Stochastic Models

In Health Care

V. G. Kulkarni
Department of Operations Research
University of North Carolina
Chapel Hill, NC 27599-3180
email: vkulkarn@email.unc.edu
home page: http://www.unc.edu/∼kulkarni

November 15, 2016

Contents

Chapter 1

Blood Banks

1.1 Background

Human blood can be classified into four types: A, B, AB and O. They were discovered by Karl Landsteiner in 1901 in the University of Vienna. He received Nobel prize for this discovery in 1930. Type O is a universal donor, and type AB is a universal receiver. Furthermore, it is also differentiated by the presence or absence of the Rh factor, presence is indicated by + and absence by -. (This is an oversimplification, but is commonly used.) Thus we have eight main categories: A+, A-, B+, B-, AB+, AB-, O+, O-. Further subcategories exist, but these suffice for our modeling purposes. Whole blood, as opposed to blood components, can be safely stored for up to 45 days, after which it is too old to be useful, and hence has to be discarded.

The relative frequencies of these eight types are as given in Table 3.1.2 below:

ABO Type	Rh Type	Percentage
O	positive	37.4
O	negative	6.6
A	positive	35.7
A	negative	6.3
B	positive	8.5
B	negative	1.5
AB	positive	3.4
AB	negative	.6

Blood banks came into existence in 1930s first in USSR once the process of storing the blood was understood. The system quickly spread over the whole world. Currently blood is collected by organizations such as Red Cross from volunteers and is stored in regional blood banks after extensive testing to ensure integrity of the blood supply. Hospitals get their blood from these regional blood banks. The blood is kept in plastic bags that identify the type of the blood and when it was collected (thus giving the age of the blood). The doctors request blood from this hospital inventory for their patients. The hospital

inventory is classified as free and assigned inventory. When a surgery is scheduled, the doctor will request a certain amount of blood for his patient. The hospital blood bank will then move that much blood from the free inventory to the assigned inventory. When the surgery is over, the unused assigned blood can be transferred back to the free inventory.

The management of blood bank inventory is closely related to the general problem of managing inventory of perishable products. The earliest work on this is the PhD dissertation by Van Zyl [157] of UNC, Chapel Hill!

1.2 Issuing Policies

Suppose the doctor requests 200 ml of type A+ blood for day after tomorrow. The question arises: which blood should be moved from the free inventory to the assigned inventory? For example, we can use the oldest type A+ blood, or the youngest type A+ blood, or the oldest or youngest type O+ blood. The policy that dictates this choice is called the issuing policy.

1.2.1 Pegels-Jelmert Model

Pegels and Jelmert [104] developed one of the earliest model to answer to evaluate a given issuing policy. We describe their model, but present our own analysis. Consider a single unit of blood (say 100 ml). Let X_n be the age of this blood on day n. Let q_i be the probability that a blood unit of age i is used for transfusion. Assume that the blood is discarded once it reaches age m. Then $\{X_n, n \geq 0\}$ is a DTMC (Discrete Time Markov Chain) with state space $\{T, D, 0, 1, \cdots, m-1\}$, where state T stands for "transfused", state D stands for discarded, and state i stands for age i, $0 \leq i \leq m-1$. The transition probabilities are

$$p_{i,T} = q_i, \;\; 0 \leq i \leq m-1,$$

$$p_{i,i+1} = 1 - q_i, \;\; 0 \leq i \leq m-2, \;\; p_{m-1,D} = 1 - q_{m-1}.$$

One can think of $[q_0, q_1, \cdots, q_{m-1}]$ as an issuing policy. For example, a policy that issues fresher blood more often the old blood (called modified LIFO) will have $q_0 \geq q_1 \geq \cdots \geq q_{m-1}$, while a policy that issues fresher blood less often the old blood (called modified FIFO) will have $q_0 \leq q_1 \leq \cdots \leq q_{m-1}$. We are interested in the following three performance measures of this issuing policy.

1. δ = the probability that a fresh unit of blood is eventually discarded.
2. τ = the expected age of the transfused blood.
3. θ = the expected age of the blood in the inventory.
4. L = the expected inventory level in steady state.

Using the DTMC model we can evaluate the above performance measures associated with a given issuing policy $[q_0, q_1, \cdots, q_{m-1}]$. Suppose we start with a unit of fresh blood. That is, $X_0 = 0$. Let $A = \min\{n \geq 0 : X_n = T\}$. If $A < \infty$, it represents the age of the unit when it gets transfused, and $A = \infty$ if the unit gets discarded. We have

$$\beta_i = P(A = i) = \Pi_{j=0}^{i-1}(1 - q_j)q_i, \;\; 0 \leq i \leq m-1$$

$$\beta_m = P(A = m) = \Pi_{j=0}^{m-1}(1 - q_j).$$

Then $\delta = \beta_m$ is the probability that the blood is discarded, and $\beta_i/(1 - \beta_m)$ is the probability that the age of the transfused blood is i $(0 \leq i \leq m-1)$. The expected age of the transfused blood is given by

$$\tau = \sum_{i=0}^{m-1} i\beta_i/(1 - \beta_m).$$

To compute the expected age of the blood in the inventory in steady state, we create a modified DTMC $\{Y_n, n \geq 0\}$ with state space $\{0, 1, \cdots, m-1\}$ and transition probabilities as follow:

$$p_{i,i+1} = 1 - q_i, \;\; 0 \leq i \leq m-2,$$

$$p_{i,0} = q_i, \;\; 0 \leq i \leq m-2,$$

$$p_{m-1,0} = 1.$$

We can think of the DTMC Y as identical to the DTMC X, except that whenever the X process gets absorbed in T or D, the Y process returns to state 0. The limiting distribution of the Y process is given by

$$\alpha_i = \frac{\beta_i/q_i}{\sum_{j=0}^{m-1} \beta_j/q_j}, \;\; 0 \leq i \leq m-1.$$

The expected age of the inventory of the blood is given by

$$\theta = \sum_{i=0}^{m-1} i\alpha_i.$$

One can derive a simpler expression for θ using renewal theory. One can think of $\{Y_n, n \geq 0\}$ as the age process in a discrete renewal process with iid inter-renewal times $\{A_k, k \geq 0\}$, with common distribution that of the first passage time A above. Then the limiting expected age is given by

$$\theta = \frac{E(A^2)}{2E(A)}.$$

To compute the expected inventory level we assume λ units of fresh blood comes to the hospital each day on average. Each unit of blood stays in the hospital on

average for $E(A)$ amount of time. Hence using Little's law we see that, L, the expected inventory in steady state is given by

$$L = \lambda E(A).$$

Pegels and Jelmert also develop another DTMC model, but we won't go into that model here. The basic calculations are the same.
One can easily compute these performance measures numerically. Typically the modified LIFO policy produces smaller θ and larger δ as compared to the modified FIFO policy.

1.2.2 Optimality of FIFO Policy

The Pegels-Jelmert model did not explicitly model the issuing policy, and did not explicitly prove that FIFO policy is optimal. This was done by Pierskalla and Roach [107] in the context of perishable inventory model. We present their results here.

Suppose we operate the system for n periods. The lifetime of the items is m periods. Let D_{ij} be the demand for items of age j in period i. This can be satisfied by items of age j or less. We consider both cases: backlogging and lost sales. Let I_{ij} be the inventory of items of age j after demands are filled in period i. Let V_j be the value of one unit of age j, and assume that $V_1 \geq V_2 \geq \cdots \geq V_m$, i.e., fresher items are more valuable. Let R_i be the value of the items used in filling all demands and the remaining inventory in period i. Thus

$$R_0 = \sum_{j=1}^{m} V_j I_{0j},$$

$$R_i = \sum_{k=1}^{i} \sum_{j=1}^{m} V_j D_{kj} + \sum_{j=1}^{m} V_j I_{i,j}.$$

Thus we derive a value of V_j no matter what age item we use to satisfy the demand for an item of age j (obviously we must use an item of age j or younger). Let S_{ij} be the amount backlog for items of age j at the end of period i in the backlogging case, and the lost sales of items of age j in period i in the lost sales case. We do not make any assumptions about the distributions of D_{ij} or how the inventory is replenished.

FIFO policy satisfies the demand for item of age j by the oldest available item of age j or less, while LIFO uses the youngest available item. We use the superscript F to denote the quantities under the FIFO policy. The main results of Pierskalla and Roach [107] are summarized below.

1. $R_i^F \geq R_i$, that is FIFO policy maximizes the total utility.

2. When backlogging is allowed

$$\sum_{j=1}^{k} S_{ij} \geq \sum_{j=1}^{k} S_{ij}^{F},$$

that is FIFO minimizes stockouts of items of age k or less for all $1 \leq k \leq m$ and $1 \leq i \leq n$.

3. In the lost sales case

$$\sum_{k=1}^{i}\sum_{j=1}^{m} S_{kj} \geq \sum_{k=1}^{i}\sum_{j=1}^{m} S_{kj}^{F},$$

that is FIFO minimizes the total lost demand up to any period $1 \leq i \leq n$.

4. $\sum_{j=1}^{m} I_{ij} \geq \sum_{j=1}^{m} I_{ij}^{F}$, that is, FIFO minimizes the total inventory at the end of each period $1 \leq i \leq n$,

5. $I_{im} \geq I_{im}^{F}$, that is FIFO minimizes the number of items reaching the maximum age m (which are then discarded if not used).

1.3 Procurement Policies

A simplified version of the blood bank inventory management problem is the perishable inventory management problem. In this problem the items in the inventory have fixed lifetime m, after which they have to be discarded. Here we consider the optimal procurement policies, that is, we decide how much order. We begin with the simplest case: $m = 1$.

1.3.1 Perishable System with $m = 1$.

This is also called the news-vendor problem. At the beginning of the day the inventory manager orders $y \geq 0$ units from the supplier at a per unit cost of c, and they arrive instantaneously. Then the manager realizes a random demand D, which is a continuous random variable with pdf f and cdf F. After the demand is satisfied the manager is left with $(y - D)^+$ units on hand, and the manager pays a per unit penalty (cost) of θ for them. These units have to be discarded, since their lifetime is over. The manager also faces a shortage of $(D - y)^+$ units, for which he pays a per unit penalty (outage cost) of r. The question is: What is the optimal order quantity y^*?
The total expected discarding plus outage cost is given by

$$L(y) = \theta E((y - D)^+) + rE((D - y)^+).$$

This is a convex function of y and is minimized at

$$y^* = F^{-1}\left(\frac{r}{\theta + r}\right).$$

This is the optimal order size.

Now suppose the manager already has x units on had at the beginning of the day. The total holding plus outage cost is then given by $L(x+y)$. Using the convexity, we can show that the optimal order quantity is

$$y^*(x) = \begin{cases} y^* - x & \text{if } x < y^* \\ 0 & \text{if } x \geq y^* \end{cases}$$

This is called the order up to y^* policy, or the base-stock policy, with y^* as the base-stock level.

Now suppose we get a random supply of S_n on day n, over which we have no control, and there is a random demand of D_n on day n. Suppose the demands are identically distributed with common cdf F. Since the inventory perishes in one day, we see that the order up to y^* policy is optimal at all times. That is, at time n it is optimal to place an order of $(y^* - S_n)^+$.

1.3.2 Perishable system with $m = 2$.

Now consider the situation where the life time is $m = 2$. This model was considered by Nahmias and Pierskalla [99]. Now the state of the system is given by x, the amount of 1 day old inventory on hand at the beginning of the day. The manager then orders y amount of fresh inventory. To meaningfully account for the outdating costs we need to look at a two period problem, with D_n as the random demand in period $n = 1, 2$. We assume FIFO policy, i.e., older items are used before the newer items to satisfy the demand. The total stock out is given by $(D_1 - x - y)^+$. The total amount discarded after two periods from the y units ordered today is given by $(y - (D_2 + (D_1 - x)^+)^+$. Thus the total one period cost is given by

$$L(x,y) = rE((D_1 - x - y)^+) + \theta E((y - (D_2 + (D_1 - x)^+)^+). \tag{1.1}$$

Assuming $D_1,\ D_2$ are iid with cdf F (and pdf f), we can write the above as

$$L(x,y) = r\int_{x+y}^{\infty} (t - x - y)f(t)dt + \theta \int_0^y F(t+x)F(y-t)dt.$$

Let $y^*(x)$ be the value of y that minimizes $L(x,y)$ for a given x. The authors prove that

1. $y^{*'}(x) = -1$ for $x \leq 0$,
2. $-1 < y^{*'}(x) < 0$ for $x > 0$
3. $y^*(x) \to 0$ as $x \to \infty$.

This completes the analysis of the one period problem.

Next we consider a finite horizon problem with discount factor $0 \leq \alpha < 1$. Let $C_n(x)$ be the minimum total discounted cost over the next n periods if the starting state is x (amount of one day old inventory). If the manager orders y units of fresh inventory, and an amount t is demanded, we are left with $y-(t-x)^+$ one-day old inventory at the beginning of the next day. This reflects the assumption of FIFO issuing policy and backlogging of unsatisfied demand. Hence we see that $C_n(x)$ satisfies the following DP equation:

$$C_n(x) = \min_{y \geq 0}\{L(x,y) + \alpha \int_0^\infty C_{n-1}((y-(t-x)^+))f(t)dt\}, \tag{1.2}$$

with

$$C_0(x) = 0.$$

Let $y_n^*(x)$ be the value of y that minimizes the RHS of Equation 1.2. The authors prove that $y_n^*(x)$ satisfy the same properties as $y^*(x)$ as above.

Next we consider a more general cost structure. Let c be the cost per unit of new inventory, and h be the holding cost of inventory per unit time, in addition to the discarding cost θ and the shortage cost r. Then the one period cost can be written as

$$L(x,y) = cy+hE((x+y-D_1)^+)+rE((D_1-x-y)^+)+\theta E((y-(D_2+(D_1-x)^+)^+). \tag{1.3}$$

The DP Equation 1.2 remains valid for this cost structure if we use the L function defined above. The structure of the optimal policy is simplified if we assume that all one day old inventory at the end of the horizon can be returned at the purchase cost c, and all unsatisfied demand has to be satisfied at a per unit cost of c, i.e, we assume that

$$C_0(x) = -cx.$$

In order to get a non-trivial answer, we shall assume that $r > (1-\alpha)c$, else it does not make sense to ever hold any inventory. The main result is as follows. Let

$$\bar{x} = F^{-1}\left(\frac{r-(1-\alpha)c}{r+h}\right).$$

Then

1. $y_n^*(x) = 0$ for $x \geq \bar{x}$,
2. $y_n^{*'}(x) = -1$ for $x \leq 0$,
3. $-1 < y_n^{*'}(x) < 0$ for $0 < x < \bar{x}$,
4. $y_n^*(x) \to 0$ as $x \to \bar{x}$.

Thus if the one day old inventory is above $\bar{x}$, it is optimal not order any more inventory. Once the one day old inventory goes below $\bar{x}$ it always stays below $\bar{x}$.

1.3.3 Perishable System with $m > 2$.

Following the development in the previous subsection, we write the state of the system at time n as

$$x_n = (x_n^{m-1}, x_n^{m-2}, \cdots, x_n^1).$$

where x_n^i be the amount of inventory on hand on day n with remaining lifetime i. $(1 \leq i \leq m-1)$. After observing x_n, but before observing the demand D_n, the manager orders y amount of fresh inventory from the supplier, and it is delivered instantaneously. The inventory is used to satisfy the demand in a FIFO order. Thus $(x_n^1 - D_n)^+$ inventory is discarded at per unit cost θ, and the next day's inventory is given by

$$\begin{aligned}
x_{n+1}^1 &= (x_n^2 - (D_n - x_n^1)^+)^+ \\
x_{n+1}^2 &= (x_n^3 - (D_n - x_n^1 - x_n^2)^+)^+ \\
x_{n+1}^3 &= (x_n^4 - (D_n - x_n^1 - x_n^2 - x_n^3)^+)^+ \\
&\vdots \quad \vdots \quad \vdots \\
x_{n+1}^{m-2} &= (x_n^{m-1} - (D_n - x_n^1 - \cdots - x_n^{m-2})^+)^+ \\
x_{n+1}^{m-1} &= (y - (D_n - x_n^1 - \cdots - x_n^{m-1})^+)^+.
\end{aligned}$$

We write the above transformation as

$$x_{n+1} = s(y, x_n, D_n).$$

Let $R_n(x, y)$ be the expected amount of the order y that is ordered at the start of period n is eventually discarded if the state of the system was x. Note that this quantity is independent of n as long as $n \geq m$. The authors provide a recursive method of computing $R_m(x, y)$. Let

$$\sigma(x) = \sum_{i=1}^{m-1} x_i$$

be the total inventory. we write the one step cost as

$$L(x, y) = cy + hE((\sigma(x) + y - D_1)^+) + rE((D_1 - \sigma(x) - y)^+) + \theta R_m(x, y). \quad (1.4)$$

The following result gives the structure of the myopic policy (one period policy) in state x.

1. $L(x, y)$ is a convex function of y for each x.
2. Let $y^*(x)$ be the global minimum of L over y. If $r > c$, $y^*(x) > 0$.
3. Let
$$\bar{x} = F^{-1}\left(\frac{r-c}{r+h}\right).$$
The $y*(x) > 0$ if $\sigma(x) < \bar{x}$, and $y^*(x) = 0$ otherwise.

For multiperiod problem, we get the following DP equation for the discounted cost case:

$$C_n(x) = \min_{y \geq 0}\{L(x,y) + \alpha \int_0^\infty C_{n-1}(s(y,x,t))f(t)dt\}, \tag{1.5}$$

with

$$C_0(x) = -c\sigma(x).$$

The structural results for the optimal order quantity $y_n(x)$ are given below (assuming $r > (1-\alpha)c$):

1. Let
$$\bar{x} = F^{-1}\left(\frac{r-(1-\alpha)c}{r+h}\right).$$
The $y_n*(x) > 0$ if $\sigma(x) < \bar{x}$, and $y^*(x) = 0$ otherwise.

2. $-1 \leq y_n^{(1)}(x) \leq \cdots \leq y_n^{m-1}(x) < 0$, where $y_n^{(i)}(x)$ is the partial derivative of $y_n(x)$ with respect to x_i. Thus increasing the stocks for newer inventory will decrease the order quantity more significantly than that of older inventory.

3. If $x^{m-1} < 0$ (i.e., inventory is backlogged), it is optimal to order up to $y_n(0)$.

Remark 1: These results remain valid if we assume that the unsatisfied demand is lost, rather than backlogged.

Remark 2: These results can be easily extended to the infinite horizon discounted cost case. In this case there exists a stationary ordering policy $y^*(x)$.

Questions 1: It would be interesting to study the optimal policies if there is a random supply S_n of fresh inventory on day n that is not under the manager's control. Will order up to $y_n^*(x)$ policy remain optimal? The function $L(x,y) + \alpha \int_0^\infty C_{n-1}(s(y,x,t))f(t)dt$ is not convex in x and hence this policy may not be optimal.

Question 2: What will happen if we use Last In First Out (LIFO) policy? See Cohen and Pekelman [20].

1.3.4 Heuristic Policies

The main drawback of the optimal policies derived above is that they are hard to compute. They are also too complicated to implement since $y_n^*(x)$ is a function of the entire vector x, and not just the total inventory $\sigma(x)$. One way to get around this difficulty is to develop an approximation for $R_m(x,y)$ that depends only on $\sigma(x)$. Then $L(x,y)$ will be a function of $\sigma(x)$ and the one period myopic

policy will be a function of $\sigma(x)$ and can be used as an approximation of the optimal policy. Let F^{*m} be the m-fold convolution fo F with itself, and let

$$H(y) = \int_0^y F^{*m}(u)du.$$

Nahmias [101] show that a simple and useful bound for $R_m(x, y)$ is given by

$$H(\sigma(x)) \leq R_m(x, y) \leq H(\sigma(x) + y),$$

Using this bound in Equation 1.4, we get

$$L(x,y) = cy+hE((\sigma(x)+y-D_1)^+)+rE((D_1-\sigma(x)-y)^+)+\theta(H(\sigma(x)+y)-H(y)). \tag{1.6}$$

Next, Nahmias approximates the function $s(y, x, t)$ (from R^m to R^m) as follows:

$$s(y, x, t) = \sigma(x) + y - t - (H(\sigma(x) + y) - H(\sigma(x) + y - t)).$$

Using this in the DP equation 1.2 converts the function $C_n(x)$ into a function of $\sigma(x)$. Then we Nahmias derives a single critical number approximation z^*, where z^* is the solution to

$$(1-\alpha)c+hF(z)-r(1-F(z))+(\theta+\alpha c)F^{*m}(z)-\alpha(\theta+c)F^{*(m+1)}(z) = 0. \tag{1.7}$$

Then the approximate policy is

$$y_n^*(x) = (z^* - \sigma(x))^+.$$

Nahmias shows that Equation 1.7 has a unique solution if F has monotone likelihood ratio (or is PF_2). (Exponential, Erlang, Normal, distributions will work.) Numerical solutions have shown this to be a fairly good approximation. Nahmias also compares it with other heuristic critical number policies and shows that the z^* policy works quite well.

1.3.5 Base Stock Policy Evaluation: Cohen Model

So far we have studied the structure of the optimal ordering policies. We saw that they tend to be too complicated to implement. Hence we looked at some heuristic approximations to the optimal policies. Here we study the performance of a given ordering policy. In particular, we study the "order up to S" policy, and present the results from Cohen [19].

