}[L_R]$			$\lvert\delta\rvert$ with sim.	
				Sim.	Approx. 1	Approx. 2	Approx. 1	Approx. 2
1	1	1	0.9171	10.1	10.1	10.9	0.0	0.8
2	1	3	0.9250	11.0	11.2	12.0	0.2	1.0
3	1	5	0.9270	11.5	11.5	12.3	0.0	0.8
4	3	3	0.9171	9.9	10.1	10.9	0.2	1.0
5	3	5	0.9250	11.2	11.2	12.0	0.0	0.8
6	5	5	0.9170	10.0	10.2	10.9	0.2	0.7

low to moderate load for priority patients ($q_2 = 0.10$). The slot pick probability p_h is uniform distributed. We also give the load ρ_S that follows from the simulation.

The mean number of regular patients in the queue in the simulation, $\mathbb{E}[L_R]_S$, was calculated by simulating a period of 100,000 slots (so that there would be $\approx$ 10,000 priority patient arrivals), preceded by a warm-up period of 1,000 slots. When in run n,

$$\left| \frac{\sum_{i=1}^{n} \mathbb{E}[L_R]_{S,i}}{n} - \frac{\sum_{i=1}^{n-1} \mathbb{E}[L_R]_{S,i}}{n-1} \right| < \epsilon, \tag{26}$$

the simulation would stop. For ϵ a value of e^{-1} was chosen, which corresponds in the case of ten minute slots to an error margin of one minute. We see that the first approximation is the most accurate with a maximum error in the six test cases of 0.2 (2 minutes).

3.4 Mean Waiting Time for Regular Patients

Even though the regular patients may experience additional delay when a priority patient takes their spot, the mean waiting time for regular patients, $\mathbb{E}[W_R]$, can still be calculated using Little's law. This is because the queuing discipline for the regular patients is FCFS and therefore the order in the queue for regular patients does not change when a priority patient arrives and picks a slot in the appointment window. The mean waiting time is therefore equal to the sojourn time, which is calculated using the mean number of regular patients present, $\mathbb{E}[L_R]$, and the mean throughput of regular patients per slot, ρ_1, minus one slot:

$$\mathbb{E}[W_R] = \frac{\mathbb{E}[L_R]}{\rho_1} - 1, \tag{27}$$

where $\rho_1 = \frac{q_1}{1-q_1}$.

4 Results

To generate the results presented in this section, we use the first (most accurate) approximation of $\pi_0 = (1-\rho)\pi^{\infty}$. We use the same parameter values as in the previous section, i.e., $q_1 = 0.45$, $q_2 = 0.10$.

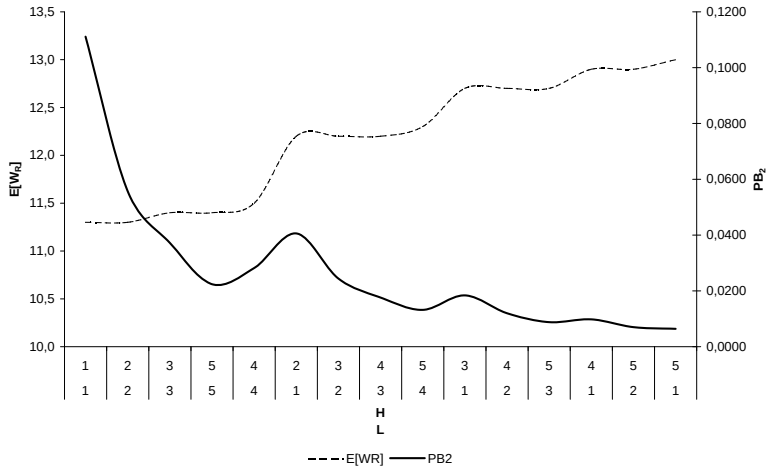
4.1 The Effect of the Size and Position of the Appointment Window

In Table 2 we see the effect of the size and position of the appointment window on the waiting time for regular patients, $\mathbb{E}[W_R]$, and the blocking probability for priority patients, $\mathbb{P}_{B_2}$. As is also apparent from Figure 6, $\mathbb{E}[W_R]$ increases and $\mathbb{P}_{B_2}$ decreases when the appointment window becomes larger and is positioned further away from the first position in the queue.