Let X_n^i be the amount of inventory with remaining life time i on day n. Thus the state of the system on day n is given by

$$X_n = [X_n^{m-1}, X_n^{m-2}, \cdots, X_n^1].$$

Let S be a given base-stock level. Suppose $\sigma(X_0) < S$. We order $S - \sigma(X_n)$ amount of fresh inventory on day n. Thus $\sigma(X_n) \leq S$ for all $n \geq 0$. Let D_n be

the demand on day n. It is satisfied in FIFO order from the existing stock, and unsatisfied demand is backlogged. The dynamics of the system is given by

$$X_{n+1} = s(S - \sigma(X_n), X_n, D_n), \;\; n \geq 0.$$

Suppose $\{D_n, n \geq 1\}$ are iid with common cdf F and pdf f. Then $\{X_n, n \geq 0\}$ is a DTMC. Cohen showed that it is Harris recurrent and aperiodic. Let $Z_n = (X_n^1 - D_n)^+$ be the mount of inventory that perishes on day n. Let A_n be the amount of by which the inventory decreases in period n (called the disposition). It is given by

$$A_n = D_n + Z_n.$$

The quantities Z_n and A_n are important for policy evaluation since the cost incurred on day n can be written as

$$C_n(S) = cE(A_n) + hE((S - D_n)^+) + rE((D_n - S)^+) + \theta E(Z_n).$$

The long run average cost is given by

$$K(S) = \lim_{n\to\infty} \frac{\sum_{k=1}^n C_n(S)}{n}.$$

One can show that

$$A_n = \max(S - \sum_{i=1}^{m-1} A_{n-i}, \; D_n).$$

Now define $Y_n = (A_{n-m+1}, \cdots, A_{n-1})$. It is clear that $\{Y_n, n \geq m\}$ is a DTMC on state space $[0, \infty)^{m-1}$, and is easier to analyze than the original DTMC $\{X_n, n \geq 0\}$. In any case, its limiting distribution is hard to compute analytically except when $m = 2$. In this case the DTMC is one-dimensional, and its limiting distribution is given by

$$G(y) = \lim_{n\to\infty} P(A_n \leq y) = \frac{F(y)(1 - F(S - y))}{1 - F(y)F(S - y)}, \;\; y \geq 0.$$

This matches with the expression obtained by van Zyl [157]. Using this one can show that $K(S)$ is a convex function of S.

The convexity of $K(S)$ for $m > 2$ was not settled by Cohen.

1.3.6 Base Stock Policy Evaluation: Chazan-Gal Model

Chazan and Gal [18] analyze the same model as Cohen [19], with two modifications: they assume lost sales, and discrete demand. They also observe the system after an order has been delivered but before the demand is satisfied, so that there are always S (a positive integer) items in the system. That is, we define $X_n^m = S - \sum_{i=1}^{m-1} X_n^i$ and consider the state of the system as the vector $X_n = [X_n^m, \cdots, X_n^1]$. The state space is

$$\{x = (x_m, x_{m-1}, \cdots, x_1) : \sigma(x) = S\}.$$

Let D_n be the demand on day n, and assume that $\{D_n, n \geq 0\}$ are iid non-negative integer valued random variables with pmf $p_k, k = 0, 1, 2, \cdots$. Let

$$Z_n = (X_n^1 - D_n)^+$$

be the number of items that perish (or discarded) in period n, and

$$Z(n) = \sum_{i=1}^{n} Z_i$$

be the total number of items discarded up to time n. Vector x majorizes vector y (written as $x \succeq y$) if

$$\sum_{i=1}^{j} x_i \geq \sum_{i=1}^{j} y_i, \quad i = 1, 2, \cdots, m.$$

Chanzan and Gal [18] prove the following monotonicity property:

Theorem 1.1 *If $X_0 \succeq \bar{X}_0$ then $Z(n) \geq \bar{Z}(n)$ for all $n \geq 1$. Here $\bar{Z}(n)$ is the total outdating process in a system starting with initial state $\bar{X}_0$.*

Let $Z_S(n)$ be the total outdating process in a system starting with initial state $(S, 0, 0, \cdots, 0)$. Then

Theorem 1.2 *$Z_S(n)$ is a convex function of S for any given sequence of demands $\{D_n, n \geq 1\}$.*

Thus the expected number of items that perish per unit time in steady state, namely,

$$\lim_{n \to \infty} \frac{Z_S(n)}{n}$$

is a convex function of S.

See Cohen and Pekelman [20] for the corresponding analysis with LIFO issuing policies.

1.3.7 Fixed Order Policy

Now we consider a policy that orders exactly S (a positive integer) units of fresh inventory in each period, regardless of the state of the system. We follow the analysis given in Brodheim et al. [16]. Let T_n be the total amount of inventory on hand at end of day n. It is clear that the age distribution of inventory is completely described by T_n as follows: At the end of the nth day there are S items of age i for $1 \leq i \leq \lfloor T_n/S \rfloor$ and $T_n - S\lfloor T_n/S \rfloor$ items of age $\lfloor T_n/S \rfloor + 1$. Let D_n be the demand on day n, and assume that $\{D_n, n \geq 1\}$ is a sequence of iid random variables with common pmf $p_k, k = 0, 1, 2, \cdots$ and mean

μ. Then $\{T_n, n \geq 0\}$ is a DTMC with state space $\{0, 1, \cdots, mS\}$ with transition probability matrix as given below for $S = 2$ and $m = 3$

$$P = \begin{bmatrix} a_2 & p_1 & p_0 & 0 & 0 & 0 & 0 \\ a_3 & p_2 & p_1 & p_0 & 0 & 0 & 0 \\ a_4 & p_3 & p_2 & p_1 & p_0 & 0 & 0 \\ a_5 & p_4 & p_3 & p_2 & p_1 & p_0 & 0 \\ a_6 & p_5 & p_4 & p_3 & p_2 & p_1 & p_0 \\ a_6 & p_5 & p_4 & p_3 & p_2 & p_1 & p_0 \\ a_6 & p_5 & p_4 & p_3 & p_2 & p_1 & p_0 \end{bmatrix}, \tag{1.8}$$

where

$$a_j = \sum_{i=j}^{\infty} p_i.$$

Let π be the limiting distribution of T_n.

When $m = \infty$ the above transition probability matrix looks like that of the embedded chain in a $G^X/M/1$ queue with arrivals occurring in batches of size S. Let ν be the limiting distribution of the this queue assuming that $\mu > S$ (to ensure stability). This is easier to compute than computing π (especially when $S = 1$) using matrix-geometric methods. Now let

$$\tilde{\pi}_i = \frac{\nu_i}{\sum_{j=0}^{mS} \nu_j}, \quad 0 \leq i \leq mS.$$

The authors show that $\tilde{\pi}$ is stochastically smaller than π, which in turn is stochastically smaller than ν. This provides nice bounds on π. Using these bounds one can compute the bounds on several performance measures such as (1) fraction of the demand that goes unmet, (2) fraction of the items that get discarded, (3) average inventory level, etc.

See Cohen and Pekelman [20] for the corresponding analysis with LIFO issuing policies.

1.4 Distribution Policies

Now we consider a two level blood banking system with one central blood bank supplying blood to several hospital blood banks. All the fresh blood is procured by the central bank and is allocated to the hospitals to satisfy the demand there. No cross shipment is allowed among the hospitals. The aim is to minimize the shortages and outdates.

1.4.1 Myopic Polices: Prastocos Model

Here we present the results by Prastacos [110] with our own notation. There are K hospital blood banks, and the demand at the kth blood bank on day n is

denoted by D_n^k. $\{D_n^k, n \geq 0\}$ are iid with common cdf F^k. The lifetime of the blood is m days, and each hospital blood bank issues blood according to FIFO policy. Prastacos [110] considers two classes of policies: rotation policies and retention policies. Under rotation policies all blood remaining at the end of the day in the hospitals is transferred back to the central blood bank and is available for distribution on the next day, while under retention policies all blood, once allocated to a hospital blood bank, stays there until used or discarded. The question is: how much blood should be allocated to each hospital, and what should be the composition of the blood (by age)?

Let Q_n be the fresh blood that arrives at the central bank at the beginning of day n. Let $X_n(j)$ be the blood of age j available at the central bank for distribution on day n. Let $Y_n^k(j)$ be the amount of blood of age j available at the hospital k at the beginning of day n. Since fresh blood is available on at the central bank, we have

$$X_n(0) = Q_n, \;\; Y_n^k(0) = 0.$$

Shortages cost p and outdates cost θ. Let S_n^k be the shortage and Z_n^k be the outdates at hospital k on day n. The shortage and outdating cost on day n is

$$p \sum_{k=1}^{K} E(S_n^k) + \theta \sum_{k=1}^{K} E(Z_n^k).$$

Prastacos [110] finds a myopic policy that minimizes this cost.

First we consider the rotation policies. Let T be the total blood and B be the amount of blood of age m available at the central bank for allocation. Suppose we allocate a total T^k of blood to hospital k, of which B^k is the blood of age m on day n. Then

$$E(Z_n^k) = E((B^k - D_n^k)^+) = \int_0^{B^k} (B^k - x) dF^k(x),$$

$$E(S_n^k) = E(D_n^k - T_k)^+) = \int_{T^k}^{\infty} (x - T^k) dF^k(x).$$

Prastacos [110] shows that the optimal allocations (B^{k*}, T^{k*}) are given by the solution to the following:

$$F^k(B^{k*}) = \lambda_1, \;\; \sum_{k=1}^{K} B^{k*} = B, \tag{1.9}$$

$$F^k(T^{k*}) = \lambda_2, \;\; \sum_{k=1}^{K} T^{k*} = T. \tag{1.10}$$

This is feasible because $T^{k*} \geq B^{k*}$ for all k. For example, if the demand at hospital k is exponentially distributed with mean k, the above equations imply:

$$B^{k*} = r_k B, \;\; T^{k*} = r_k T, \;\; 1 \leq k \leq K,$$

where

$$r_k = \frac{\mu_k}{\sum_{j=1}^{K} \mu_j}, \;\; 1 \leq k \leq K.$$

For the retention policies, the optimal mypoic policy is a bit more complicated but quite close to the one in rotation case. Let I_n^k be the total inventory at hospital k on day n. Let

$$T_n = Q_n + \sum_{k=1}^{K} I^k$$

be the total inventory in the region. Let T^{k*} satisfy Equation 1.10 with $T = T_n$. We say that hospital k is understocked if $I^k \leq T^{k*}$, and over stocked otherwise. It is not optimal to allocate any blood to an overstocked hospital. Suppose there r understocked hospitals. Number the hospitals so that $F^k(I^k)$ is an increasing sequence. Now compute Q^{k*} for $k = 1, \cdots, r$ by solving the following:

$$F^1(I^k + Q^{k*}) = \lambda,$$

such that

$$\sum_{k=1}^{r} Q^{k*} = Q_n.$$

It is optimal to allocate Q^{k*} amount of fresh blood to hospital k, $k = 1, 2, \cdots, r$.

Remark 1: The proof of optimality is not clear.

Remark 2: The author ignores the outdating cost in the derivation of myopic policy in the retention case, since only fresh blood is being allocated, and thus cannot affect the outdating cost. But this cannot be right. It might make more sense to account for the future outdates from the current allocation, much the same way it was done in deriving the one stage optimal procurement policies.

remark 3: These nice results fall apart if the issuing policy is LIFO. For example, the optimal policy in the FIFO case is independent of p and θ, which is no longer the case in the LIFO regime. See Prastacos [111] for more details.

1.4.2 Proportional Policies

A similar two level blood bank structure is considered by Cohen et al. [21]. Unlike the myopic policies considered by Prastacos [111], they concentrate on stationary critical number (order up-to) policies for procurement. They consider

two allocation policies: under policy I, each hospital bank gets blood of each age in proportion to the total demand from that hospital. Under policy II, each hospital gets blood of each age in a fixed predetermined proportion, regardless of the amount ordered. They first compute the expected total discounted cost over infinite horizon as function of the critical numbers and show that the cost function is convex in the critical numbers (under certain stationarity conditions). Thus the problem of finding optimal critical numbers is a convex non-linear program. The resulting policies are difficult to compute, although relatively easy to implement.

1.5 Cross-matching Policies

One main characteristic of blood transfusion process that is ignored in the above model is the stage of cross-matching. A surgeon typically orders a fixed amount of blood of particular type to be reserved for a particular day. The hospital blood bank than reserves this amount of blood (moves it from the free inventory to the cross-matched inventory). However, a surgeon typically orders more blood than is used for the patient, and later (after what is called cross-match release period) releases it back to the blood bank, where it is returned to the free inventory. The cross match release period is there to guard against the patient needing a post-operative emergency surgery in a short period. The fraction of the assigned blood that actually gets used is called the transfusion to cross-match ratio, and typically ranges from 1/4 to 2/3. It plays an important role in inventory management in blood banks.

1.5.1 Deterministic Model

We present a simple deterministic model inspired by Jagannathan and Sen [67]. Let m (a positive integer) be the life time the blood. The fresh blood is of age 1, and blood is discarded when it reaches age $m+1$. Suppose δ units of blood are needed every day. We follow an order up to δ policy. Thus there is no overstocking, since we stock exactly what we need to satisfy the cross-matching demand. A fraction p of the assigned blood is used, and the remaining blood is released after d ($1 \le d \le m$) days. We assume proportional returns, that is, a fraction p of each age group of assigned blood is returned.
Let x_{in} be the amount of blood of age i at the beginning of day n in the free inventory, after the order has arrived, and before the blood is assigned. Let $x_n = [x_{1n}, x_{2n}, \cdots, x_{mn}]'$. Then the dynamics of the system is described by the equation

$$x_{1,n} = \delta - (1-p) \sum_{i=1}^{m-d} x_{i,n-d}, \tag{1.11}$$

$$x_{kd+1,n} = (1-p) x_{(k-1)d+1,n-d}, \quad k = 1, 2, \cdots, [m/d], \tag{1.12}$$

where $[m/d] = \text{floor}((m-1)/d)$. All other $x_{i,n} = 0$. On the nth day we use $p\delta$ units of blood, and release $(1-p)\sum_{i=1}^{m-d} x_{i,n-d}$ to the free inventory after d

days. Total outage on day n is $x_{[m/d]d+1,n}(1-p)$. Letting $n \to \infty$ in Equations 1.11 and 1.12 and simplifying, we get

$$x_1 = \lim_{n\to\infty} x_{1,n} = \frac{\delta p}{1-(1-p)^{\lceil m/d \rceil}}, \tag{1.13}$$

$$x_{kd+1} = \lim_{n\to\infty} x_{1,n} = \frac{\delta p(1-p)^k}{1-(1-p)^{\lceil m/d \rceil}}. \tag{1.14}$$

Hence the limiting outage per day is given by

$$w = \frac{\delta p(1-p)^{\lceil m/d \rceil}}{1-(1-p)^{\lceil m/d \rceil}}. \tag{1.15}$$

This matches Theorem 1 of Jagannathan and Sen [67] when m/d is an integer. Otherwise, this results is exact, while that in [67] is approximate. The average age of the transfused blood is given by

$$\sum_{i=1}^{m} i x_i.$$

What happens if the transfusion policy is FIFO, that is the surgeon uses the oldest p fraction of the assigned blood? In this case we find it more convenient to use the following notation:

$$y_{in} = \sum_{j=i}^{m} x_{jn}, \quad 1 \le i \le m.$$

Thus y_{in} is the free inventory (before assignment) of age i and higher. All the free inventory gets assigned, and FIFO transfusion policy implies that the oldest $p\delta$ of that gets used. Thus after transfusion on day n we are left with $z_{i,n} = \max(y_{i,n} - p\delta)$ amount of inventory of age i or more. This is returned on day $n+d$ as inventory that is aged by d days. Thus z_{m-d+1} amount of inventory is discarded since it is now too old. The rest is added back to the free inventory. Thus we get

$$y_{i,n+d} = z_{i-d,n} - z_{m-d+1,n}, \;\; d+1 \le i \le m.$$

Then the order up to δ procurement policy implies that $y_{1,n+d} = \delta$ for all $n \ge 0$. Hence we get

$$y_{i,n+d} = y_{d+1,n+d}, \;\; 2 \le i \le d.$$

This completes the description of the dynamics of the system. Note that the system states for d unrelated sequences $\{Y_{kd+r,n}, k \ge 0\}$ for $r = 0, 1, 2, \cdots, d-1$. One can show that each of these subsequences can show periodic behavior.

Question: Can you identify the periodic limiting behavior?

Question: What happens if the surgeon uses the LIFO policy?

Now suppose we order up to $y > \delta$. Now of the free inventory we assign δ inventory. Thus the assigned inventory on day n of age i or more is given by

$$z_{i,n} = \min(y_{i,n}, \delta).$$

This reduces the free inventory to $\max(y_{i,n} - \delta, 0)$. Of the assigned inventory we use up $p\delta$ in a FIFO order. Hence we are left with $w_{i,n} = \max(z_{i,n} - p\delta, 0)$ unused inventory of age i or more that is returned after d days. Then $w_{m-d+1,n}$ inventory is discarded as too old, and the rest is added to the free inventory on day $n + d$. All this yields:

$$\begin{aligned} y_{1,n+d} &= y, \\ y_{i,n+d} &= \max(y_{i,n+d-1} - \delta, 0) + w_{d+1,n} - w_{m-d+1}, \;\; 2 \le i \le d, \\ y_{i,n+d} &= \max(y_{i,n+d-1} - \delta, 0) + w_{i-d,n} - w_{m-d+1}, \;\; d+1 \le i \le m. \end{aligned}$$

One can show that this also shows a periodic behavior as a function of n. However, we do not have d disconnected sequences, since they are connected through the unused free inventory.

Jagannathan and Sen [67] consider a mixed case when $y > \delta$: the cross-matching follows FIFO policy but transfusion follows proportional policy. In this case they show that fresh blood does not get matched for $(y - \delta)/(p\delta + w)$ days, where w is the outdating rate as given in Equation 1.15.

1.5.2 Stochastic Demand.

Next we consider a stochastic models of same scenario. There are many ways to introduce stochasticity in the model: one can make the demands random, or the released amount random, or the release duration random. Jagannathan and Sen [67] consider a specific stochastic model. We present their model and results here first:

Suppose D_n is the cross-matching demand on day n, where $\{D_n, n \ge 0\}$ is a sequence of non-negative iid random variables with common mean δ. We follow a base-stock policy of always ordering up to y. Assume proportional release model with transfusion ratio p. Let R_n be the quantity ordered on day n to bring the inventory up to y. Let w be the mean outdates on any given date. Then we get

$$R_n = (D_n - (D_{n-d}(1-p) - w)). \tag{1.16}$$

Hence

$$\nu = E(R_n) = p\delta + w, \; \sigma^2 = \mathrm{Var}(R_n) = \sigma_D^2(1 + (1-p)^2).$$

Let Λ be the number of days fresh blood stays in the bank before getting cross-matched (due to FIFO issuing policy). Then, using first passage times in Brownian motion, Jangannathan and Sen derive the following approximations:

$$E(\Lambda) = (y - \delta)/\nu, \quad \mathrm{Var}(\Lambda) = (y - \delta)\sigma^2/\nu^3.$$

They also provide a method approximating the mean outdate w, and the expected shortage $E(y - D_n)$. They show by simulation that their approximation is reasonable.

Remark 1: The analysis approximate and not very convincing. For example, the appearance of a deterministic quantity w in Equation 1.16 is really inexplicable.

Next we describe two other stochastic models.

1.5.3 Binomial Release Model

We assume that the demand is a fixed positive integer δ, and we order up to δ in each period. Thus all inventory is assigned in each period. Each unit is used for transfusion with probability p, or released with probability $1 - p$ after d days. Let X_i^n be the number of units of age i that are assigned on day n. Let $k = \lfloor (m-1)/d \rfloor$. We see that

$$X_i^n = 0 \quad \text{if } i \neq jd + 1$$

$$X_{jd+1}^n \sim \text{Bin}(X_{(j-1)d+1}^{n-d}, (1-p)), \quad \text{for} j = 1, 2, \cdots, k$$

$$X_1^n = \delta - \sum_{j=1}^{k} X_{jd+1}^{n-d}.$$

Let X_i be the limit in distribution of X_i^n. We can show that the marginal distributions of X_i are as follows

$$X_1 \sim \text{Bin}(\delta, p/(1-(1-p)^{k+1})),$$

$$X_{jd+1} \sim \text{Bin}(\delta, p(1-p)^j/(1-(1-p)^{k+1})), \;\; 1 \leq j \leq k.$$

Note that $E(X_i)$ matches the values of x_i derived in Equation 1.12 and the expected outage $w = E(X_{kd+1})(1-p)$ matches Equation 1.15.

Question: What happens if the demand on day n is a random variable D_n?

1.5.4 Proportional Outdating, State-independent Procurement

Here we present a model that uses a one-dimensional stochastic process rather than a d-dimensional one. Suppose we always order a fixed amount y of fresh blood in each period, and the demand on day n is D_n. Let X_n be the amount of blood inventory of all ages at the beginning of day n. We assign D_n amount (creating backlog if needed). Of that a fraction p is transfused, a fraction r is outdated and the remaining fraction $1 - p - r$ is returned to the inventory after d days. Hence we have

$$X_{n+1} = X_n + y - D_n + (1 - p - r)D_{n-d}.$$

Note that X_n is allowed to go negative. We implicitly assume that excess demand is satisfied from an emergency source. Solving recursively, we get

$$X_{n+1} = X_1 + ny - (p+r)\sum_{m=1}^{n-d} D_m - \sum_{m=n-d+1}^{n} D_m.$$

Thus,

$$E(X_{n+1}) = E(X_1) + n(y - (p+r)\delta) - d\delta, \quad n > d.$$

Thus $E(X_{n+1})$ increases linearly if $y > (p+r)\delta$, and decreases linearly in n if $y < (p+r)\delta$. If $y = (p+r)\delta$, we have

$$E(X_{n+1}) = E(X_1) - (p+r)d\delta, \quad n > d.$$

Thus the mean is independent of n. However, the variance increases linearly in n, since

$$\mathrm{Var}X_{n+1} = n\sigma^2(p+r) + d\sigma^2(1-p-1), \quad n > d.$$

Thus, $\{X_n, n \geq 1\}$ is either transient or null recurrent.

Another possibility is to assume that excess demand is lost. In this case we have the following recursion:

$$X_{n+1} = (X_1 + ny - (p+r)\sum_{m=1}^{n-d} D_m - \sum_{m=n-d+1}^{n} D_m)^+.$$

Thus any unsatisfied demand is satisfied from an emergency source, and the unused portion of the blood is returned to the inventory (whether it is from the inventory or the emergency source).

1.6 Multi-Class Policies

One prominent aspect of blood bank operation not adequately addressed in the literature is the possibility of using type O blood for any class, or satisfying the demand for type AB blood from any class. There is work in supply chain literature about multi-class inventory models, but it generally does not include out-dating and cross matching. So there is a large open area in this direction where further research can be done.

Chapter 2

Organ Transplantation

2.1 Background

The most common organs that are transplanted are: heart, kidneys, liver, lungs, and pancreas. Organ transplantation became routine since the discovery of immunosuppressant drugs that resolved the most vexing problem of transplant operation: the body can reject the transplanted organ as a foreign material.

There are many non-medical issues involved in organ transplantation:

1. Ethical: how to define death (so that organs can be removed), how to obtain permission for organ donation, equity among different patient groups, etc.
2. Economic: how to pay for organs, how to prevent black-market for organs, etc.
3. Logistical: how to store the organs, how to match the supply and demand, efficiency (finding the best match), etc.

We shall concentrate on the last topic.

In 1984 congress passed NOTA: National Organ Transplantation Act, which brought into existence UNOS: United Network for Organ Sharing (www.unos.org). It runs the OPTN (Organ Procurement and Transplantation Network), the national clearing house for organs procured by the local OPOs: Organ Procurement Organizations. There are currently 112 OPOs and they form the Association of OPOs (www.aopo.org). These 112 OPOs are organized into 11 regions, see optn.transplant.hrsa.gov/latestData/stateData.asp?type=region. When an organ becomes available at an OPO, it is first made available to the patients in that locality, if no match is found locally, it is made available regionally, and if there is no regional match, it is made available nationally.

Clearly, we need to define what we mean by a "match". An organ can be transplanted only if the donor and recipient are blood-type compatible (RH-factor compatibility is not needed,) the recipient is not pre-sensitized (will not reject the organ), the physical state of the patient is conducive to high probability of success, and the tissue types of the donor and the recipient match (this is different than the blood type matching). UNOS recommends an algorithm to take all these factors into account and rank the patients on the waiting list by assigning them priority points. The higher the priority point, the higher the rank of the patient. A brief description of the allocation algorithm is as follows (See Zenios [154]). First, the organ is offered to an identical blood-type zero-antigen-mismatched local patient, then regionally and then nationally. Then it is offered to a blood-type compatible zero-antigen mismatched patient using the same geographic hierarchy. Finally, the organ is offered to all other blood-type compatible candidates ranked according to their total number of points. Note that the recipient can refuse to accept the organ when it is made available to him/her by the allocation algorithm. In that case the organ is made available to the next person on the ranked list.

There is a serious shortage of organs available for transplants, and currently there is a long waiting list of patients waiting for an organ. For example, (see optn.transplant.hrsa.gov/data/default.asp) as of July 16, there were 92,729 patients waiting for Kidneys, 1,270 for Pancreas, 15,967 for Liver, and 3,185 for Heart. See optn.transplant.hrsa.gov/latestData/viewDataReports.asp for a large and interesting source of data. It shows how many organs became available, how many got transplanted, how long is the waiting list, classified according to to year, region, organ, and several other criteria.

We shall look at various analytical models that have been proposed in the OR literature to analyze the organ transplant problem. Zenios [154] is a good reference paper for this area.

2.2 Choice Model: Single Patient.

A significant fraction of the organs are rejected by the first patient that they are offered. In this section we study models of decision making by a single patient.

2.2.1 David-Yechiali Model

David and Yechiali [27] were the first to develop the stylized model the decision process involved. We describe it below.

Suppose a patient joins the waiting list for kidney transplant at time 0, and has a random lifetime T (time spent on dialysis), with cdf G. Suppose kidney offers arrive according to a renewal process with iid inter-arrival times with cdf H. Let X_n be the value of the n offer. This could be the probability of success-

ful transplantation if the he accepts the nth offer. Assume that $\{X_n, n \geq 0\}$ are iid non-negative random variables with common cdf F. If he accepts the offer of value X at time t, he gets a net reward of $\beta(t)X$, where $\beta(t)$ is a decreasing function of t representing diminishing value of waiting. For example $\beta(t) = e^{-\beta t}$ represents the continuous discounting with discount factor β. If no offer is accepted up to time T, then he gets a reward of 0.

We first consider a simple case where T is $\exp(r)$, the kidneys arrive according to a $\mathrm{PP}(\mu)$ and the continuous discount factor is β. This implies that the optimal policy is stationary: when a kidney of value x arrives, and the patient is still waiting for a kidney, the decision to accept or reject the kidney depends only on x. Let $V(x)$ be optimal discounted value if the patient has not yet accepted a kidney, and one of value x becomes available. The value function V satisfies the DP (Dynamic Programming) equation

$$V(x) = \max\{x, \gamma\},$$

where

$$\gamma = \frac{\mu}{r + \mu + \beta} \int_0^\infty V(y) dF(y).$$

One can show that γ is given by the unique solution to

$$(\beta + r)\gamma = \int_\gamma^\infty (1 - F(y)) dy.$$

For example, when $X \sim U(0, 1)$, we get

$$\gamma = \alpha/(1 + \sqrt{1 - \alpha^2}),$$

where $\alpha = \mu/(\mu + r + \beta)$. Thus the optimal policy is to accept the first kidney that comes by with value greater than γ.

Now we consider the case of general G. Let

$$\bar{G}(s|t) = P(T > s + t | T > t).$$

Let $V(t, x)$ be the maximum reward the patient can obtain if he has not accepted a kidney up to time t and a kidney of value x becomes available at time t. The value function V satisfies the following DP equation:

$$V(t, x) = \max\{\beta(t)x, \lambda(t)\},$$

where

$$\lambda(t) = \int_{s=0}^\infty \bar{G}(s|t) \int_{y=0}^\infty V(t + s, y) dF(y) dH(s).$$

The optimal policy is to accept the offer of value x at time t if $\beta(t)x \geq \lambda(t)$, and reject it otherwise.

Now suppose G is an IFR (increasing failure rate) distribution, that is, $\bar{G}(s|t)$ is a decreasing function of t for all s. Alternatively, the hazard rate $r(t) = g(t)/(1 - G(t))$ is an increasing function of t. Then one can show that $\lambda(t)$ above is a decreasing function of t. In fact, the IFR assumption is necessary for this. Thus, if the patient rejects a kidney of value x at time t, he will reject at all later times, and if accepts it at time t, he will accept it at all earlier times.

David and Yechiali [27] show that in this case the function $\lambda(\cdot)$ satisfies the following differential equation:

$$\lambda'(t) = r(t)\lambda(t) - \mu\beta(t) \int_{\lambda(t)/\beta(t)}^{\infty} (1 - F(x))dx.$$

For example, suppose G is a Gamma(2,θ) distribution with mean 5 years (average time on dialysis). Suppose the kidneys arrive according to a Poisson process with rate of 6.4 per year, and belong to five types A, B, C, D, and E with probability .0094, .0941, .3134, .4073, and .1758 respectively. The probability of one year survival from these kidney types is .44, .47, .49, .62 and .70 respectively. Thus the distribution the random variable X is given by

$$P(X = .44) = .1758, \;\; P(X = .47) = .4073, \;\; P(X = .49) = .3134,$$

$$P(X = .62) = .0941, \;\; P(X = .70) = .0094.$$

The optimal policy can be shown to be (i) From 0 to 1.83 years of "dialysis age"-wait for A or B - match, (ii) From 1.83 to 8.05 years-wait for A or B or C-match, (iii) Beyond 8.05 years-wait for A or B or C or D-match. Never accept an E match.

2.2.2 Ahn-Hornberger Model

The David and Yechiali model assumes that the patient is always active, that is, she is available to accept an organ if a good one becomes available. However, in practice the patient state keeps changing, and in some states she may not be able to use an excellent organ even if it becomes available (she could be running a fever, or on vacation).

We describe a semi-Markov decision process model inspired by Ahn and Hornberger [2] that accounts for such a situation. Suppose the patient state at the beginning of each month can be classified as: 1. Actively waiting for transplant, 2. Inactive (ineligible to receive transplant), 3. Received a transplant, 4. Transplant failed, 5. Death. In states 1, 2 and 4 we observe the patient every month. Once the patient enters state 3, we observe the patient next when he moves to state 4 or 5. The decision to accept a transplant can be made only in state 1. Let X_n be the state of the patient at the nth observation epoch. Note that we have two decisions {1 = Accept a transplant organ, 2 = reject a transplant organ} in state 1. In other states we do not have a decision. Let Y_n

be the quality of the kidney that becomes available at the beginning of the nth month. If this kidney is accepted, the probability of one year survival is Y_n. Let f be the pdf of Y_n over $[0,1]$. Next we associate a quality-of-life score q_i with state i. A typical assessment is given below:

$$q_1 = .65,\ q_2 = .60,\ q_3 = .90,\ q_4 = .60,\ q_5 = 0.$$

If a kidney is not accepted, the $\{X_n, n \geq 0\}$ process changes states according to the transition probability matrix:

$$\begin{bmatrix} p_{11} & p_{12} & 0 & 0 & p_{15} \\ p_{21} & p_{22} & 0 & 0 & p_{25} \\ 0 & 0 & 0 & p_{34} & p_{35} \\ p_{41} & 0 & 0 & p_{44} & p_{45} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

If a kidney is accepted, the $\{X_n, n \geq 0\}$ process changes states according to the transition probability matrix:

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ p_{21} & p_{22} & 0 & 0 & p_{25} \\ 0 & 0 & 0 & p_{34} & p_{35} \\ p_{41} & 0 & 0 & p_{44} & p_{45} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Let $V(1,y)$ be the optimal QALE (Quality-adjusted-life-expectancy) if the starting state is 1 and a kidney of quality y becomes available. $L(y)$ be the expected lifetime if a kidney of quality y is transplanted, at the end of which either the kidney fails with probability p_{34} or the patient dies with probability p_{35}. Let $V(i)$ be the maximum QALE if the starting state is i, and the optimal policy is followed, $i = 2, 4, 5$. (Clearly, $V(5) = 0$.) Suppose τ_i is the expected time spent in state i until the next kidney becomes available, or the patient dies. Then we have

$$\begin{aligned} V(1,y) &= \max\{V(3,y),\ \ q_1 + p_{11}\int_0^1 V(1,y)f(y)dy + p_{12}V(2)\}, \\ V(2) &= q_2 + p_{21}\int_0^1 V(1,y)f(y)dy + p_{22}V(2), \\ V(3,y) &= q_3 L(y) + p_{34}V(4), \\ V(4) &= q_4 + p_{41}\int_0^1 V(1,y)f(y)dy + p_{44}V(4), \\ V(5) &= 0. \end{aligned}$$

All the data needed for the above can be estimated from available sources. See references in Ahn and Hornberger [2]. The optimal policy is described by a single number d, so that the patient accepts a kidney if and only if its quality is at least d, and the patient is in state 1. One can compute this d by value

iteration. Ahn and Hornberger [2] present a more direct approach. Note that the special particulars of the patient and the kidneys are all encapsulated in the function $f(y)$.

Recommended for presentation: Davies et al [31].

2.3 Choice Model: Multiple Patients

In the previous section we studied the decision making situation of a single patient facing an infinite stream of organ offers. Now we consider the situation where there are n patients who face a stream of n organs sequentially.

2.3.1 Derman-Lieberman-Ross Model.

We present a model studied by Derman, Lieberman and Ross [33] in the organ transplant context. Suppose there are n patients. Let p_k be the probability that if an organ is given to patient k, she will have a successful transplant. We get a stream of n organs sequentially, with values $X_1, X_2, \cdots, X_n$ that are iid random variables with cdf G. If the ith organ is matched to the kth patient, we derive a value of $p_k X_i$. The aim is to maximize the total value derived from the n matches.

As a first step, suppose all organs are available simultaneously. First rank the patients so that $p_1 \leq p_2 \leq \cdots \leq p_n$ and the n organs so that $X_1 \leq X_2 \leq \cdots \leq X_n$. Then it is optimal to match the ith organ to the ith patient. The total value derived is

$$\sum_{i=1}^{n} p_i X_i.$$

This follows from Hardy's Theorem.

The problem is more complicated if the organs arrive sequentially. In this case Derman et al show that there are critical points

$$-\infty = a_{0,n} \leq a_{1,n} \leq \cdots \leq a_{nn} = +\infty$$

so that when the first organ arrives with value $X_1 = x \in (a_{i-1,n}, a_{in}]$ it is optimal to assign the organ to the ith patient ($i = 1, 2, \cdots, n$), and remove that patient from further consideration. Then we proceed with the remaining $n-1$ patients, and so until all patients are matched. This can be proved by induction. The constants a_{in} can be computed recursively as follows:

$$a_{0n} = -\infty,\ a_{nn} = \infty$$

for $i = 1, 2, \cdots, n$,

$$a_{i,n+1} = \int_{a_{i-1,n}}^{a_{in}} x dG(x) + a_{i-1,n} G(a_{i-1,n}) + a_{in}(1 - G(a_{in}).$$

(Interpret $\pm\infty \cdot 0 = 0$.) In a later paper, Albright and Derman [4] prove the following asymptotics, that makes the computation of a_{in} easy:

$$\lim_{n\to\infty} n(1 - G(a_{i,n})) = i, \quad 1 \le i \le n.$$

This model assumes that all the patients are present at time 0, and kidneys arrive sequentially. What happens if the patients also arrive sequentially? This case is studied by Righter [116].

Recommended for presentation: Righter [116].

2.3.2 Albright Model

Albright [3] considers a combination of the Derman-Lieberman-Ross [33] model and the David-Yechiali [27] model. We present a special case of his results. In order for this problem to make sense, we must use discounting, else it will be best to wait for the best organ, since one will eventually arrive.

First consider the case of n identical patients waiting for the organs that arrive according to PP(μ). The value of the kth arriving organ is X_k, where $\{X_k, k \ge 1\}$ are iid with common cdf F. If the k organ is used for transplant, we get a reward of X_k. Let $V_n(x)$ be the total expected discounted reward (continuous discount factor $\beta > 0$) if there are n patients still waiting and the value of the currently available organ is x. $V_n(x)$ satisfies the following DP equation:

$$V_n(x) = \max\{x + \frac{\mu}{\mu+\beta}\int_0^\infty V_{n-1}(y)dF(y), \frac{\mu}{\mu+\beta}\int_0^\infty V_n(y)dF(y)\}, \quad V_0(x) = 0.$$

This can be rewritten as

$$V_n(x) = \max\{x, \gamma_n - \gamma_{n-1}\} + \gamma_n$$

where

$$\gamma_n = \frac{\mu}{\mu+\beta}\int_0^\infty V_n(y)dF(y).$$

Let $y_n = \gamma_n - \gamma_{n-1}$. Thus the optimal strategy when there are n patients waiting is to accept an organ of value x if $x \ge y_n$, and reject it otherwise. The y_n's are decreasing and can be recursively computed as follows:

$$y_1 + y_2 + \cdots + y_n = \frac{\mu}{\beta}\int_{y_n}^\infty (1 - F(y))dy, \quad n \ge 1. \tag{2.1}$$

Albright [3] extends this analysis to renewal arrivals.

Next we consider the case where the patients are not identical. We use the reward structure from Derman et al [33]. The patient i has a parameter p_i

($p_1 \leq p_2 \leq \cdots \leq p_n$). If an organ of value x is assigned to patient i we get a reward of $p_i x$. Albright proves the following structure of the optimal policy. When an organ of value $x \in (y_{n-i+1}, y_{n-i}]$ ($y_0 = \infty$) arrives, assign it to patient i. If $x \leq y_n$, reject it. (These are the same y_n's as computed using Equation 2.1!)

2.3.3 Righter Models.

Righter [114] calls the above model A. One can interpret the continuous discounting as follows: there is no discounting, but the process ends (say all the unassigned customers die) after an exp(β) amount of time. She then considers the case where each customer has an exp(β) life-time, independent of all other customers. This is called model B. Let $V_n(x)$ be the total expected reward if there are n (with identical p_i's, iid lifetimes) patients still waiting and the value of the currently available organ is x. Also, let V_n be the total reward expected if initially there are n patients available. Then, $V_0 = 0$ and

$$V_n = \frac{\mu}{\mu + n\beta} \int_0^\infty V_n(y) dF(y) + \frac{n\beta}{\mu + n\beta} V_{n-1}.$$

Furthermore, $V_n(x)$ satisfies the following DP equation:

$$V_n(x) = \max\{x + V_{n-1}, V_n\}, \quad V_0(x) = 0.$$

Thus the optimal policy is to accept the organ of value x if $x \geq y_n$, where y_n can be shown to be given by

$$y_1 + y_2 + \cdots + y_n = \frac{\mu}{n\beta} \int_{y_n}^\infty (1 - F(y)) dy, \;\; n \geq 1. \tag{2.2}$$

If the success probabilities of these n customers are $p_1 \leq p_2 \leq \cdots \leq p_n$, then when an organ of value $x \in (y_{n-i+1}, y_{n-i}]$ ($y_0 = \infty$) arrives, it is optimal to assign it to patient i. If $x \leq y_n$, reject it.

In a follow up paper Righter [115] considers the case where the patient state changes stochastically and derives the control limit policies and many structural properties of the control limits. Here we present a very simplified version of that analysis that is appropriate for our case.

Suppose the state of each patient alternates between two states (1=active, 2=inactive) according to independent identically distributed CTMCs (Continuous Time Markov Chains) with rates $q_{12} = \theta_1$ and $q_{21} = \theta_2$. In state i the death rate is β_i. The patient can accept a transplant in state 1 but not in state 2. As a first step assume that the reward of accepting an organ of value x is x for all patients. We say that the system is in state (i, j) if there i patients in active state and j in inactive state. As before, let $V(i, j)$ be the total expected reward if the initial state is (i, j). Let $V(i, j, x)$ be the maximum total reward

that can be obtained in state (i,j) if an organ of value x is available. Then, we have $V(0,0)=0$ and

$$V(0,j)=\frac{1}{j\theta_2+j\beta_2+\mu}[j\theta_2 V(1,j-1)+j\beta_2 V(0,j-1)+\mu V(0,j)],\ \ j\geq 1$$

$$\begin{aligned}V(i,j) &= \frac{1}{i(\theta_1+\beta_1)+j(\theta_2+\beta_2)+\mu}[j\theta_2 V(i+1,j-1)+j\beta_2 V(i,j-1)\\ &\quad +i\theta_1 V(i-1,j+1)+i\beta_1 V(i-1,j)+\mu\int_0^\infty V(i,j,y)dF(y)],\ \ i\geq 1.\end{aligned}$$

Furthermore

$$V(i,j,y)=\max\{x+V(i-1,j),V(i,j)\},\ \ i\geq 1.$$

Thus there are critical numbers $y_{i,j}$ such that the optimal policy in state (i,j) is to accept the organ if $x\geq y_{i,j}$, else reject it. Righter [115] studies the structural properties of $y_{i,j}$ and develops methods of computing them.

What if the patient success rates $p_1,p_2,\cdots$ were different? Then in state (i,j) we first number the i active patients as 1,2,...,i so that $p_1\leq p_2\leq\cdots\leq p_i$. We use the same $y_{0,j}=\infty\geq y_{1,j}\geq\cdots\geq y_{i,j}$ as above, and if the organ value is in $(y_{i-k+1,j},y_{i-k,j}]$, it is optimal to assign it to patient k, and if it is less than $y_{i,j}$ it is optimal to reject the organ.

2.3.4 David-Yechiali Model.

In a followup to their 1985 paper [27], David and Yechiali [29] consider a decision model involving N patients that are available at time 0, and $M\geq N$ organs that arrive sequentially. Suppose the patients and the organs are characterized by a single trait (say blood type) that can take K values, say $1,2,\cdots,K$. Let a_i be the trait of the ith patient, and assume that these are known. Let X be the trait of the first incoming organ, chosen randomly from the population. Let $f_i=P(X=a_i)$ be the relative frequency of the the trait of the ith patient in the population. Assume that $f_1\leq f_2\leq\cdots\leq f_N$. Thus trait of patient 1 is the most rare, and the trait of patient N is most common. Now, if $X=a_i$ for some i, then the incoming organ matches patient i, and if $X\neq a_i$ for all i, then the incoming organ is a mismatch. If we assign the organ to patient i and $X=a_i$, then we get a reward of R, and if $X\neq a_i$ we get a reward of $r<R$. Suppose $M>N$. Then we have three options when an organ comes in: assign it to a matching patient (assuming one is available), assign it to a non-matching patient, or reject it (i.e., each patient must be assigned an organ). When $M=N$, the third option is unavailable. The aim is to maximize the expected total reward obtained.

It is clear that if a match is available (i.e $X=a_i$ for some i) then it is optimal to assign the organ to a matching patient, since then we get the most

reward from any organ. If $M > N$, and no matching patient is available, we need to decide whether to assign the organ to a non matching patient, or to reject it. If patient i is assigned a random organ, then the expected reward is $f_i R + (1 - f_i)r = \zeta_i$. This expected reward is the smallest for $i = 1$. Clearly $\zeta_1 \geq r$, hence, it is better to accept a mismatch and assign it to patient 1. Thus the optimal policy is:

Assign the incoming organ to a matching patient if one is available. If no matching patient is available and $M > N$, reject the organ. If no matching patient is available and $M = N$, assign it to the patient with the rarest trait.