Table 2: Results for various positions and sizes of the appointment window

L	H	$\mathbb{E}[L_R]$	$\mathbb{E}[W_R]$	$\mathbb{P}_{B_2}$
1	1	10.1	11.3	0.1111
1	2	10.8	12.2	0.0406
1	3	11.2	12.7	0.0184
1	4	11.4	12.9	0.0098
1	5	11.5	13.0	0.0064
2	2	10.1	11.3	0.0557
2	3	10.8	12.2	0.0244
2	4	11.2	12.7	0.0120
2	5	11.4	12.9	0.0070
3	3	10.1	11.4	0.0372
3	4	10.8	12.2	0.0176
3	5	11.2	12.7	0.0088
4	4	10.2	11.5	0.0282
4	5	10.8	12.3	0.0132
5	5	10.2	11.4	0.0224

Figure 6: Waiting time for regular patients, $\mathbb{E}[W_R]$, versus blocking probability for priority patients, $\mathbb{P}_{B_2}$



4.2 Comparison with the Non-Priority Queue

We compute $\mathbb{E}[L_R]$ for the same queuing system, but now the queue discipline is FCFS for both regular and priority patients (we still refer to priority patients, even though these (care pathway) patients do not have priority anymore), and there is no blocking of priority patients. The expected number of patients at the facility, $\mathbb{E}[L]$, is given by $\lim_{z\to 1} P'_{\Pi}(z)$, where it is easy to derive that $P_{\Pi}(z)$ in this case is given by:

$$P_{\Pi}(z) = (1-\rho)\frac{G(z)(1-z)}{G(z)-z}, \tag{28}$$

so that

$$\mathbb{E}[L] \quad = \quad (1-\rho)\frac{2G'(1)(1-G'(1)) + G''(1)}{2(G'(1)-1)^2}. \tag{29}$$

In case of absence of the priority patients we have that $G'(1) = \rho_1$ and $G''(1) = \rho_1^2$, and thus $\mathbb{E}[L] = \mathbb{E}[L_R] = 2.7$ (note that ρ in (29) is equal to ρ_1). If the priority patients also arrive at

the facility, $G(z)$ is the product of the two probability generating functions of the independent geometric arrival processes, and thus $G'(1) = \rho_1 + \rho_2$, and $G''(1) = 2\rho_1^2 + 2\rho_1\rho_2 + 2\rho_2^2$ (note that ρ in (29) is equal to $\rho_1 + \rho_2$). We obtain $\mathbb{E}[L] = 11.9$, and $\mathbb{E}[L_R] = \frac{\rho_1}{\rho_1+\rho_2}\mathbb{E}[L] = 10.5$, $\mathbb{E}[W] = 11.8$. So even though priority patients are not blocked, the mean waiting time for regular patients is shorter. Only a priority discipline where a single slot is reserved for priority patients results in a slightly shorter waiting time for regular patients (see Table 2).

5 Discussion

In this paper we analyzed the single server queue in discrete time with two types of patients. Both patient types arrive according to a geometric arrival process and have a service requirement of 1 slot. Priority patients claim upon arrival an empty slot, h, ('appointment') in a pre-defined appointment window, and have absolute priority over regular patients. We have derived the blocking probability for priority patients and the mean waiting time for regular patients. The methodology we developed is mainly meant as a capacity planning tool, so that managers can study the effect of for instance the values of the lower and upper bound of the appointment window. In reality, a steady state situation, especially in an environment that does not offer 24/7 service such as an outpatient clinic, will maybe not be reached. However, given the managerial insights that the methodology gives, we still feel it can be very valuable in these cases.