David and Yechiali [29] provide a rigorous proof of this policy. They also show that this policy minimizes the total reward stochastically, not only in expectation.

Recommended for presentation: David and Yechiali [28].

2.3.5 Su-Zenios Model.

Su and Zenios [128] consider another sequential stochastic assignment model. We describe a discrete version of it. We have an arbitrarily large number of patients waiting for organs that arrive sequentially. Suppose there are m different patient types and k different organ types. If we assign a organ of type j to patient of type i, we get a reward of r_{ij}. Assume that we are not allowed to discard an available organ. We have a large number of patients, and the relative frequency of type i among the patient population is p_i, and that of type organ j is q_j. We can describe a policy (called the first-best policy) by the numbers a_{ij} which represents the joint probability that an incoming organ is of type j and it gets assigned to patient of type i. Thus an incoming organ of type j is assigned to a patient of type i with probability $a_{ij} / \sum_{i=1}^{m} a_{ij}$. Then the optimal assignment a_{ij} can be obtained as the solution to the following transportation problem:

$$\max \sum_{i=1}^{m} \sum_{j=1}^{k} a_{ij} r_{ij}$$

such that

$$\sum_{j=1}^{k} a_{ij} = p_i, \quad i = 1, \cdots, m,$$

$$\sum_{i=1}^{m} a_{ij} = q_j, \quad j = 1, \cdots, k,$$

$$a_{ij} \geq 0.$$

They next consider an extension where the organ types are a continuous variable with cdf F, while patient types are a finite number m. Then the reward of

assigning an organ of type x to a patient of type i is denoted by $R_i(x)$. We say that $R_i \geq R_j$ if $R_i(x) - R_j(x)$ is an increasing function of x. Suppose $R_m \geq R_{m-1} \cdots \geq R_1$. Then we can define $m+1$ numbers $a_0, a_1 \cdots a_m$ as follows:

$$a_0 = -\infty, \;\; F(a_i) = p_1 + p_2 + \cdots + p_i, \;\; 1 \leq i \leq m-1, \;\; a_m = \infty.$$

Then the asymptotic optimal policy is to assign an incoming organ of type in the interval $A_i = (a_{i-1}, a_i]$ to patient of type i.

Su and Zenios [128] next allow the patients to reject at most one assigned organ. In order to penalize a patient from rejecting too many organs they assume that patients are in a queue within each type and when an organ is assigned to type i, it is offered to the first patient in the queue of type i patients. If that patient rejects it, he is moved down to the position K of that queue, and that organ is wasted (admittedly not a realistic assumption, but under optimal policies, the advantages of allocating the rejected organs to someone else are minuscule.) Thus if $K = 1$, there is no penalty. We assume that patients use a discount factor δ to discount the future values.

Now suppose we follow a policy given by the partitions $A_i, \;\; 1 \leq i \leq m$, i.e., an organ with $x \in A_i$ is offered to the first patient of type i. Suppose $P(X \in A_i) \leq p_i$. If she accepts it, she gets a reward of $E(R(X)|X \in A_i)$. If she rejects it now and accepts the next one offered to her, the discounted value of the next offer is

$$v_K = \left(\frac{\delta P(X \in A_i)}{1 - \delta + \delta P(X \in A_i)} \right)^K E(R(X)|X \in A_i).$$

Now suppose,

$$\inf_{x \in A_i} R_i(x) \geq v_K.$$

Then it is clear that this patient will eventually accept an organ that is offered to her. Thus as the number of patients becomes large, all patients will eventually accept an organ. Hence the optimal policy (called the second best) can be obtained by solving the following optimization problem:

$$\max_{A_1, \cdots A_m} \sum_{i=1}^{m} E(R_i(X) 1_{\{X \in A_i\}})$$

$$\text{such that} \;\; \inf_{x \in A_i} R_i(x) \geq v_K, \;\; 1 \leq i \leq m,$$

$$\text{and} \;\;\;\; P(X \in A_i) \leq p_i, \;\; 1 \leq i \leq m.$$

The discrete version of this program can also be written down.

Recommended for presentation: SU and Zenios [127].

2.4 Live Donors

Most of the organ used in transplants are cadaveric organs, that is they are harvested from dead persons. However, it is possible to have live donor transplants in special cases like kidneys, where a live donor can agree to donate one his two good kidneys to a patient in need of a new kidney. Clearly a kidney from a live donor is preferable to a kidney from a cadaver. When a patient needs a kidney she can join the waiting list for a new kidney, which typically comes from a dead person (cadaveric kidney). Alternatively, she may find someone living who is willing to donate her one of his two kidneys. If this donor kidney is physiologically acceptable to her (tissue and blood-type compatible), she has to decide when to go through the transplant. However, if the donor is not compatible she and this incompatible donor can join a kidney-exchange program, where they wait for another incompatible pair to show up, so that the first donor is compatible with the second recipient and the second donor is compatible with the first recipient. Then the exchange takes place and both recipients end up getting live kidneys, rather than cadaveric kidneys. There is a third possibility: she can join the cadaveric kidney waiting list while the live donor gives his kidney to some anonymous recipient. In return for this act of altruism on part of her living donor, she gets to go to the head of the line of patients waiting for cadaveric kidneys. Here we build a stylized model to help us decide whether the incoming pair should join the kidney exchange queue or the cadaveric queue. Clearly it matters who is making the decisions. We develop two models: in the first the decisions are made by a centralized manager who aims to maximize total benefits to the society, in the second model each incoming pair maximizes its own benefit.

2.4.1 Compatible Live Donor

Here we consider the problem of optimal timing of a transplant assuming the patient has found a compatible live donor. We describe a discrete time Markov decision process model inspired by Alagoz et al [1].

Suppose the states of a live patient can be classified as $1, 2, \cdots, H$. The higher the state, the worse is the patient's health. We also consider a state $H+1$ to represent death, and state T to represent "Transplant done". In each state $i \in \{1, 2, \cdots, H\}$, there are two possible decisions: Wait (W) or Transplant (T). If the decision is to wait in state i, the patient gets a reward of $r(i, W)$, which can be thought of quality adjusted value of one period of waiting, and then the state changes to $j \in \{1, 2, \cdots, H+1\}$ with probability p_{ij} independent of the past. If a transplant decision is made in state i, the patient gets a reward of $r(i, T)$, which can be thought of as the QUAL for the patient after receiving a live kidney, and the patient moves to state T. Once in state $H+1$ or T, the patient stays there forever without any further rewards. Suppose the discrete discount factor is α.

Let $V(i)$ be the optimal reward starting in state i. It satisfies the DP equation is:

$$V(i) = \max\{r(i,T), r(i,W) + \alpha \sum_{j=1}^{H} p_{ij} V(j)\}, \;\; 1 \le i \le H.$$

Clearly, $V(H+1) = V(T) = 0$. Since the state-process will eventually get absorbed in states $H+1$ or T, and there are no rewards in these states, there is an optimal policy for the discounted as well as the undiscounted case. Suppose $r(i,W)$ and $r(i,T)$ are decreasing functions of i (assume $r(H+1,T) = r(H+1,W) = 0$, and let $P = [P_{ij}]_{i,j=1,2,\cdots,H}$. We present two results below:

Theorem 2.1 Suppose P matrix is upper triangular, and

$$(1 - \alpha p_{ii})(r(i,T) - \alpha r(i+1,T)) \ge r(i,W), \;\; 1 \le i \le H.$$

Then $V(i) = r(i,T)$ for $i \in \{1, 2, \cdots, H\}$ and the optimal decision in state i is to transplant.

This makes sense, since the patient health deteriorates continuously and the decrease in the benefit of transplant from waiting one more period is more than the reward of waiting one period.

Next suppose P is stochastically increasing (also called IFR: increasing failure rate), that is

$$z(i) = \sum_{j=k}^{H} p_{ij}, \;\; 1 \le i \le H$$

is an increasing function of i for every k. In other words, the deterioration accelerates in a stochastic sense. One important property of a stochastically increasing matrix is that it preserves monotonicity: if a vector v increases in its components, then so does Pv. Using this property one can show the following:

Theorem 2.2 Suppose P is stochastically increasing, $r(\cdot, W)$ is non-increasing, and $r(\cdot, T)$ satisfies

$$\frac{r(i,T) - r(i+1,T)}{r(i,T)} \le \alpha(p_{i+1,H+1} - p_{i,H+1}), \;\; 1 \le i \le H.$$

Then there exists a state i^* such that it is optimal to transplant if the patient state is $i^* \le i \le H$, and wait if $1 \le i \le i^* - 1$.

The condition on $r(\cdot, T)$ has an intuitive explanation: "as the patient gets sicker, the reduction in the benefit of waiting is greater than the reduction in the benefit of performing the transplant", to quote Alagoz et al [1].

The policy given in the above theorem is called a control limit policy. Algoz et al [1] show numerically that a control limit policy performs well even if the conditions in the theorem are not strictly met, as long as the violations are "small".

2.4.2 Kidney-Exchange Programs

Here we consider the problem faced by a patient who has found a live donor, but the donor is not compatible. In this section we present a model inspired by Zenios [153].

Suppose there are two types of pairs, the donor in the first type can donate to the other, and the donor in the second can donate to the first. Suppose type pairs of type i arrive according to $\text{PP}(\lambda_i)$ $(i = 1, 2)$. Let $X_i(t)$ be the number of type i pairs waiting at time t. Note that we cannot have both queues positive. Hence define

$$Q(t) = X_1(t) - X_2(t).$$

Thus $X_1(t) = Q(t)^+$ and $X_2(t) = Q(t)^-$. In the absence of any control, $\{Q(t), t \geq 0\}$ is a birth and death process on $\{0, \pm 1, \pm 2, \cdots\}$ with birth rate λ_1 in each state and death rate λ_2 in each state. This process is transient if $\lambda_1 \neq \lambda_2$ (unbalanced system) and null recurrent if $\lambda_1 = \lambda_2$ (balanced system).

Now suppose a pair waiting in the exchange gets reward h_0 per unit time (Quality of Life), a patient getting a live kidney experiences reward rate h_2 per unit time forever, and a patient getting a cadaveric kidney experiences a reward at rate h_1 per unit time forever. Suppose $h_0 < h_1 < h_2$ and the rewards are discounted at rate α.

When a new pair arrives, the system manager can decide to let the pair join the kidney-exchange queue, or the cadaveric queue. Let $V(i)$ be the optimal reward to all kidney recipients starting from state i. (We ignore the effect on the donors, since their quality of life decreases marginally.) Scale the time so that $\lambda_1 + \lambda_2 + \alpha = 1$. Then we have

$$V(i) = h_0 i + \lambda_1 \max\{V(i+1), h_1/\alpha + V(i)\} + \lambda_2(2h_2/\alpha + V(i-1)), \quad i \geq 1,$$

$$V(i) = -h_0 i + \lambda_2 \max\{V(i-1), h_1/\alpha + V(i)\} + \lambda_1(2h_2/\alpha + V(i+1)), \quad i \leq -1,$$

$$V(0) = \lambda_1 \max\{V(1), h_1/\alpha + V(0)\} + \lambda_2 \max\{V(-1), h_1/\alpha + V(0)\}.$$

Note that this assumes that the patient gets immediate kidney replacement on the cadaveric queue, which is an approximation. We do not model the cadaveric queue. Using methods of MDP (Markov Decision Processes) we can show that the optimal policy is characterized by two constants, s_1 and s_2, as follows: In state i allow a type 1 pair to join the kidney exchange queue if $i < s_1$, else send it to the cadaveric queue. Similarly, allow a type 2 pair to wait in the exchange queue if $i > -s_2$, else send it to the cadaveric queue. It is possible to derive explicit expressions for the total expected discounted reward as a function of s_1 and s_2, and then one can choose the values of s_1^S and s_2^S that maximize this reward. (The superscript S is for social optimality.)

Next suppose each pair decides to join the exchange queue or get an instant cadaveric transplant based on whichever produces the larger reward. Consider

a pair of type 1 that sees i pairs of type 1 waiting. (If it sees the other type waiting, then clearly it will get a live exchange transplant and leave.) Let $V_1(i)$ be the reward of waiting in the exchange queue. Then the pair will wait if

$$V_1(i) \geq h_1/\alpha.$$

One can explicitly compute $V_1(i)$ as follows:

$$V_1(i) = \frac{h_0}{\alpha} + \left(\frac{\lambda_2}{\lambda_2+\alpha}\right)^{i+1} \frac{h_2 - h_0}{\alpha}.$$

Hence a pair of type 1 will wait in the queue if the number of type 1 pairs waiting in front of it is less than

$$s_1^I = \frac{\ln(\frac{h_1-h_0}{h_2-h_0})}{\frac{\lambda_2}{\lambda_2+\alpha}} - 1.$$

We can similarly compute s_2^I such that type 2 pairs join if $i > -s_2^I$. (The superscript I is for individual optimality, or for incentive compatible.)

For example, when $\lambda_1 = \lambda_2 = 5$, $s_1^S = s_2^S = 28$, while $s_1^I = s_2^I = 487$. In general, if the pairs are allowed to act in their own self interest, most of them will decide to wait in the exchange, and produce very long waiting periods. The social optimal policy works much better, but violates individual rights!

Zenios [153] also studies the diffusion approximation to this problem and comes up with a more tractable model, which works quite close to the socially optimal model.

Recommended for Presentation: Kurt et al [79].

2.5 Multi-class Queueing Models

Next we present a queueing model of the organ transplant system, following Zenios [152]. The patients belong to K classes and the organs belong to J different classes. Class k patients arrive according to PP(λ_k), $(1 \leq k \leq K)$ and form K separate queues. Organs of class j arrive according to PP(μ_j), $(1 \leq j \leq J)$. An incoming organ of class j is assigned to a patient of class k with probability v_{jk}, $(\sum_k v_{jk} = 1)$. If there is no patient of class k waiting and an organ is assigned to it, that organ is lost. Patients of class k renege (die) at rate θ_k.

Let $X_k(t)$ be the number of class k patients waiting in the system at time t. $\{X_k(t), t \geq 0\}$ is a birth and death process with birth rate in state n given by $\alpha_n = \lambda_k$ and death rate in state n given by $\beta_n = \sum_j \mu_j v_{jk} + n\theta_k$. This process is always stable and its limiting distribution

$$\pi_k(n) = \lim_{t\to\infty} P(X_k(t) = n)$$

is given by

$$\pi_k(n) = C(\rho_k \nu_k)^n \prod_{j=0}^{n} \left(\frac{1}{\nu_k + j} \right),$$

where

$$\rho_k = \frac{\lambda_k}{\sum_j \mu_j v_{jk}}, \quad \nu_k = \frac{\mu_k}{\theta_k},$$

and C is a normalizing constant. From this we can compute all other performance measures such as L_k the expected queue length, D_k the expected waiting time, W_k the expected waiting time of the customers who receive a transplant, ϕ_k the fraction of patients who get transplants, etc.

Simpler approximations can be obtained by taking fluid limits (i.e., studying a system with arrival rate $n\lambda_k$ and $n\mu_k$ and letting $n \to \infty$.) We give the final expressions below (assuming $\lambda_k > \sum_j \mu_j v_{jk}$):

$$D_k = \frac{1}{\theta_k} \left(1 - \frac{\sum_j \mu_j v_{jk}}{\lambda_k} \right),$$

$$W_k = \frac{1}{\theta_k} \ln \left(\frac{\lambda_k}{\sum_j \mu_j v_{jk}} \right),$$

$$\phi_k = \frac{\sum_j \mu_j v_{jk}}{\lambda_k}.$$

Using these approximations Zenios [152] identifies the allocation policy (i.e., v_{jk}'s) that equalizes the W_k's in the asymptotic regime.

Recommended for presentation: Boxma et al [13].

Chapter 3

Stochastic Models of Disease Progression

3.1 Multi-state Models of Disease Progression

Disease progression models are motivated by the observed phenomena that we can identify different stages of a disease in a given patient based on symptoms and tests, for example, colo-rectal cancer, leukemia, renal diseases, AIDS, diabetes, etc. The particulars of the categorization of the stages of a disease state depend on the disease. We denote the state of the disease in a given patient at time t by $X(t)$ and assume that $\{X(t), t \geq 0\}$ is a multi-state stochastic process. Three most commonly used stochastic processes are

1. Continuous Time Markov Chains (CTMC). Here we assume that $\{X(t), t \geq 0\}$ is a Continuous Time Markov Chain (CTMC) on state space $\Omega = \{0, 1, 2, \cdots, S\}$, with rate matrix Q. It is common to denote death by state S, thus making it an absorbing state. Most commonly used CTMCs are modified pure birth processes with Q matrix structured as below ($S = 4$)

$$Q = \begin{bmatrix} -\lambda_0 - \theta & \lambda_0 & 0 & 0 & \theta \\ 0 & -\lambda_1 - \theta & \lambda_1 & 0 & \theta \\ 0 & 0 & -\lambda_2 - \theta & \lambda_2 & \theta \\ 0 & 0 & 0 & -\lambda_3 - \theta & \lambda_3 + \theta \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.1}$$

or modified birth and death processes Q matrix structured as below ($S =$

4)

$$Q = \begin{bmatrix} -\lambda_0 - \theta & \lambda_0 & 0 & 0 & \theta \\ \mu_1 & -\lambda_1 - \mu_1 - \theta & \lambda_1 & 0 & \theta \\ 0 & \mu_2 & -\lambda_2 - \mu_2 - \theta & \lambda_2 & \theta \\ 0 & 0 & \mu_3 & -\lambda_3 - \mu_3 - \theta & \lambda_3 + \theta \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.2}$$

2. Semi-Markov Processes (SMP). Here we assume that $\{X(t), t \geq 0\}$ is an SMP on state space $\Omega = \{0, 1, 2, \cdots, S\}$. Such a process has piece-wise constant sample paths, with jumps at an increasing sequence of times $\{T_n, n \geq 0\}$. The probabilistic evolution is described as follows

$$P(X(T_{n+1}) = j, T_{n+1} - T_n \leq x | X(T_n) = i, X(t), 0 \leq t < T_n) = G_{ij}(x).$$

The matrix $G(x) = [G_{ij}(x)]$ is called the kernel of the SMP. SMPs allow us to move away from the assumption of exponential sojourn times in the states of the CTMCs and hence afford a better fit for the data. Typically we use SMPs with same structure as the CTMCs described above.

3. Fluid Process. When the state of the disease can be identified as a continuous level of a biomarker (for example, PSA level, or Hematocrit level), one can model $\{X(t), t \geq 0\}$ as a fluid process described by differential equations, whose parameters may vary randomly.

The main practical impediment in using such models is that the patient state is observed periodically and not continuously. Thus we do not know the entire sample path of the process $\{X(t), t \geq 0\}$, but have a finite number of observations $X(t_i)$ at time $0 \leq t_1 < t_2 < \cdots < t_n$. The first question is how to estimate the parameters of the process based on these observations. There are many practical problem that arise: missing data, inaccurate data, censored data, among others. The second question is how to determine the observation points t_i if they are under our control. The third question is how to evaluate the usefulness of the model. This depends upon how we can test the predictions made by the model. Finally, the fourth question is how to use the model to improve the treatments for the disease. We shall look at some of these issues here with specific examples.

3.1.1 Disease Progression Models for Cancer

We follow the development in the book by Frank [40]. The idea of modeling the development of cancer in human body as progressing through stages was initiated in a seminal paper by Armitage and Doll [5]. First some terminology:

Cancer *incidence rate* is the number of new cases of cancer identified in a year in a given age group divided by the number of individuals in that age group. Thus typically the incidence rate will be given by number of cases per

100000 population in that age group. It has been observed the logarithm of the incidence rate grows approximately linearly with the logarithm of the age group over the age range from 25 to 65 or so. Thus if the incidence rate is I and the age group is t, then we have approximately

$$\log(I) = c + b\log(t).$$

One way to explain this behavior is to posit that the cancer proceeds through stages. For example, a simple model for colo-rectal cancer is assumed to go through the following six stages:

1. Normal Epithelium
2. Displastic Crypt and Early Adenoma
3. Intermediate Adenoma
4. Late Adenoma
5. Carcinoma
6. Metastasis

There is a cell-level reason why the disease progresses from one stage to the next. We will not go into those details here. One can model the cancer progression as a pure birth process with state space $\{0, 1, \cdots, n\}$. An individual starts in state 0, and is identified as a cancer case when he enters state n. Let λ_i be the birth rate in state i, $0 \leq i \leq n-1$. State n is absorbing. Let $X(t)$ be the stage that the cancer is in at time t given that $X(0) = 0$. Let

$$p_i(t) = P(X(t) = i | X(0) = 0).$$

Then we know that $p_i(t)$ satisfy the following forward Chapman-Kolmogorov equations:

$$p_0\prime(t) = -\lambda_0 p_0(t),$$
$$p_i\prime(t) = \lambda_{i-1}p_{i-1}(t) - \lambda_i p_i(t), \quad 1 \leq i \leq n-1,$$
$$p_n\prime(t) = \lambda_{n-1}p_{n-1}(t).$$

The initial conditions are

$$p_0(0) = 1, \quad p_i(0) = 0, \quad 1 \leq i \leq n.$$

When $\lambda_i = \lambda$ for $i = 0, 1, \cdots, n-1$, we get the following solution:

$$p_i(t) = e^{-\lambda t}(\lambda t)^i / i!, \quad 0 \leq i \leq n-1.$$

Since state n is absorbing, we get

$$p_n(t) = 1 - \sum_{i=0}^{n-1} p_i(t).$$

This also gives the total fraction of the population that develops this cancer by age t. The incidence rate is given by

$$I(t) = \frac{p_n\prime(t)}{1 - p_n(t)} = \frac{\lambda(\lambda t)^{n-1}/(n-1)!}{\sum_{i=0}^{n-1}(\lambda t)^i/i!}.$$

Taking logs, we get

$$\log(I(t)) = n\log(\lambda) + (n-1)\log(t) - \log(1 - p_n(t)).$$

If we focus on the region where cancer is rare (say age range below 60), we can ignore the last term in the above equation, and hence we get the observed log-log relationship.

The slope of the log-log plot gives the percentage change in the incidence per percentage change in age. It is called the log-log acceleration and denoted by ILA(t). That is,

$$ILA(t) = \frac{d\log(I(t))}{d\log(t)} = \frac{d(I(t))/I(t)}{dt/t} = n - 1 - \lambda t(S_{n-2}(t))/S_{n-1}(t)),$$

where

$$S_i(t) = \sum_{k=0}^{i}(\lambda t)^k/k!.$$

If the log-log model is strictly valid, the ILA is constant over time. However, it has been observed that ILA does vary with age, especially after 60. More detailed analysis of the deviation from linear log-log relationship has been done by Frank [41]. There are many different variations one can consider. First, one can posit that there are multiple regions in the tissue, and each region develops along the pure birth model described above. Cancer is identified when one of these regions reaches the final stage. Second, there might be multiple pathways of cancer progression, each with different number of stages and different progression rates, and cancer is detected when any of these pathways reaches its final stage. Third, the birth rates can be time varying, giving different accelerations in different age regions. Four, one can consider a mortality rate from each state from all causes other than cancer, thus producing a modified pure birth process. One can consider any combination of these approaches.

3.1.2 Chronic Myelogenous Leukemia

We follow the material in Klein et al [75]. Chronic myelogenous leukemia (CML) is a type of blood cancer where the bone marrow produces too many white blood cells. They model CML by a three state Semi-Markov Process (1 = stable phase, 2 = accelerated phase, 3 = death). The disease moves from stage 1 to 2 or 3, and from 2 to 3. Let F_i be the cdf of the sojourn time in state i, and let r_i be the hazard rate of F_i. We have an additional information about the patients that

can influence this distribution, namely, the Adenosine deaminase (ADA) level (measured in units of 10^{-8} moles inosine/hour/10^6 blood cells) in the patient's blood. This is called a biological marker. Let $X(t)$ be the state of the disease and $Z(t)$ be the level of ADA in the patient's blood at time t. We assume that

$$r_i(t) = \lambda_i exp(\beta_i Z(t)),$$

where t is the time spent in state i. Then

$$F_i(t) = 1 - exp(-\int_0^t r_i(u))du.$$

The aim is to estimate the λ's and the β''s based on the observed data. Clearly we do not observe the patients continuously in time. We observe a patient at time $0 = t_0 < t_1 < t_2 \cdots < t_n$. Let $X(t_i)$ be the state of the disease and $Z(t_i)$ be the level of ADA in the patient's blood at time t_i. We assume that

$$Z(t) = Z(t_i), \;\; t_i \leq t < t_{i+1},$$

that is the Z process is assumed to be constant between observations. We need the following probabilities to compute the likelihood of the data:

$$P_{11}(t_{n+1}-t_n|z) = P(X(t_{n+1} = 1|X(t_n) = 1, Z(t_n) = z) = exp(-r_1(t_n)(t_{n+1}-t_n)),$$

$$P_{12}(t_{n+1} - t_n|z) = P(X(t_{n+1} = 2|X(t_n) = 1, Z(t_n) = z) =$$

$$[exp(-r_2(t_n)(t_{n+1} - t_n)) - exp(-r_1(t_n)(t_{n+1} - t_n))](r_1(t_n)/(r_1(t_n) - r_2(t_n)),$$

$$P_{13}(t_{n+1} - t_n|z) = P(X(t_{n+1} = 3|X(t_n) = 1, Z(t_n) = z) =$$

$$1-P(X(t_{n+1} = 1|X(t_n) = 1, Z(t_n) = z)-P(X(t_{n+1} = 2|X(t_n) = 1, Z(t_n) = z),$$

$$P_{22}(t_{n+1}-t_n|z) = P(X(t_{n+1} = 2|X(t_n) = 2, Z(t_n) = z) = exp(-r_2(t_n)(t_{n+1}-t_n)),$$

$$P_{23}(t_{n+1}-t_n|z) = P(X(t_{n+1} = 3|X(t_n) = 2, Z(t_n) = z) = 1-exp(-r_2(t_n)(t_{n+1}-t_n)),$$

$$P(X(t_{n+1} = 3|X(t_n) = 3, Z(t_n) = z) = 1.$$

Each patient is seen for different number of times and at different times. For example, Klein et al [72] have the following data for one patient

n	t_n	$X(t_n)$	$Z(t_n)$
0	0	1	6.3
1	51	2	9.9
2	63	2	11.7
3	107	2	21.8
4	124	3	0.0

The contribution of this patient data to the likelihood is

$$P_{12}(51|6.3)P_{12}(12|9.9)P_{22}(44|11.7)P_{23}(17|21.8).$$

Klein et al [72] use the data from 55 patients and numerically compute the maximum likelihood estimates given below

$$(\lambda_1, \lambda_2, \beta_1, \beta_2) = (6.237 \cdot 10^{-5}, 5.293 \cdot 10^{-3}, 2.257 \cdot 10^{-1}, 8.906 \cdot 10^{-3}).$$

3.1.3 Hepatocellular Carcinoma

Kay [72] reports a modified birth and death process model of the Hepatocellular Carcinoma, a type of liver cancer. The concentration of alpha-fetoprotein (AFP) in the blood can be taken as a biomarker for the cancer, the higher the AFP level, the higher the severity of the cancer. Typically the patient starts with a base line observation of the AFP level. As long it stays less 5% above that, the patient is said to be in state 1, if it is above 5% of that he is said to be in state 2. State 3 is death. Kay assumes that the $\{X(t), t \geq 0\}$ is a CTMC on state space $\{1, 2, 3\}$ with rate matrix

$$Q = \begin{bmatrix} -\lambda_1 - \mu_1 & \lambda_1 & \mu_1 \\ \lambda_2 & -\lambda_2 - \mu_2 & \mu_2 \\ 0 & 0 & 0 \end{bmatrix}. \tag{3.3}$$

For this CTMC it is easy to compute

$$p_{ij}(t) = P(X(t) = j | X(0) = i).$$

Let

$$\rho_1 = \frac{1}{2}\{-(\mu_1 + \mu_2 + \lambda_1 + \lambda_2) + \sqrt{(\mu_1 - \mu_2 + \lambda_1 - \lambda_2)^2 + 4\lambda_1\lambda_2}\},$$

$$\rho_2 = \frac{1}{2}\{-(\mu_1 + \mu_2 + \lambda_1 + \lambda_2) - \sqrt{(\mu_1 - \mu_2 + \lambda_1 - \lambda_2)^2 + 4\lambda_1\lambda_2}\}.$$

Then, for $i = 1, 2$, we have

$$p_{ii}(t) = \frac{1}{\rho_1 - \rho_2}[(\mu_i + \lambda_i + \rho_1)e^{\rho_2 t} - (\mu_i + \lambda_i + \rho_2)e^{\rho_1 t}],$$

$$p_{i,j}(t) = \frac{\lambda_i}{\rho_1 - \rho_2}(e^{\rho_1 t} - e^{\rho_2 t}), \;\; j \neq i,$$

$$p_{i3}(t) = 1 + \frac{\mu_i + \rho_2}{\rho_1 - \rho_2}e^{\rho_1 t} + \frac{\mu_i + \rho_1}{\rho_1 - \rho_2}e^{\rho_2 t}.$$

Typical data for a patient looks like (here $Z(t_n)$ is the % increase in AFP value observed at time t_n, with AFP value of "na" implying death)

n	t_n	$Z(t_n)$
0	0	4.3
1	31	9.9
2	75	11.7
3	152	21.8
4	194	na

Now fix an α and define $X(t) = 1$ if $0 < Z(t) < \alpha$ and $X(t) = 2$ if $Z(t) \geq \alpha$, and $X(t) = 3$ if $Z(t) = na$. Thus the classification of the disease stage will depend on α. Kay [72] uses data from 85 patients, and reports the estimates of λ_i and μ_i $(i = 1, 2)$ for two values of α: 5% and 10%.

parameter	$\alpha = 5$	$\alpha = 10$
λ_1	.0081	.0049
λ_2	.0014	.0008
μ_1	.0010	.0018
μ_2	.0083	.0091

It shows that using higher value of α makes stage 2 mortality rate higher, as expected. Another interesting question is: is it useful to measure AFP levels at all? That is, does it provide any more information about the seriousness of the disease? This can be done by testing the hypothesis $H_0 = \mu_1 = \mu_2$. The data rejects this hypothesis conclusively. That is the mortality rates are indeed significantly influenced by the AFP levels, hence it is useful to measure them.

3.1.4 Diabetic Retinopathy

Diabetic Retinopathy is a deterioration of the retina due to diabetes. The working group of ophthalmologists classify the disease severity into six grades. Marshall and Jones [91] model the progression of the disease as a 4 state birth and death Markov chain, with state for grade I (no retinopathy), state 2 for grades II-III (early retinopathy involving micro-aneurisms), state 3 for grades IV-V (pre-proliferative retinopathy), and state 4 for grade VI (proliferative retinopathy). The rate matrix of the CTMC is

$$Q = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & 0 \\ \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 & 0 \\ 0 & \mu_3 & -\lambda_3 - \mu_3 & \lambda_3 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.4}$$

One can further include the effect of covariates z(other information about the patients, such as age, duration of diabetes, blood pressure, gender, smoker or not, cholesterol, etc) as a proportional hazard rate model:

$$\lambda_i(z) = \lambda_i \exp(\beta_i \prime z)$$

$$\mu_i(z) = \mu_i \exp(\theta_i \prime z).$$

Note that the covariates are fixed at the beginning of the experiment, and thus, the disease state of a patient with covariate vector z will be a CTMC with rate matrix $Q(z)$ obtained by replacing the rates in Q by the covariate-dependent rates. For a one-dimensional covariate this implies a total of ten (five+five) parameters to be estimated from the data. This can be reduced to two+five if we assume

$$\beta_i = \beta, \quad \theta_i = \theta.$$

This can be reduced further to one+five if we assume

$$\beta = -\theta.$$

The data consists of 277 patients with 882 patient visits. The time and the retinopathy grade is recorded at each visit. Marshal and Jones [91] report the

maximum likelihood estimates for these three models and identify the significant covariates.

3.1.5 AIDS

Acquired immunodeficiency syndrome (AIDS) is a disease of the human immune system caused by the human immunodeficiency virus (HIV). It manifested in 1981 and so far has caused over 30 million deaths. Tremendous research has been done related to AIDS. Here we present the results of one such study published by Longini et al [83]. They model AIDS as a pure birth CTMC with five states: (1) infected but antibody-negative; (2) antibody-positive but asymptomatic; (3) pre-AIDS symptoms and/or abnormal haematologic indicator; (4) clinical AIDS and (5) death. The model is described by four transition rates $q_{i,i+1} = \lambda_i$ for $i = 1, 2, 3, 4$. The time to reach state 4 is called the incubation period. Using the standard CTMC theory we can compute the transition probabilities $p_{ij}(t)$ for the above pure birth process. When all the λ's are distinct, we get

$$p_{ik}(t) = (-1)^{k-i}\lambda_i \cdots \lambda_{k-1} \sum_{j=i}^{k} e^{-\lambda_j t} / \prod_{l=i, l\neq j}^{k} (\lambda_j - \lambda_l), \;\; 1 \leq i \leq k \leq 4,$$

$$p_{i5}(t) = (-1)^{k-i}\lambda_i \cdots \lambda_4 \sum_{j=i}^{4} (1 - e^{-\lambda_j t})/\lambda_j \prod_{l=i, l\neq j}^{k} (\lambda_j - \lambda_l), \;\; 1 \leq i \leq 4.$$

Let T be the incubation period of AIDS. Its density is given by $\lambda_3 p_{13}(t)$. (This is the rate at which the CTMC enters state 4.) The mean and variance of T is given by

$$E(T) = \sum_{i=1}^{3} \frac{1}{\lambda_i},$$

$$\text{Var}(T) = \sum_{i=1}^{3} \frac{1}{\lambda_i^2}.$$

Based on their data the authors obtain the following estimates (in 1/months):

$$\lambda_1 = .4571, \;\; \lambda_2 = .0190, \;\; \lambda_3 = .00159, \;\; \lambda_4 = .0420.$$

Let τ_i be the mean time (in years) until death starting from state i

$$\tau_1 = 11.8, \;\; \tau_2 = 11.6, \;\; \tau_3 = 7.2, \;\; \tau_4 = 2.0.$$

3.1.6 Breast Cancer

Shwartz [124] gives a detailed model of breast cancer. It tracks the state of the breast cancer by two observable quantities: the size of the tumor $S(t)$ and the number of lymph nodes involved $N(t)$. It is believed that tumor grows

exponentially in size, with the patient-dependent growth rate. Hence Shwartz assumes that

$$S(t) = S(0)e^{\Lambda t}$$

where the rate Λ is a random variable with a distribution fitted from observed data. Interestingly, he assumes that $\{N(t), t \geq 0\}$ is a Non-homogeneous Poisson Process (NPP) with rate function

$$n(t) = b_0 + b_1 S(t) + b_2 S'(t),$$

where the parameters $b_0,\ b_1,\ b_2$ are fitted from the data. This gives a complete description of the evolution of the state $(S(t), N(t))$ of the disease. In addition one can add the "death" state with that can be reached from any of these states. Using this model Shwartz obtains the optimal screening schedules for a patient. We shall explain it in the next section.

Recommended for presentation: More recently, Perez et al [105] study a three state pure birth model of breast cancer, with the progression from no relapse, to relapse, to death. They use this to study the effectiveness of treatments, as well influence of patient categories on survival functions. Later in Perez et al [106] they extend the analysis to non-homogeneous pure birth model.

3.2 Optimal Screenings and Treatments

A disease progression model is a useful tool to obtain optimal state dependent treatment regimens for the diseases. We need frequent screening of the patient to determine what state the patient is in. Typically different treatments may be appropriate in different states. The use of a treatment in a state will affect future disease progression for the patient, and generate different costs and benefits. Thus Markov Decision Processes offer a valuable method of obtaining optimal treatments for patients. A useful reference for this material is the paper by Schaefer et al [118].

We first describe the Markov decision process model briefly below. Let X_n be the state of the patient at time n. Let S be the state space, i.e., $X_n \in S$ for all $n \geq 0$. Let A_n be the action chosen at time n after observing the state X_n. Let A be the action space, i.e., $A_n \in A$ for all $n \geq 0$. The evolution of the system is described as follows:

$$P(X_{n+1} = j | X_n = i, A_n = a, X_{n-1}, A_{n-1}, \cdots, X_0, A_0) = p_{ij}(a), \quad i, j \in S,\ a \in A.$$

Let R_n be the reward earned at time n and let

$$r_n(i, a) = E(R_n | X_n = i, A_n = a).$$

A policy tells us what action to choose in what state, i.e., it is a sequence of functions $f_n : S \to A$. A policy $f = [f_0, f_1, \cdots]$ chooses action $f_n(i)$ when

$X_n = i$. The aim is to find a policy that maximizes the total reward earned up to time N, called the horizon. A policy that does this is called an optimal policy. Let $v_n(i)$ be the maximum reward incurred over time $n, n+1, \cdots, N$ if $X_n = i$. We have

$$v_n(i) = \max_{i \in A}\{r_n(i,a) + \sum_{j \in S} p_{ij}(a) v_{n+1}(j)\}, \;\; 0 \leq n < N, \; i \in S,$$

$$V_N(i) = g(i),$$

where $g(i)$ is the terminal reward of ending up in state i in the end. The above equation can be solved recursively in a backward fashion to obtain v_n's for each i. Let $f_n(i) \in A$ be such that

$$r_n(i, f_n(i)) + \sum_{j \in S} p_{ij}(f_n(i)) v_{n+1}(j) = \max_{i \in A}\{r_n(i,a) + \sum_{j \in S} p_{ij}(a) v_{n+1}(j)\}.$$

Then $f_n(i)$ gives the optimal action in state i at time n, and $f = [f_0, f_1, \cdots, f_{N-1}]$ is the optimal policy.

We explain by a few specific examples below.

3.2.1 Hereditary Spherocytosis

Spherocytosis is a genetic disorder in which the red blood cells in a patient are spherical in shape rather than the normal concave disc shape. If the condition is mild, one can simply live with it, but in more sever case (anemic spherocytosis) there are several possible treatment options. Common treatment for such a condition is prophylactic (preventive) splenectomy (removal of the spleen) or prophylactic cholecystectomy (removal of the gall bladder). The decision to implement one or both of these procedures depends on the state of the patient, and the relative benefits of these procedures vs. not doing anything. Magni et al [88] formulate this as a discrete time Markov decision process and derive optimal policies. We briefly explain their model and the results below.

The state of the patient is given by the state of the patients spleen and gallbladder. The gallbladder can be in one of the following five states:

1. No gallstones present,
2. Presence of symptomatic gallstones,
3. Gallstones present with mild biliary colic (pain caused by blockage of bile duct) symptoms,
4. Gallstones present with severe biliary colic symptoms,
5. Gallbladder removed.

The spleen can be in one the following six states:

1. Not removed
2. Removed one year ago
3. Removed two years ago
4. Removed 3 years ago
5. Removed 4 years ago
6. Removed 5 or more years ago.

In addition they also include the state "death". Thus the state space has 31 states. The following four action are available

1. Do nothing
2. Perform preventive spleen removal
3. Perform preventive gallbladder removal
4. Perform preventive spleen and gallbladder removal.

Thus the action space consists of these four actions. The transition probabilities $p_{ij}(a)$ and the rewards (in Quality Adjusted Life Years) are computed from the data given in Marchetti et al [87]. Using those data the authors produce the dynamic optimal policy, slightly restated here: "The best treatment for a patient without gallstones is to wait. Every year a clinical check-up is scheduled, but if gallstones do not develop, no surgery is suggested until the age of 15, when splenectomy is suggested. On the other hand, if gallstones appear (before 15 years of age), both cholecystectomy and splenectomy are suggested. Most (88%) of the patients, however, do not develop gallstones until they are 15, and, therefore, splenectomy is performed at that age. Furthermore, they should undergo cholecystectomy as soon as asymptomatic gallstones are discovered by echography if they are under 55. Otherwise, cholecystectomy should be delayed until the first bout of colic."

3.2.2 Ischemic Heart Disease

Ischemia is the restriction in blood supply to tissues that results in reduction of the supply of oxygen that is needed to keep the tissue alive. Ischemic heart disease is the result of reduction of blood supply to the heart muscle, usually caused by atherosclerosis (thickening of the walls) of the coronary arteries. Typically the exact state of the heart disease is not known. The doctor can order an investigative action which provides more information about the patient's heart, or a treatment action that changes the health state of the patient. We briefly describe the model formulated by Hauskrecht and Fraser [60] to analyze this situation.

The state of the patient is described by a 10-vector as follows

1. Severity of atherosclerosis (none, mild, severe)
2. Severity of ischemia (none, mild, sever)
3. Acute Myocardial Infarction (i.e., heart attack)) (yes, no)
4. Ventricular Function (Normal, Decreased)
5. Bypass (not done, done)
6. Angioplasty (not done, done)
7. Chest Pain (none, mild, severe)
8. Resting EKG ischemia (Yes, no)
9. Catheter coronary artery result (Not Available (NA), normal, mild, severe)
10. Stress test result (NA, normal, mild, severe)

Thus there are a total of 31104 states of the patient's health! In addition there the "death" state for the patient!

As mentioned before there are two investigative options: 1. stress test or 2. angiogram, and four treatment options: 1. no action, 2. medication, 3. angioplasty, or 4. bypass.

Let X_n be the state of the patient at time n, and A_n be the action at time n. We assume that the transition probabilities

$$p_{x,x'}(a) = P(X_{n+1} = x' | X_n = x, A_n = a)$$

are known. We assume that the reward in state $x = [x_1, x_2, \cdots, x_{10}]$ under action a can be written as

$$r(x, a) = \sum_{i=1}^{10} r(x_i) + r'(a).$$

This simplifies the cost estimation problem considerably. The data about the transition probabilities and the rewards are obtained from Wong et al [143]. The authors consider the case of infinite horizon ($N \to \infty$) and employ an ingenious discounting criterion with action dependent discount factor $\gamma(a)$ (=1, if the action is investigative, or treatment 3 or 4, which take relatively short periods of time, and =.95 for long term treatments like 1 and 2 that are supposed to last three months.) The objective is to maximize the total expected discounted reward over infinite horizon.

The authors model this as a partially observable MDP, with the added structural property that only component 1 and 2 are truly unobservable, while others

are observable. (We think this is not really needed, one can solve it approximately as a standard MDP with 8-dimensional vector state.) Solving the model is fairly computationally intensive and we refer to the paper for more details. For example, for a patient with the state vector [?, ?, No, Normal, No, No, Mild, No, NA, NA] the algorithm suggest the optimal action to be the investigative action of stress-test.

3.2.3 Breast Cancer Screening Policies: Shwartz Model

The material in this section is based on Shwartz [124]. We consider a policy that recommend the initial screening at age A, and then subsequent screenings after every T years. We want to evaluate the influence of choosing different A and T.