Throughout the paper we assumed that when h was already taken, the claim of the new arrived priority patient is advanced to slot $(h-1, \ldots, L)$, until a free slot was found. It is straightforward to analyze the queue where the claims are set back to slots $(h+1, \ldots, H)$. Also the possibility to choose any distribution for the slot pick probability, p_h, introduces a lot of flexibility. The choice for the distribution of p_h will especially influence the mean waiting time for regular patients. For example, the case where $p_H = 1$, $p_h = 0 \quad \forall \quad h \neq H$, makes maximal use of the appointment window in the case that the slots are advanced when a picked slot is already claimed, and thus $\mathbb{E}[W_R]$ will be larger than in the case that $p_H \neq 1$.

The effect of increasing H gradually reduces when H becomes larger, and will lead to computational issues. Currently, the computations for $\mathbb{E}[L_R]$ using a software program such as Matlab become already quite involved for $H \approx 10$. This is not necessarily a problem and allows for analysis of many problem instances, but deserves attention in future research. The symmetry in D might be useful to simplify the analysis and size of the solution space. Note that simulation has the same computational limitations.

Of course, the size of appointment window $(L, \ldots, H)$ has a significant influence on both the priority patient blocking probability, $\mathbb{P}_{B_2}$, and the regular patient waiting time, $\mathbb{E}[W_R]$. When the window size $H-L+1$ is decreased, $\mathbb{P}_{B_2}$ will increase but $\mathbb{E}[W_R]$ will decrease. It is obvious that the trade-off between these two competing performance measures lies exactly here. A rule of thumb that comes into mind from the Subsection 4.2 and the graph in Figure 6, is that by reserving one slot for priority patients a few slots (3–5) from the first queue position, results in acceptable outcomes for both the waiting time for regular patients and the blocking probability for priority patients. However, a mean waiting time of over 11 slots (also in the case without priorities) is quite long, so the load of the system should be subject of study as well. In future research we plan to further investigate the exact trade-off and come to a rule of thumb. Furthermore, we plan to expand this research to the multi-queue variant of the problem.

References

[1] Allen D (2009) *From boundary concept to boundary object: the practice and politics of care pathway development*. Social Science & Medicine 69(3):354-361

[2] Allen D, Rixson L (2008) *How has the impact of 'care pathway technologies' on service integration in stroke care been measured and what is the strength of the evidence to support their effectiveness in this respect?* International Journal of Evidence-Based Healthcare 6(1):78-110

[3] Bruneel H (1993) *Performance of discrete-time queueing systems.* Computers Operations Research 20(3):303-320

[4] Dickson D, Ford RC, Laval B (2005) *Managing real and virutal waits in hospitality and service organizations.* Corneel Hotel and Restaurant Administration Quarterly 46(1):52-68

[5] Disney's Fastpass, Wikipedia the free encyclopedia. Retrieved from `http://en.wikipedia.org/w/index.php?title=Disney\%27s_Fastpass&oldid=442738236` on August 30, 2011

[6] Kemeny JG, Snell JL (1976) *Finite Markov Chains.* 2nd ed. Springer, New York, NY, USA

[7] Kostami V, Ward AR (2009) *Managing service systems with an offline waiting option and customer abandonment.* Manufacturing & Service Operations Management 11(4):644-656

[8] Neuts MF (1989) *Structured stochastic matrices of $M/G/1$ type and their applications.* Marcel Dekker New York, NY, USA

[9] Pullman M, Rodgers S (2010) *Capacity management for hospitality and tourism: a review of current approaches.* International Journal of Hospitality Management 29:177-187

[10] Takahashi Y, Hashida O (1991) *Delay analysis of discrete-time priority queue with structured inputs.* Queueuing Systems 8:149-164

[11] de Vuyst S, Wittevrongel S, Bruneel H (2005) *Delay differentiation by reserving space in queue.* Electronics Letters 41(9)

[12] Walraevens J, Steyaert B, Bruneel H (2002) *Delay characteristics in discrete-time $GI/G/1$ queues with non-preemptive priority queueing discipline.* Performance Evaluation 50:53-75