Suppose $q(s)$ is the probability that a tumor of size s is missed by a screening (false-negative). (Shwartz considers the possibility that successive screening outcomes may not be independent.) Let $c(t)h$ be probability that the disease manifests clinically between time t to $t+h$. Shwartz assumes that

$$c(t) = c_0 + c_1 S(t) + c_2 S'(t),$$

where $S(t)$ is the size of the tumor at time t and $c_0,\ c_1,\ c_2$ are fitted from data. Thus we may catch the hidden cancer by screening, or it may surface symptomatically between two screenings at rate $c(t)$. Once the disease is identified it can be teated by one of the procedures. To account for the benefits of diagnostic screening, we need further models. Let $d(t)$ be the death rates from non-breast-cancer causes, and $d_b(t)$ be death rate from untreated cancer at time t since the initial screening. Let $d_r(t)$ be the death rate from breast cancer t years after it has been treated (due to relapse or complications, say).

With this data, it is possible to compute $L(A,T)$ = expected life time if policy (A,T) is followed. One can then study the effect of A and T. One can also compare it to $L(A)$, the expected lifetime of a woman of age A if no screening is done at all, and we treat the breast cancer only after it manifests symptomatically. The difference would give a quantitative feel for the benefits of the screening policy. For example, Shwartz reports: $L(40) = 36.88$ years, that is if no screening was done the avearge remaining lifetime of a woman of age 40 is 36.88 years (this was in 1978). $L(40,10) = 36.91,\ L(40,5) = 36.97,\ L(40,1) = 37.06$. Thus yearly screening would add roughly 2 months to the average life expectancy.

Recommended for presentation: Maillart et al [89] describe a nice model of breast cancer screening using partially observable MDPs.

3.2.4 HIV Therapy for AIDS

In this section we shall study a stochastic model of AIDS progression and use it to determine the optimal time when the HIV therapy should be started. The material is based on Schecter et al [122].

HIV attacks the white blood cell (specifically the CD4 variety) and diminishes their count. Thus one can use the CD4 blood count as a surrogate for the severity of AIDS. Over the years retroviral therapy has been developed to counteract the presence of HIV. Unfortunately HIV can develop resistance to these drugs fairly quickly. To address this issue it is now standard to administer HAART (Highly Active AntiRetroviral Theray), which uses a mixture of several drugs to slow the resistance and reduce mortality. However the therapy also has down sides: increased toxic side effects, increased evolutionary pressure towards resistance, etc. This creates the question: when is the optimal time to start the therapy. Schecter et al [122] use MDP models to address this question.

In the simple model, we assume the patient visits the doctor every month and his CD4 count is taken. This is then classified as belonging to N possible categories, labeled $\{1, 2, \cdots, N\}$, the higher states indicating higher ranges of CD4 count. Schecter et al use the following four regions of CD4 level: [0,50), [50,200),[200,350),[350,∞) to define the four states. They also include a state 0 to indicate death. Based on the state of the patient, the doctor (in agreement with the patient) has to decide whether to initiate the therapy or to wait one more month. If the patient initiates the therapy, the patient earns a reward (expected quality adjusted lifetime from then on), and the problem terminates. Else the problem is revisited next time, with a possibly deteriorated condition. Let p_{ij} be the probability that the state changes from i to j in one month if the treatment is not started. ($p_{00} = 1$) The aim is to maximize the total QUALIS of the patient. Suppose the the quality of life coefficient in state i is $r(i)$. Then we assume that $r(0) = 0$ and $r(i) \leq r(i+1)$, $0 \leq i \leq N-1$. Let $R(i)$ be the lump sum reward earned if we decide to start the therapy in state i ($R(0) = 0$). Let $v(i)$ be the maximum total reward earned starting from state i. The value function v satisfies the following Bellman Equation:

$$v(0) = 0,$$

$$v(i) = \max\{r(i) + \sum_{j=1}^{N} p_{ij} v(j), R(i)\}, \quad 1 \leq i \leq N.$$

We can obtain useful structural results if we assume that P matrix is stochastically increasing, (or increasing failure rate). That is,

$$\sum_{j=s}^{N} p_{ij} \leq \sum_{j=s}^{N} p_{i+1,j}, \quad 0 \leq i \leq N, 0 \leq s \leq N.$$

Intuitively, this implies that the rate of deterioration of the disease increases with time. Also assume that $r(i)$ and $R(i)$ are non-negative and nondecreasing

in i. Under these conditions, one can show that there is an i^* such that it is optimal to start the treatment as soon as the state is i^* or less (i.e., worse).

Schecter et al report the optimal policies using actual data from 25000 patients. We refer the reader to the paper for further details.

Chapter 4

Stochastic Models of Epidemics

4.1 Background

An epidemic occurs when a disease enters a susceptible population from outside, spreads through the population infecting a significant fraction of it, and eventually dies out as the population becomes immune to the disease. This is in contrast to endemic diseases which always exist in the population. The most common epidemics in humans are the annual flu epidemics, recent SARS epidemic, West Nile virus, AIDS, Malaria, Plague, etc. Epidemics also affect animal and plant populations. The most common topis in epidemiology are: source of the epidemic, isolation and cure, controlling the spread, vaccination, quarantine policies, cost control, etc. Many of these topics are helped by building mathematical models of how epidemics spread. There are many books on mathematical epidemiology. The one by Bailey [8] is a standard textbook. The review papers by Isham [64], Greenwood and Gordillo [53] and the book by Daley and Gani [26] are good sources for the stochastic models in epidemiology. We shall concentrate on a few important models here, again emphasizing the stochastic models over deterministic models.

A mathematical model of epidemics generally has three main components:

1. Population Model: whether its size is fixed or not, if it is homogeneous or stratified, whether the mixing is uniform or not, etc.

2. Disease model: how does the epidemic disease progress? A typical disease progression model of an individual is as follows: An individual is initially susceptible (S) to a disease. At time t_1 he gets exposed (E), but there are no outwardly symptoms. At time $t_2 > t_1$ he becomes infectious (I), that he can now transmit the disease to a susceptible individual by some means of contact. At time $t_3 > t_2$ the individual exhibits symptoms of the

disease. Finally at time $t_4 > t_3$, the individual recovers (R) and becomes immune to the disease, or recovers but again becomes susceptible to the disease, or dies.
In contrast to the above, there are carrier-borne epidemics, where the infection is spread by carriers of the disease, which are the individuals that are infectious, but show no symptoms, and hence stay in the population until detection or death. The susceptible persons that they infect may become infected (with symptoms) or carriers themselves (with no symptoms).
Another possibility is vector borne epidemics, where the disease spreads not by direct contact between humans, but via a third party, such as a mosquito or flea.

3. Mathematical assumptions: deterministic or stochastic, discrete time or continuous time, Markovian or not, etc. A common mathematical assumption is the *Law of Mass Action*, which states that the rate of interaction between two homogenous groups is proportional to the product of the size of the two groups.

It is common to describe an epidemic model by its disease progression model. An SI model assumes that a susceptible individual becomes infections and stays that way forever. This is called the *simple epidemic model*. Similarly we have the SIS model, the SIR model, the SIRS model, the SEIRS model, etc.

4.2 The I model

The I models tracks the infectious individuals and ignores the rest. This is the simplest model possible.

4.2.1 Deterministic Model

Let $i(t)$ be the number of infectious individuals at time t. Suppose each infectious individual generates new infections at rate β and dies at rate γ. Thus we have

$$i'(t) = \beta i(i) - \gamma i(t).$$

The solution is

$$i(t) = i(0)e^{(\beta-\gamma)t}.$$

Thus the epidemic explodes ($i(t) \to \infty$) if $\beta > \gamma$, and dies out ($i(t) \to 0$) if $\beta < \gamma$, and becomes endemic ($i(t) = i(0)$) if $\beta = \gamma$.

4.2.2 Stochastic Model: Discrete Time

Such a model is also called the branching process model. Let I_n be the number of infectious individuals at time $n \geq 0$. Suppose each infectious person infects an

iid number of new individuals (with mean μ and variance σ^2) and then becomes immune (or dies). Let $i_n = E(I_n)$, and $\sigma_n^2 = \text{Var}(I_n)$. Then one can show that

$$i_{n+1} = \mu i_n,$$

$$\sigma_{n+1}^2 = \sigma^2 i_n + \mu^2 \sigma_n^2.$$

Solving these we get

$$i_n = i_0 \mu^n,$$

$$\sigma_n^2 = i_0 \sigma^2 \mu^{n-1} \frac{\mu^n - 1}{\mu - 1}.$$

Thus if $\mu < 1$, the mean and variance of the number of infected eventually converges to zero; if $\mu = 1$, the mean stays unchanged but variance explodes to infinity, and if $\mu > 1$ the mean and variance both explode to infinity. The expected total number of individuals infected during the life time of the epidemic is given by

$$\sum_{n=0}^{\infty} i_n = \frac{i_0}{1-\mu}, \quad (\mu < 1).$$

Thus, if $\mu \geq 1$, the total number infected goes to infinity. Finally, note that the epidemic becomes extinct once $I_n = 0$, since zero is an absorbing state. Let η_i be the probability that epidemic eventually becomes extinct starting with $I_0 = i$. Let $\eta = \eta_1$. Since an epidemic started by $I_0 = i$ individuals is equivalent to i independent epidemics started in parallel each by a single individual, we see that

$$\eta_i = \eta^i, \quad i \geq 0.$$

Let p_k be the probability that a single individual infects k new individuals ($k \geq 0$). Then we can show that η satisfies

$$\eta = \sum_{k=0}^{\infty} p_k \eta^k.$$

This has a unique solution in $(0, 1)$ if $\mu > 1$. If $\mu \leq 1$, then $\eta = 1$ is the only solution, that is, the epidemic dies out with probability 1.

What if the infectious individual remains infectious with probability p after each period? (The above model is the special case $p = 0$). This can be reduced to the standard branching process model with μ replaced by $\mu + p$ and σ^2 replaced by $\sigma^2 + p(1-p)$.

4.2.3 Stochastic Model: Continuous Time

Let $I(t)$ be the number of infectious individuals at time t. Each individual has an Exp(γ) lifetime and produces new infections according to a $PP(\beta)$ during its lifetime. All individuals behave independently of each other. Then $\{I(t), t \geq 0\}$

can be seen to be a birth and death process with birth rates $\lambda_n = n\beta$ and death rates $\mu_n = n\gamma$. It is also called a linear growth process. Let

$$p_i(t) = P(I(t) = i|I(0) = 1).$$

One can show that (using $\rho = \beta/\gamma$)(see Kulkarni [78])

$$p_0(t) = \frac{1 - \exp((\beta - \gamma)t)}{1 - \rho \exp((\beta - \gamma)t)},$$

$$p_i(t) = \rho^{i-1}(1 - \rho) \exp((\beta - \gamma)t) \frac{(1 - \exp(\beta - \gamma)t)^{i-1}}{(1 - \rho \exp((\beta - \gamma)t))^{i+1}}, \quad i \geq 1.$$

If $I(0) = k > 1$, one can think of $I(t)$ as a sum of k iid linear growth processes, each starting with one individual.

Next let $m_k(t) = E(I(t)|I(0) = k)$. Then we can derive

$$m_k(t) = k \exp((\beta - \gamma)t).$$

Similarly, let $v_k(t) = \mathrm{Var}(I(t)|I(0) = k)$. Then we can derive

$$v_k(t) = k \exp((\beta - \gamma)t) \frac{1 - \exp(-(\beta - \gamma)t)}{\beta - \gamma}.$$

Thus the mean is the same as the solution of the deterministic model. However, the behavior of the stochastic model is fundamentally different than that of the deterministic one. For example, we can get information about the variance $v_k(t)$ from the stochastic model, which is absent in the deterministic model. Also, let η_k be the probability that the epidemic starting with $I(0) = k$ eventually dies out. One can show that $\eta_k = 1$ if $\rho \leq 1$. Otherwise we get

$$\eta_k = \rho^{-k}, \quad \rho > 1.$$

In the deterministic model the probability of dying out is zero if $\rho > 1$ and 1 if $\rho \leq 1$.

4.3 SI Model

Now we consider the situation where an infectious individual infects a susceptible individual at rate β, and then dies (or is removed) at rate γ. Thus the disease progression model is $S \to I$. Sometimes this model is appropriate for carrier borne epidemics.

4.3.1 Deterministic Model: Weiss Model

We follow the development in Weiss [139]. Let $s(t)$ be the number of susceptible individuals at time t, $i(t)$ be the number of infected individuals (carriers) at time

t. The dynamics of the system is given by

$$\begin{aligned} s'(t) &= -\beta s(t) i(t), \\ i'(t) &= -\gamma i(t). \end{aligned}$$

Thus the carriers (the infectious individuals) die (or are identified and quarantined and treated) at rate γ per carrier, while the infection is spread from the carrier population to the susceptible according to the law of mass action with proportionality constant β. These equations can be easily solved to get

$$\begin{aligned} i(t) &= i_0 e^{-\gamma t}, \\ s(t) &= s_0 e^{-\sigma i_0 (1 - e^{-\gamma t})}, \end{aligned}$$

where $\sigma = \beta/\gamma$. Thus $i(t)$ goes to zero as $t \to \infty$, while the limiting number of survivors is

$$s(\infty) = s_0 e^{-\sigma i_0}.$$

4.3.2 Stochastic Model: Weiss Model

Next we follow the development in Weiss [139] and describe their stochastic equivalent of the deterministic model of the previous section. Let $S(t)$ be the number of susceptible individuals at time t, $I(t)$ be the number of infected individuals (carriers) at time t. Suppose a carrier stays infectious for an $\exp(\gamma)$ amount of time and then is removed from the population. Any two individuals come into to contact with each other after an $\exp(\beta)$ amount of time. If a carrier individual comes in contact with a susceptible individual, the susceptible individual gets infected, and is instantaneously removed. Let $S(0) = m$ and $I(0) = n$. Thus $\{(S(t), I(t)), t \geq 0\}$ is a CTMC on state space $\{(i,j) : 0 \leq i \leq m, 0 \leq j \leq n\}$ with transition rates

$$\begin{aligned} q_{(i,j),(i-1,j)} &= \beta i j, \\ q_{(i,j),(i,j-1)} &= \gamma j. \end{aligned}$$

Clearly states $(i, 0)$ are absorbing and other states are transient. Let

$$\pi_k(m,n) = \lim_{t \to \infty} P(S(t) = k | S(0) = m, I(0) = n).$$

Weiss [139] derives the following closed form solution:

$$\pi_k(m,n) = \binom{m}{k} \sum_{j=k}^{m} (-1)^{j-k} \binom{m-k}{j-k} (j\sigma + 1)^{-n}, \quad 0 \leq k \leq m.$$

From this we can show that

$$E(S(\infty)) = \lim_{t \to \infty} E(S(t) | S(0) = m, I(0) = n) = m(1+\sigma)^{-n}.$$

For small values of σn, $E(S(\infty)) \sim s(\infty)$, else, $E(S(\infty)) > s(\infty)$. Also, the stochastic model gives the variance in the number of survivors, a deterministic model cannot do that.

4.3.3 Deterministic Model: Fixed Population

One can develop another type of SI model, where the infectious individuals stay infectious permanently, that is, $\gamma = 0$. However, to make the model realistic, we have assume that the total size of the population is fixed at N. See Chapter 2 of Daley and Gani [26]. Thus, the dynamics is given by

$$\begin{aligned} s'(t) &= -\beta s(t) i(t), \\ i'(t) &= \beta s(t) i(t). \end{aligned}$$

Thus $s'(t) + i'(t) = 0$, and hence $s(t) + i(t) = s(0) + i(0) = N$. We can eliminate $s(t)$ and write

$$i'(t) = \beta i(t)(N - i(t)).$$

The solution is given by

$$i(t) = \frac{i(0)N}{i(0) + s(0)\exp(-\beta N t)}.$$

Thus if $i(0) > 0$, then $i(t) \to N$ as $t \to \infty$, following a logistic curve.

4.3.4 Stochastic Model: Fixed Population

We now consider the stochastic process $\{(S(t), I(t)), t \geq 0\}$ that is the equivalent to the Deterministic model of the previous subsection. The state space is given by $\{(i, j) : i + j = N\}$ The transition rates are given by

$$q_{(i,j),(i-1,j+1)} = \beta i j.$$

Thus, $\{I(t), t \geq 0\}$ is a pure birth process on $\{0, 1, \cdots, N\}$ with birth rates

$$\lambda_n = \beta n(N - n).$$

One can use the general theory of pure birth processes to compute the distribution of $I(t)$ as follows (see Kulkarni [78]):

$$P(I(t) = n | I(0) = i) = \sum_{k=i}^{n} A_{k,n} e^{-\lambda_n t}, \;\; 0 \leq i \leq n \leq N,$$

where

$$A_{k,n} = \frac{1}{\lambda_n} \prod_{r=i, r \neq k}^{n} \frac{\lambda_r}{\lambda_r - \lambda_k}.$$

The above expression is valid when the birth rates λ_n are all distinct. Otherwise the expressions get complicated. See Daley and Gani [26]. Surprisingly, there is no simple expression for $E(I(t))$.

4.4 SIS Model

Some diseases, like flu, common cold, etc, do not bestow immunity from future infections. In such a case, a susceptible individual may get infected and once the infection is cured will become susceptible again. Thus the disease progression is $S \to I \to S$. We follow the development in Weiss and Dishon [140]

4.4.1 Deterministic Model: Fixed Population

The dynamics is given by

$$\begin{aligned} s'(t) &= -\beta s(t)i(t) + \gamma i(t), \\ i'(t) &= \beta s(t)i(t) - \gamma i(t). \end{aligned}$$

Thus $s'(t) + i'(t) = 0$, and hence $s(t) + i(t) = s(0) + i(0) = N$. Eliminating $s(t)$ we get

$$i'(t) = (\beta N - \gamma)i(t) - \beta i(t)^2,$$

which can be solved to get

$$i(t) = \frac{(\beta N - \gamma)i(0)}{\beta i(0) + (\beta(N - i(0)) - \gamma)\exp(-(\beta N - \gamma)t)}.$$

Thus, assuming $i(0) > 0$, we have

$$\lim_{t\to\infty} i(t) = \frac{\beta N - \gamma}{\beta}, \quad \text{if } \beta N > \gamma.$$

If $\beta N \leq \gamma$, $i(t) \to 0$ as $t \to \infty$.

4.4.2 Stochastic Model: Fixed Population

Let $I(t)$ be the number of infectious individuals at time t. Clearly, $\{I(t), t \geq 0\}$ is a birth and death process on $\{0, 1, \cdots, N\}$ with birth rates $\lambda_n = \beta n(N - n)$ and death rates $\mu_n = \gamma n$. Thus state 0 is absorbing and all other states are transient. Using standard results from Birth and Death processes (see Kulkarni [78] and Weiss and Dishon [140]), one can also compute the expected time until absorption into state 0, that is the expected lifetime of the epidemic. Equating birth and death rates we get

$$\lambda_n = \mu_n \Rightarrow \beta n(N - n) = \gamma n \Rightarrow n = n^* = \frac{\beta N - \gamma}{\beta}.$$

Thus if $\beta N > \gamma$, the drift $\delta_n = \lambda_n - \mu_n$ is positive for $n \leq n^*$ and negative for $n > n^*$. Thus n^* is a "stable state", any deviation from it will push the system back to it. If $\beta N \leq \gamma$, then $n^* \leq 0$ and 0 is the only stable state. However, in the stochastic system, state zero is always absorbing and the system will eventually get there and stay there.

4.5 The SIR Model

Next we consider the SIR model, also called the general epidemic model, for Susceptible-Infected-Recovered model of the disease progression. We assume the following model of the spread of an epidemic: disease infects a susceptible individual (from outside the population) and converts him/her into an infected individual. An infected individual can infect other susceptible individuals by contact. Once the infected individual recovers, he/she is immune from further infections by the same disease.

4.5.1 Deterministic Model

The earliest deterministic SIR model was analyzed by Kermack and McKendrick [77]. It forms a building block of all stochastic models that follow later on.

Let $s(t)$ be the number of susceptible individuals at time t, $i(t)$ be the number of infected individuals at time t and $r(t)$ be the number of recovered individuals at time t. Kermack and McKendrick [77] postulate the following continuous dynamics of the system:

$$\begin{aligned} s'(t) &= -\beta s(t) i(t), \\ i'(t) &= \beta s(t) i(t) - \gamma i(t), \\ r'(t) &= \gamma i(t). \end{aligned}$$

Here β is called the infection rate and γ is called the recovery rate. (When s, i and r are interpreted as fractions it is customary to use the parameter $\lambda = N\beta$ in place of β, where N is the population size.) Note that the new infections occur with rate proportional to the product of $s(t)$ and $i(t)$. We begin with the initial conditions $s(0) = s_0 > 0$, $i(0) = i_0 > 0$ and $r(0) = r_0 \geq 0$. Clearly

$$s'(t) + i'(t) + r'(t) = 0$$

hence

$$s(t) + i(t) + r(t) = s_0 + i_0 + r_0 = N.$$

Thus the size of the population remains constant. This is a reasonable approximation if the rates of births and deaths of the individuals are much less compared to the infection and recovery rates. This is called the SIR model. For many variations and extensions of this model, see the wikipedia article on Epidemic Models at en.wikipedia.org/wiki/Epidemic_model.

We quote from Hethcote [63] the main properties of the solution $\{(s(t), i(t)), t \geq 0\}$, using $\sigma = \beta/\gamma$,

Theorem 4.1 1. If $\sigma s_0 < 1$, $i(t)$ decreases monotonically from i_0 to 0 as $t \to \infty$.

2. If $\sigma s_0 > 1$, $i(t)$ increases from i_0 to

$$i_{max} = N - r_0 - N(1 + \ln(\sigma s_0))/\sigma,$$

and then decreases monotonically to 0 as $t \to \infty$.

3. $s(t)$ decreases monotonically from s_0 to s_∞ as $t \to \infty$, where s_∞ is the unique solution in $(0, 1/\sigma)$ to

$$N - r_0 - s_\infty + \ln(s_\infty/s_0)/\sigma = 0.$$

4. $r(t)$ increases monotonically from r_0 to $N - s_\infty$ as $t \to \infty$.

One can think of $\sigma s(t)$ as the replacement number at time t. As long as it is greater than one, the the number of infected increases, and when it falls below 1, the epidemic begins to subside. (Thus $\sigma s(t)$ plays the role of μ from the branching process model.)

4.5.2 Stochastic Model: Discrete Time

We next discuss a discrete time stochastic model of the evolution of epidemic, proposed by Reed and Frost in 1928. Now let S_n, I_n and R_n represent the number of susceptible, infected and removed individuals at the beginning of the nth period. Each individual stays infectious for one period of time, and in the period can infect any susceptible individual with probability p, in an independent fashion.. Thus we have

$$S_{n+1} = \text{Bin}(S_n, (1-p)^{I_n}), \quad n \geq 0,$$

$$I_{n+1} = S_n - S_{n+1} \sim \text{Bin}(S_n, 1 - (1-p)^{I_n}), \quad n \geq 0,$$

$$R_{n+1} = R_n + I_n, \quad n \geq 0.$$

We can rewrite these equations as the following difference equations:

$$S_{n+1} - S_n = -\text{Bin}(S_n, 1 - (1-p)^{I_n}), \quad n \geq 0,$$

$$I_{n+1} - I_n = \text{Bin}(S_n, 1 - (1-p)^{I_n}) - I_n, \quad n \geq 0,$$

$$R_{n+1} - R_n = I_n, \quad n \geq 0.$$

This clearly shows that

$$S_n + I_n + R_n = N,$$

where N is the populations size. Let s_n, i_n and r_n denote the expected values of S_n, I_n and R_n. If p is small, we can approximate $(1-p)^{I_n}$ by $1 - pI_n$. Using this approximation and taking expectations we get

$$\Delta s_n = -p s_n i_n, \;\; \Delta i_n = p s_n i_n - i_n, \;\; \Delta r_n = i_n,$$

where $\Delta f_n = f_{n+1} - f_n$ is the discrete derivative. These equations are similar to the equations of the SIR model, and have similar properties. Note that we have used $E(S_n I_n) = s_n i_n$, which is true only if S_n and I_n are independent, but this is not true in this model. Thus the expected value equations above are approximate.

In the above model, the expected infections caused by a single infected individual is pS_n, which is bounded above by pN, which can be large. One way to limit this is to assume that each individual makes at most a random number K of contacts in one period (with replacements uniformly among the $N-1$ remaining individuals). Thus the probability that a susceptible individual escapes infection in period n is $(1 - I_n/(N-1))^K$. Thus

$$S_{n+1} = \text{Bin}(S_n, 1 - (1 - I_n/(N-1))^K), \quad n \geq 0,$$

$$I_{n+1} = S_n - S_{n+1} \sim \text{Bin}(S_n, (1 - I_n/(N-1))^K), \quad n \geq 0,$$

$$R_{n+1} = R_n + I_n, \quad n \geq 0.$$

4.5.3 Stochastic Model: Continuous Time

Let $S(t)$, $I(t)$ and $R(t)$ represent the number of susceptible, infected and removed individuals at time t. Suppose an infected individual stays infectious for an $\exp(\gamma)$ amount of time and then becomes immune (and thus removed from further infections.) Any two individuals come into to contact with each other after an $\exp(\beta)$ amount of time. If an infectious individual comes in contact with a susceptible individual, the susceptible individual gets infected. The removed individuals stay removed forever. The total number of individuals in the society stays fixed at N, i.e.,

$$S(t) + I(t) + R(t) = N, \quad t \geq 0.$$

We can model $\{(S(t), I(t)), t \geq 0\}$ as a CTMC on state space $\{(i,j) : 0 \leq i+j \leq N\}$ with infinitesimal rates as follow:

$$q_{(i,j),(i-1,j+1)} = \beta ij,$$

$$q_{(i,j),(i,j-1)} = \gamma j.$$

The states $(i, 0)$ are absorbing, and the other states are transient. This model is considered by Daniels [25].

We have

$$E(S(t+h) - S(t) | S(t) = i, I(t) = j) = -\beta hij + o(h),$$

$$E(I(t+h) - I(t) | S(t) = i, I(t) = j) = \beta ijh - \gamma jh + o(h).$$

Writing $s(t) = E(S(t))$ and $i(t) = E(I(t))$, we can write the above equations as

$$s'(t) = -\beta E(S(t)I(t)),$$

$$i'(t) = \beta E(S(t)I(t)) - \gamma i(t).$$

If we could write $E(S(t)I(t)) = s(t)i(t)$, these would be the same equations as in the SIR model. But, $S(t)$ and $I(t)$ are dependent and hence $E(S(t)I(t)) \neq s(t)i(t)$. Hence the solutions of the SIR equations do not yield the expected values of the the $S(t)$ and $I(t)$ processes. The deterministic SIR equations are called the fluid equivalent of the stochastic system.

Computing the limiting behavior of the (S, I) process is not trivial. We restate some results from Daniels [25]. Fix a $0 \leq k \leq N$ and define

$$a_{i,j} = \lim_{t\to\infty} P(S(t) = k, I(t) = 0 | S(0) = i, I(t) = j).$$

Clearly $a_{i,j} = 0$ if $i < k$, and $a_{i,0} = \delta_{i,k}$, where $\delta_{i,k} = 1$ if $i = k$ and zero otherwise. Using first step analysis we get, for $j > 0$, $i > k$,

$$a_{i,j} = \frac{i}{\rho + i} a_{i-1,j+1} + \frac{\rho}{\rho + i} a_{i,j-1}, \tag{4.1}$$

$$a_{k,j} = \frac{\rho}{\rho + k} a_{k,j-1}, \tag{4.2}$$

where $\rho = \gamma/\beta$. Solving Equation 4.2 recursively, we get

$$a_{k,j} = \left(\frac{\rho}{\rho + k}\right)^j.$$

Using this in Equation 4.1 we can recursively obtain $a_{i,j}$. Closed form expressions and asymptotic forms are given in Daniels [25].

In the above model we have assumed that every contact with an infected individual makes a susceptible individual infected. As mentioned before, Weiss [139] considers a carrier borne epidemic where such a contact confers immunity on the susceptible individual. Thus in his model the infected and the susceptible numbers both decline with time.

Downton [34] considers a generalization of both models above: he assumes that a contact between an infected and a susceptible individual infects the susceptible individual with probability π and imparts immunity with probability $1 - \pi$. The Danils model is a special case with $\pi = 1$ and the Weiss model is a special case $\pi = 0$. Picard [109] uses clever Martingale methods to derive interesting properties of the Downton model.

Recommended for Presentation: Weiss [139], Downton [34], Picard [109], Nasell [98], Kuske et al [80].

4.6 Optimal Control of Epidemics

There are many reasons for controlling epidemics: there is economic cost from lost productivity, health cost to the society from exploding epidemics, etc. There are many ways of controlling epidemics: education, vaccination, identification and quarantines, etc. Roughly speaking education is an effort to reduce β, the infection rate, by asking people to wash hands, not to spit in public, use condoms, etc. Vaccination is an effort to reduce the number of susceptibles, and quarantine is an effort to remove the infectious individuals before they can spread the disease. In vector borne diseases, such as malaria, one can control the population of the intermediate group such as mosquitoes. We shall study a few simple models here, although there is a vast literature on this topic.

4.6.1 Optimal Vaccination: Static Policy

First consider the case where we can implement a mass vaccination program whereby we can instantly vaccinate a given number (or fraction) of the susceptible population. What fraction should we choose? The static SIS model of Section 4.5.1 gives insight into how much of the population needs to be vaccinated (and thus rendered immune) in order to prevent epidemics. Suppose we want to ensure that the number of infected individuals never increases. Thus we need to ensure that $\sigma s_0 < 1$. Thus if a fraction f is vaccinated, we have $s_0 = N(1-f)$. Hence we must have $\sigma N(1-f) < 1$, or $f > 1 - 1/(N\sigma)$.

This analysis ignores the cost of the vaccination program. An alternative is to assume that each infection costs c dollars and each vaccination costs v dollars. Suppose initially we have i_0 infected and no recovered individuals. We decide to vaccinate n susceptible individuals (out of a total of $N - i_0$). Then $s_0 = N - n - i_0$. The total number of infected individuals in such an epidemic is given by $r_\infty(n) = r_\infty$ which is given by the unique positive solution to

$$r_\infty + \ln((N - r_\infty)/(N - n - i_0))/\sigma = 0.$$

Thus the total cost of this policy is

$$C(n) = nv + cr_\infty(n).$$

One can then choose an n that minimizes the above cost.

4.6.2 Optimal Vaccination: Dynamic Programming

When instantaneous vaccination is not feasible, we need to decide at what rate we should vaccinate the population. This leads to a dynamic control problem. We follow Hethcote and Walton [62] and study a deterministic SIS epidemic model described by the differential equations in Section 4.5.1. Our aim is to find an optimal vaccination policy described by a function α, where $\alpha(t)$ is the

vaccination rate at time t. Let $v(t)$ be the total number vaccinated by time t ($v(0) = 0$). Thus the system of differential equations now becomes

$$\begin{aligned} s'(t) &= -\beta s(t)i(t) - \alpha(t), \\ i'(t) &= \beta s(t)i(t) - \gamma i(t), \\ r'(t) &= \gamma i(t), \\ v'(t) &= \alpha(t). \end{aligned}$$

Suppose it costs $c(u)$ per unit time to perform u vaccinations per unit time. Suppose each infectious individual costs h per unit time. Then the optimal control problem can be formulated as

$$\min \int_0^\infty e^{-\delta t}[c(\alpha(t)) + hi(t)]dt$$

subject to the above differential equations and the initial conditions

$$s(0) = s_0, i(0) = i_0, r(0) = r_0, v(0) = v_0.$$

Here $\delta > 0$ is the continuous discount factor. (Hethcote and Weiss [62] consider a different objective function than the one given above.) They solve the problem by using dynamic programming algorithm. We illustrate their method below.

We discretize the time horizon as $0, d, 2d, \cdots,$, and the vaccination rates are constrained to take values in a finite set $A = \{\alpha_1, \alpha_2, \cdots, \alpha_N\}$. We assume that $\alpha(t)$ is constant over the intervals $[nd, (n+1)d)$. Furthermore, let $s_n = s(nd+)$, etc. Then the above differential equations implicitly give a mapping Γ_a for each $a \in A$ such that, if $\alpha(t) = a$ for $t \in [nd, (n+1)d)$,

$$(s_{n+1}, i_{n+1}) = \Gamma_a(s_n, i_n)$$

as follows:

$$\begin{aligned} s_{n+1} &= s_n - \beta s_n i_n d - ad, \\ i_{i+1} &= i_n + \beta s_n i_n d - \gamma i_n d. \end{aligned}$$

Let $C(s, i, n)$ be the total minimum discounted cost of operating the system optimally up to time nd, starting from state (s, i) at time 0. The Dynamic Programming equation is given by

$$C(s, i, n) = \min_{a \in A}\{c(a)d + h(i) + e^{-\delta d}c(\Gamma_a(s, i), n - 1)\}.$$

Here $h(i)$ can be approximated by hid if d is small. The above equation can be solved by successive approximation and the optimal vaccination rate a over an interval can be obtained as a function of (s, i) at the beginning of the interval.

4.6.3 Optimal Vaccination: Optimal Control

We can also use the tools from optimal control theory to solve this problem. This is done by Sethi and Staats [121] for the SI model with fixed population (see Section 4.3.3, and we summarize their results below. They use linear cost $c(u) = cu$, and assume that $0 \leq u \leq U < \infty$. Thus, the maximum vaccination rate is U. Thus they solve:

$$\min \int_0^\infty e^{-\delta t}[cu(t) + hi(t)]dt$$

subject to

$$\begin{aligned} s'(t) &= -\beta s(t)i(t) - u(t), \\ i'(t) &= \beta s(t)i(t), \\ 0 \leq u(t) &\leq U. \end{aligned}$$

Following the standard method of solving this type of optimal control problem we first form the Hamiltonian

$$H = -(cu + hi) + \lambda\beta si + \mu(-\beta si - u),$$

Where λ and μ are the adjoint variables (similar to Lagrange multipliers). The optimal solution satisfies the above differential equations, and also the following differential equations

$$\lambda' = -\partial H/\partial i = \delta\lambda + h - \lambda\beta s + \mu\beta s$$

$$\mu' = -\partial H/\partial s = \delta\mu - \lambda\beta i + \mu\beta i.$$

The main idea is to choose an u that minimizes H. Since the only u-dependent term in H is $-(c + \mu)u$, it is clear that the optimal control is

$$u(t) = \begin{cases} U & \text{if } \mu(t) + c > 0 \\ 0 & \text{if } \mu(t) + c < 0 \end{cases}$$

If $\mu(t) + c = 0$, any u would do, such a case is called singular. Using this Sethi and Staats show that in the case with no discounting, ($\delta = 0$), $\mu(t)$ is a monotonically decreasing function of t, and hence there exists a $t^* \geq 0$ such that the optimal control is to use $u^*(t) = U$ for $t \leq t^*$ and $u^*(t) = 0$ for $t \geq t^*$. They provide many results about bounds on t^*.

Recommended for Presentation: Sethi and Staats [121], Lefevre [82], Yan eta al [147].

4.6.4 Optimal Quarantine

Another method of controlling an epidemic is to isolate or quarantine the susceptible individual when he gets infected. This removes them from causing further

infections among the susceptibles. We follow the development in Sethi [124]. They assume the SIS epidemic model (see Section 4.5.1), whose dynamics is given by

$$\begin{aligned} s'(t) &= -\beta s(t)i(t) + \gamma i(t), \\ i'(t) &= \beta s(t)i(t) - \gamma i(t). \end{aligned}$$

Here we treat s and i as fractions of the population. We assume that the symptoms are visible immediately after infection sets in. Suppose we isolate a fraction $1 - u(t)$ of the infectious individuals at time t, so that $u(t)i(t)$ number of infectious individuals are actively spreading the disease to the $1 - i(t)$ susceptible individuals. We shall deal with the SIS deterministic epidemic model of Section 4.5.1. The dynamics is given by

$$i'(t) = \beta u(t)i(t)(1 - i(t)) - \gamma i(t),$$

and the number of susceptibles is given by $s(t) = 1 - i(t)$. Suppose it costs h dollars per unit time per infected individual and w dollars per unit time to quarantine an individual. Thus the objective function is to minimize the total discounted cost

$$\int_0^\infty e^{-\delta t}[w(1 - u(t))i(t) + hi(t)]dt.$$

Sethi [124] shows that the optimal quarantine policy is 0-1, i.e., it is either optimal to quarantine everybody all the time ($u^*(t) = 0$ for all $t \geq 0$) or quarantine nobody ($u^*(t) = 1$ for all $t \geq 0$). In particular,

$$\theta = \frac{w + h}{h} \frac{\beta}{\gamma + \delta} \leq 1 \Rightarrow u^*(t) = 1 \quad \forall t \geq 0.$$

If $\theta > 1$, which of the two extreme policies is optimal depends on the value of $i(0)$. Note that this assume that it is feasible to at instantaneously quarantine $1 - u(0)$ fraction at time 0. It will be interesting to develop a model that explicitly accounts for upper limit on how many patients can be quarantined per unit time.

Recommended for Presentation: Gani, Yakowitz and Blount [43], who consider a stochastic model of quarantine of HIV positive individuals in a prison system. Behncke [11], Reluga [113].

4.7 More Topics in Epidemiology

We have broached only the bare minimum of epidemic models. There is a vast literature on epidemics on non-homogeneous populations, epidemics on networks, game theoretic considerations in vaccinations, to name a few. The students are encouraged to explore this area further if it interests them. There are also interesting papers on game-theoretic models in control of epidemics, see Reluga [113].

Chapter 5

Queueing Models for Resource Planning

5.1 Background

A hospital has several types of resources: beds, doctors, nurses, emergency rooms, operating rooms, medical equipment (such MRI machines), etc. The usual objective is to set the levels of these resources to a certain utilization level, such as: occupancy ratio of hospital beds should be at least 75%, the MRI machine should be busy at least 90% of the time, etc. Sometimes, staffing is specified as a ratio: the nurse to patient ratio should be at most 8:1 in general ward, for example. These requirements generally ignore the effect on the waiting times of the patients. Recently there have been recommendations that the probability of waiting for a bed should be at most 1%, for example. This has motivated the use of queueing models in capacity planning. We shall concentrate on these models in this chapter. The survey papers by Green [49] and Fomundam and Hermann [39] are good starting points for this material. We shall briefly review the queueing models as needed.

5.2 Birth and Death Queues: Background

Let $\{X(t), t \geq 0\}$ be a Continuous Time Markov Chain (CTMC) on $\{0, 1, 2, \cdots\}$ with infinitesimal generator $Q = [q_{ij}]$. It is called a birth and death process with birth rate $\{\lambda_i, i \geq 0\}$ and death rates $\{\mu_i, i \geq 1\}$ if the positive transition rates are given by:

$$q_{i,i+1} = \lambda_i, \;\; i \geq 0, \; q_{i,i-1} = \mu_i, \;\; i \geq 1.$$

The birth and death process is positive recurrent if

$$\sum_{i=0}^{\infty} \rho_i < \infty$$

where

$$\rho_0 = 1, \;\; \rho_i = \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i}, \;\; i \geq 1.$$

The limiting distribution of a positive recurrent birth and death process is given by

$$p_j = \lim_{t \to \infty} P(X(t) = j) = \frac{\rho_j}{\sum_{i=0}^{\infty} \rho_i}, \;\; j \geq 0.$$

The two most commonly used birth and death queues are the $M/M/c$ queue and the $M/M/c/c$ queue. We collect some important results about them below.

5.2.1 $M/M/c$ System

The $M/M/c$ queue has c servers and infinite waiting space. The arrivals occur according to a PP(λ) and the service times are iid exp(μ). Let $X(t)$ be the number of customers in the system at time t. It is a birth and death process with birth rates

$$\lambda_n = \lambda, \;\; n \geq 0$$

and death rates

$$\mu_n = \min(n, c)\mu, \;\; n \geq 1.$$

This system is stable if $\rho = \lambda/(c\mu) = r/c < 1$, where $r = \lambda/\mu$. The limiting distribution is given by

$$p_j = \frac{r^j}{j!} p_0, \quad 0 \leq j < c,$$

$$p_j = \frac{\rho^j c^c}{c!} p_0, \quad j \geq c,$$

where p_0 is given by

$$p_0 = \left[\sum_{j=0}^{c-1} \frac{r^j}{j!} + \frac{r^c}{c!} \frac{1}{(1-\rho)} \right]^{-1}.$$

The expected number of busy servers in steady state is given by λ/μ, hence the "occupancy rate" (fraction of servers busy) is ρ. The probability that a customer has to wait for service is given by $p_c/(1-\rho)$. This quantity is sometimes called the Erlang-C formula, denoted by $C(c, r)$, where $r = \lambda/\mu$. We have

$$C(c, r) = \frac{\frac{r^c}{c!} \frac{c}{c-r}}{\sum_{j=0}^{c-1} \frac{r^j}{j!} + \frac{r^c}{c!} \frac{c}{c-r}}. \tag{5.1}$$

The expected number of customers waiting for service is

$$L_q = \frac{\rho}{(1-\rho)} C(c, r),$$

and the expected number in the system is

$$L = L_q + r.$$

The expected queueing time (time to start service) is

$$W_q = L_q/\lambda = \frac{1}{\mu}\frac{1}{c-r}C(c,r),$$

and the expected waiting time in the system is

$$W = L/\lambda = W_q + \frac{1}{\mu}.$$

One can choose ρ or W_q or W as the performance measure in determining c.

5.2.2 Halfin-Whitt Asymptotics for the $M/M/c$ Queue

The blocking probability is given by Equation 5.1 for the $M/M/c$ queue. In the $M/M/c$ queue the blocked customers wait until a server becomes available. It is intuitively clear that if we fix λ and increase c, the blocking probability will go to zero, and if we fix c and increase λ (or equivalently, increase the ratio λ/μ)), the blocking probability will go to one. Thus, if we want to study large systems with the blocking probability strictly between zero and one, say α, we must have large λ/μ and large c. Such an asymptotic regime is called Halfin-Whitt regime, first studied by Halfin and Whitt [58]. We summarize the related results below. Let ϕ and Φ be the pdf and cdf of a standard Normal random variable and

$$h(x) = \frac{\phi(x)}{1-\Phi(x)} \tag{5.2}$$

be the hazard rate of a standard normal random variable. We call the asymptotic regime where

$$HW = \{r \to \infty,\ \ c \to \infty,\ \ c = r + \beta\sqrt{r} + o(\sqrt{r})\},$$

where $\beta > 0$ is a constant, the Halfin-Whitt asymptotic regime.

Theorem 5.1 In the Halfin-Whitt asymptotic regime,

$$\lim_{HW} C(c,r) = \frac{h(-\beta)}{\beta + h(-\beta)}.$$

Proof: Let X be a $\mathrm{P}(r)$ random variable. Then, in the Halfin-Whitt asymptotic regime we have

$$\begin{aligned} e^{-r}\frac{r^c}{c!} &= P(c-1 < X \le c) \\ &= P((c-1-r)/\sqrt{r} < (X-r)/\sqrt{r} \le (c-r)/\sqrt{r}) \\ &\to \frac{1}{\sqrt{r}}\phi(\beta). \end{aligned}$$

Similarly

$$\begin{aligned}\sum_{j=0}^{c-1} e^{-r}\frac{r^j}{j!} &= P(X < c-1) \\ &= P((X-r)/\sqrt{r} \le (c-r-1)/\sqrt{r}) \\ &\to \Phi(\beta).\end{aligned}$$

Using these limits we get

$$\begin{aligned}C(c,r) &= \frac{\frac{r^c}{c!}\frac{c}{c-r}}{\sum_{j=0}^{c-1}\frac{r^j}{j!} + \frac{r^c}{c!}\frac{c}{c-r}} \\ &= \frac{e^{-r}\frac{r^c}{c!}\frac{c}{c-r}}{\sum_{j=0}^{c-1} e^{-r}\frac{r^j}{j!} + e^{-r}\frac{r^c}{c!}\frac{c}{c-r}} \\ &\to \frac{\frac{1}{\sqrt{r}}\phi(\beta)\frac{\sqrt{r}}{\beta}}{\Phi(\beta) + \frac{1}{\sqrt{r}}\phi(\beta)\frac{\sqrt{r}}{\beta}} \\ &= \frac{\phi(\beta)}{\beta\Phi(\beta) + \phi(\beta)} \\ &= \frac{h(-\beta)}{\beta + h(-\beta)},\end{aligned}$$

where the last equality follows from $\phi(\beta) = \phi(-\beta)$ and $\Phi(\beta) = 1 - \Phi(-\beta)$. This completes the proof.

Note that the hazard rate function $h(x)$ of a standard normal random variable is an increasing function of x, increasing from 0 at $x = -\infty$ to $\sqrt{2/\pi}$ at $x = 0$, and growing to $+\infty$ as $x \to \infty$. Hence $\frac{h(-\beta)}{\beta + h(-\beta)}$ decreases from 1 at $\beta = 0$ to zero as $\beta \to \infty$. Hence there is a unique β that satisfies

$$\frac{h(-\beta)}{\beta + h(-\beta)} = \alpha, \tag{5.3}$$

for any given $\alpha \in (0,1)$. The solution $\beta = \beta(\alpha)$ to the above equation is a decreasing function of α. Thus to design an $M/M/c$ queue with a given large $r = \lambda/\mu$ so as to achieve a given probability of blocking α, we first compute $\beta(\alpha)$ that satisfies Equation 5.3 and use

$$c = r + \beta(\alpha)\sqrt{r}.$$

Note that the smaller the α, the larger the $\beta(\alpha)$, and hence the large the c. This makes intuitive sense. This system is highly utilized, since the fraction of busy servers is given by

$$r/c = r/(r + \beta\sqrt{r}) \approx 1 - \beta/\sqrt{r} \approx 1.$$

That is almost all servers are busy. Furthermore, the expected queueing time is given by

$$W_q = \frac{\alpha}{\mu}\frac{1}{\beta\sqrt{r}},$$

which is also fairly small, since r is large. Thus we have a designed an efficient system that gives high quality of service (as measured by α and W_q. hence this design is sometimes called Quality and Efficiency Driven (QED) design.

How is it possible that in the Halfin-Whitt regime, all servers are almost always busy, but the queueing time is not overly long? This seems to go against the folk theorem that if you keep the server busy, the wait will be long (efficiency vs. throughput tradeoff). The answer is: the tradeoff is a feature of a system with finite number of servers. However, the Halfin-Whitt regime holds when the number of serves is very large.

5.2.3 $M/M/c/c$ System

The $M/M/c/c$ queue has c servers and no waiting space. The arrivals occur according to a PP(λ) and the service times are iid exp(μ). A customer that finds all servers busy is lost. Let $X(t)$ be the number of customers in the system at time t. It is a birth and death process with birth rates

$$\lambda_n = \lambda, \;\; 0 \le n < c$$

and death rates

$$\mu_n = n\mu, \;\; 1 \le n \le c.$$

This system is always stable. The limiting distribution is given by

$$p_j = \frac{r^j/j!}{\sum_{k=0}^{c} r^k/k!}, \quad 0 \le j \le c,$$

where $r = \lambda/\mu$. The expected fraction of customers lost is given by p_c. This is also called the Eralng-B formula, and is denoted by

$$B(c, r) = \frac{r^c/c!}{\sum_{k=0}^{c} r^k/k!} \tag{5.4}$$

The expected number of busy servers (also equal to the number of customers in the system) in steady state is

$$L = r(1 - p_c).$$

One can choose p_c or L as the performance measure in determining c. It is an interesting result in queueing theory that these formulas remain valid for general service time distributions with mean $1/\mu$, i.e., for $M/G/c/c$ queues.

5.3 $M/G/1$ Model

Another common queueing model is the $M/G/1$ model. Here the customers arrive according to a PP(λ). The service times are iid with mean τ and variance σ^2. Let $X(t)$ be the number of customers in the system at time t. $\{X(t), t \geq 0\}$ is stable if $\rho = \lambda\tau < 1$. The expected number in the system in a stable system is given by

$$L = \lim_{t\to\infty} E(X(t)) = \rho + \frac{\rho}{2(1-\rho)}(1 + \sigma^2/\tau^2).$$

The probability that the server is busy is given by ρ. The expected waiting time in the system is given by $W = L/\lambda$.

Keller and Laughhunn [73] use the $M/G/1$ model to determine the optimal number of doctors to use in a clinic. There are n identical doctors available, where n has to be decided. They assume that patients arrive according to a PP(λ) and the service times are iid with mean τ and variance σ^2 regardless of which doctor sees them. Suppose the patients are evenly divided among the doctors, thus each doctor gets a dedicated Poisson stream of patients with rate λ/n. Each doctor costs c_d per day, while the cost of waiting for the patients is c_p dollars per patient per day. The total cost per day is

$$C = c_d n + nL(n)c_p,$$

where $L(n)$ is the number of patients in each of the n queues, given by (with $\rho = \lambda\tau$)

$$L(n) = \rho/n + \frac{\rho}{2(n-\rho)}(1 + \sigma^2/\tau^2).$$

Now one can choose an n that minimizes the above. One can show that the optimal n^* is given by

$$n^* = \frac{1}{\rho}\left[1 + \sqrt{\frac{c_p}{2c_d}(1 + \sigma^2/\tau^2)}\right].$$

Clearly we must choose one of the two integers nearest to n^*.

5.4 $M/G/c/c$ Model

The standard $M/G/c/c$ loss model has been found to be quite useful in modeling a health care facility with a finite number of beds, and where patients do not queue up if no beds are available. (They are sent somewhere else.) See McManus et al [93] and Gorunescu et al [45]. Note that the queue length process in an $M/M/c/c$ system is a birth and death process. We assume that the arrivals for a PP(λ), the service times are iid with common mean τ, there are c beds, and patients who do not find an empty bed upon arrival are lost. Let $X(t)$ be the number of occupied beds at time t, and let

$$p_j = \lim_{t\to\infty} P(X(t) = j), \quad 0 \leq j \leq c.$$

It is well known that, regardless of the actual distribution of the service time, the limiting distribution is given by

$$p_j = \frac{\rho^j/j!}{\sum_{i=0}^{c} \rho^i/i!}, \ \ 0 \le j \le c,$$

where $\rho = \lambda\tau$. One can show that the fraction of patients lost is given by p_c due to PASTA (Poisson Arrivals See Time Averages.) The expected number of occupied beds in steady state is given by

$$L = \rho(1 - p_c).$$

Now suppose each empty bed costs h per day, each lost patient costs π and each admitted patient generates a revenue of p per day spent in the system. Then the long run average net revenue is given by

$$R(c) = -h(c - L) + pL - \pi p_c = (p + h)L - hc - \pi p_c.$$

Thus one can choose c to maximize the above quantity. McManus et al [93] find that time spent by a patient in the clinic is well described by an exponential distribution, while Gorunescu et al [45] fit a mixture of exponentials.

5.4.1 Asymptotics for $M/G/c/c$ queue

Following the development in the $M/M/c$ queue, we can study the asymptotic behavior of the $M/G/c/c$ queue, motivated by the following design issue: suppose we are given $r = \lambda/\mu$, and a desired blocking probability α. We want to find the smallest value of c such that the blocking probability $B(c, r)$ given by Equation 5.4 is at most α.

Since the Halfin-Whitt regime analyzed in Section 5.2.2 had highly desirable properties in the $M/M/c$ queue, we firsst analyze the behavior of the $M/G/c/c$ queue under that regime. The main result is given in the theorem below.

Theorem 5.2 Suppose

$$c = r + \beta\sqrt{r}.$$

Then, as $r \to \infty$,

$$P(\text{Blocking}) \sim \frac{h(-\beta)}{\sqrt{r}}$$

where h is the normal hazard rate as defined in Equation 5.2.

Proof: The blocking probability in an M/G/c/c queue is given by $B(c, r)$ of Equation 5.4. The theorem follows by following the proof of Theorem 5.1.

One can use the above theorem to design an $M/G/c/c$ system to achieve a particular probability of blocking (say α) by setting

$$\frac{h(-\beta)}{\sqrt{r}} = \alpha.$$

Solving this we get

$$\beta = \beta(r) = -h^{-1}(\sqrt{r}\alpha).$$

The suggested design is

$$c = r + \beta(r)\sqrt{r}. \tag{5.5}$$

Note that β will be negative if $r > \sqrt{2/\pi}/\alpha$. One can see numerically that the above c behaves almost linearly in r.

Let $c = c(r, \alpha)$ be a solution to

$$B(c, r) = \alpha.$$

It is numerically observed that

$$c = (1 - \alpha)r + \frac{1 - \alpha}{\alpha} \tag{5.6}$$

provides a better approximation to $c(r, \alpha)$ than the HW approximation of Equation 5.5 for large r. Note that the c given by Equation 5.6 is essentially the efficiency driven regime (ED).

5.5 The Finite Population Model

The earliest papers applying queueing models to hospital waiting lists arose out the experience of National Health Services in UK. One is by Duncan and Curnaow [35], and another is by Worthington [144]. We briefly describe the models by Worthington here. He models a hospital waiting list as a finite source queue $M/G/s/N/N$ with finite user population N. This can be thought of as the number of beds in an hospital (always full) and each person produces requests for service at a fixed rate α. If the service times are iid exponential(β), we see that $X(t)$, the number of patients waiting for (or in) service becomes a birth and death queue on state space $\{0, 1, \cdots, N\}$ with birth rates

$$\lambda_n = (N - n)\alpha, \;\; 0 \leq n < N,$$

and death rates

$$\mu_n = \min(n, s)\beta, \quad 1 \leq n \leq N.$$

When the service times are general, there are no exact results (or even efficient numerical algorithms) to analyze the system. Assuming the expected service time is $1/\beta$ and the coefficient of variations (standard deviation/mean) is cv,

the author claims that the limiting distribution of $X(t)$ can be reasonably approximated by a Normal distribution with mean

$$L = N - s\beta/\alpha,$$

and variance

$$V = \frac{s\beta}{2\alpha}(1 + cv^2).$$

This approximation works well when $L > 2.6\sqrt{V}$ (i.e., the queue is rarely empty) and $cv^2 < 3$ (the service time is not too variable). The author validates the model with actual data.

5.6 Bed Assignment Model

In the $M/G/c/c$ model we assume that all patients are treated equally, i.e., a patient is admitted if there is a bed available. Here we present a model with two classes of customer, inspired by that of Esogbue and Singh [37]. Suppose patients can be classified as belonging to two classes, 1 and 2. Class i patients arrive according to PP(λ_i) and and require Exp(μ_i) service times. There are a total of N beds, and the admission policy is described a single number $m < N$ as follows: at most m beds can be occupied by type 2 patients at any time. Thus $N - m$ beds are reserved for type 1 patients. Suppose the cost of turning away a patient of type i is c_i ($c_1 > c_2$), while the cost of an empty bed is c per bed per unit time. The aim is to choose m to minimize the average cost per unit time.

Let $X_i(t)$ be the number of patients of type i in the system at time t. The $\{(X_1(t), X_2(t)), t \geq 0\}$ is a CTMC, and its steady state distribution can be computed numerically. Let

$$p_{ij} = \lim_{t\to\infty} P(X_1(t) = i, X_2(t) = j), \;\; 0 \leq j \leq m, \;\; 0 \leq i \leq N - j.$$

Then the long run average cost is given by

$$\Psi(m) = c_1 \sum_{j=0}^{m} p_{N-j,j} + c_2 \sum_{i=0}^{N-m} p_{i,m} + c \sum_{j=0}^{m} \sum_{i=0}^{N-j} (N - i - j) p_{ij}.$$

One can then pick the optimal m that minimizes $\Psi(m)$.

Now consider a special case where $\mu_1 = \mu_2$, see Shonick and Jackson [123]. Then the above policy admits class 2 patients as long as the total number of patients is less than m, else rejects them, and admits type 1 patients as long as there is a bed available. Now define $X(t)$ be the total number of occupied beds. It is easy to see that $\{X(t), t \geq 0\}$ is a birth and death process on state space $\{0, 1, \cdots, N\}$, with birth parameters

$$\lambda_n = \lambda_1 + \lambda_2, \;\; 0 \leq n < m,$$

$$\lambda_n = \lambda_1, \;\; m \leq n < N,$$

and death parameters

$$\mu_n = n\mu, \;\; 0 \leq n \leq N.$$

Thus one can easily compute the limiting probabilities as follows:

$$p_n = \lim_{t\to\infty} P(X(t) = n) = \frac{\rho_n}{\sum_{j=0}^{N} \rho_j}, \;\; 0 \leq n \leq N,$$

where

$$\rho_n = ((\lambda_1 + \lambda_2)/\mu)^n/n!, \;\; 0 \leq n \leq m,$$
$$\rho_n = ((\lambda_1 + \lambda_2)/\mu)^m(\lambda_1/\mu)^{n-m}/n!, \;\; m < n \leq N.$$

The long run average cost is given by

$$\Psi(m) = c_1 p_N + c_2 \sum_{n=m}^{N} p_n + c \sum_{n=0}^{N} (N-n)p_n.$$

One can choose m to minimize $\Psi(m)$.

5.7 Acute Care Facility Model

We present a model inspired by Weiss and Mclain [141]. Consider an acute care facility (ACF) where patients arrive according to a PP(λ'). When a patient is discharged from the ACF, three possible outcomes occur:

1. The patient leaves (goes home, dies) with probability θ_1,
2. The patient needs an alternate level care (ALC) in an extended care facility (ECF) and finds a place in such a facility immediately, with probability θ_2,
3. The patient needs ALC in an ECF, but the ECF is full, with probability θ_3.

The patients in the third category are force to stay in the ACF until a place becomes available in the ECF. They are called ALC (Alternative Level Care) patients and experience "administrative delay". The aim is to build a tractable model of $X(t)$, the number of patients experiencing administrative delay at time t. Weiss and McLain model $\{X(t), t \geq 0\}$ as a birth and death process with arrival rate $\lambda = \lambda'\theta_3$ and death rates given by

$$\mu_n = 0, \;\; n \leq n^*,$$

$$\mu_n = \alpha + n\mu, \;\; n > n^*.$$

One can think of $\{\mu_n, n \geq 0\}$ as modeling the placement process by which the ALC patients are placed into ECF. Thus suppose places in the ECF become

available according to a PP(α), and the ALC patients leave the ALC status (because they get better or die) after exp(μ) time while waiting for the ECF. This captured by setting $n^* = 0$.

The birth and death process is stable if $\mu > 0$, or if $\mu = 0$ and $\alpha > \lambda$. Now let $p_n = \lim_{t\to\infty} P(X(t) = n)$, $n \geq n^*$. Note that $P_n = 0$ for $n < n^*$. Define $\rho_{n^*} = 1$ and

$$\rho_n = \prod_{n=n^*+1}^{n} \frac{\lambda}{\alpha + n\mu}, \quad n > n^*.$$

Using the theory of birth and death processes we get

$$p_n = \frac{\rho_n}{\sum_{n=n^*}^{\infty} \rho_n}, \quad n \geq n^*.$$

We can show that, if $\mu > 0$,

$$L = \lim_{t\to\infty} E(X(t)) = \frac{\lambda}{\mu} + n^* p_{n^*} - \frac{\alpha}{\mu}(1 - p_{n^*}).$$

The average administrative delay is given by L/λ, from Little's law.

If $\mu = 0$ and $\lambda < \alpha$, $\{X(t) - n^*, t \geq 0\}$ is a stable $M/M/1$ queue.

Weiss and McLain [141] use actual data to estimate α, n^* and μ and validate the model by comparing the mean administrative delay.

An alternative model considers two units: the ACF and the ECF. Patients arrive at the ACF according to PP(λ'), and stay there for iid lengths of time with common mean τ_a. ACF has infinite capacity. A fraction θ_1 of them leave after their service is completed in the ACF. The rest need to spend iid random amount of time (with common mean τ_e) in the ECF. The ECF has a finite capacity C, and if the ECF is full the patient stays in the ACF until a space opens up in the ECF at which time he is transferred there. Now let $X_1(t)$ be the number of patients in the ACF who are getting acute care, $X_2(t)$ be the number of patients in the ACF who are waiting for a space to open up in the ECF, and $X_3(t)$ be the number of patients in the ECF, at time t. Then $\{X_1(t), t \geq 0\}$ is the queue length process in an $M/G/\infty$ process with arrival rate λ' and mean service time τ_a. $\{X(t) = X_2(t) + X_3(t), t \geq 0\}$ is the queue length process in an $M/G/\infty$ queue with arrival rate $\lambda = \lambda'(1 - \theta_1)$ and mean service time τ_e. We have

$$X_2(t) = \max(X(t) - C, 0), \quad X_3(t) = \min(X(t), C).$$

Now, in steady state $X(t)$ is a P(ρ) random variable, where $\rho = \lambda\tau_e$. Then we have

$$L_2 = \lim_{t\to\infty} E(X_2(t)) = \sum_{k=C}^{\infty} (k - C)e^{-\rho}\frac{\rho^k}{k!},$$

$$L_3 = \lim_{t\to\infty} E(X_3(t)) = \sum_{k=0}^{C} k e^{-\rho}\frac{\rho^k}{k!} + C\sum_{k=C+1}^{\infty} e^{-\rho}\frac{\rho^k}{k!}.$$

Thus the expected administrative delay is given by L_2/λ. The probability that a patient gets a place in the ECF when needed is

$$\theta_2 = \sum_{k=0}^{C-1} e^{-\rho}\frac{\rho^k}{k!}.$$

Now suppose it costs h_2 dollars per unit time to keep a patient in the ACF waiting for a room in the ECF, while it costs $h_3 < h_2$ dollars per unit time to keep a patient in the ECF. Then the total cost per unit time is given by

$$\Psi(C) = h_2 L_2 + h_3 L_3.$$

Hence we can choose the optimal C to minimize the cost of the ACF-ECF system.

5.8 Snowball Effect Model

Here we present a model studied by Selen et al [119]. It tries to capture the phenomenon observed commonly in health care systems: the longer a patient waits for service the longer is the service time. (This is true up to a certain time. If the wait is excessive, the patient may die and need zero service time!)

They model the service system as an s-server system where patients arrive according to a PP(λ). If an incoming patient sees an idle server she is called a non-delayed patient, and needs an exp(μ_H) amount of service. However, if she finds all servers busy when she arrives, she is called a delayed patient, and needs exp(μ_L) amount of service. We assume that $\mu_L < \mu_H$, indicating that a delayed patient needs larger service times. All service times are independent. Let $X(t)$ be the total number of patients in the system and $Y(t)$ be the number of non delayed patients in the system. Thus $\{(X(t), Y(t)), t \geq 0\}$ is a CTMC on state-space $S = \{(i,j) : i \geq 0, 0 \leq j \leq s\}$. We call it the slowdown system.

For $0 \leq i < s$, the transition rates are

$$(i,j) \to (i+1, j+1) : \lambda,$$
$$(i,j) \to (i-1, j-1) : j\mu_H,$$
$$(i,j) \to (i-1, j) : (i-j)\mu_L.$$

For $i \geq s$, the transition rates are

$$(i,j) \to (i+1, j) : \lambda,$$
$$(i,j) \to (i-1, j-1) : j\mu_H,$$
$$(i,j) \to (i-1, j) : (s-j)\mu_L.$$

Let $\rho_H = \lambda/(s\mu_H$ and $\rho_L = \lambda/(s\mu_L)$. The system is stable if $\rho_L < 1$. Clearly, if $\mu_H = \mu_L = \mu$, the system reduces to a standard $M/M/s$ queue. They use $M/M/s$ queue with $\mu = \mu_H$ (call it the fast system) as the lower bound and with $\mu = \mu_L$ (call it the slow system) as an upper bound to bound the number in the system and the waiting times in the slowdown system. They show that ignoring the effect of slowdown can lead to serious under provisioning the system (using smaller s than needed.) They also show that the stationary distribution can exhibit multi-modal behavior.

It is interesting to consider the asymptotic behavior of the system as s, the number of servers goes to infinity. The Quality and Efficiency Driven (QED) regime is captured by the following parameterization:

$$\lambda^{(s)} = s\mu_L(1 - \beta/\sqrt{s}),$$
$$\mu_H^{(s)} = \mu_L(1 + \gamma/\sqrt{s}),$$

where $\beta > 0$ and $\gamma > 0$ are fixed parameters. Let $W^{(s)}$ be the queueing time in the system with s servers in steady state, and let Φ be the cdf and ϕ be the pdf of a standard normal.Then one can show that $P(W^{(s)} > 0)$ converges to $P(W > 0)$ as $s \to \infty$, and although exact expression for $P(W > 0)$ is not known, the following bounds are derived in [119]:

$$\left(1 + \beta\frac{\Phi(\beta)}{\phi(\beta)}\right)^{-1} \geq P(W > 0) \geq \left(1 + (\beta + \gamma)\frac{\Phi(\beta + \gamma)}{\phi(\beta + \gamma)}\right)^{-1}.$$

These bounds can be used to determine an appropriate s for a queueing system with given parameters λ, μ_L and μ_H. One can choose appropriate β to ensure desired upper bound on $P(W > 0)$, say .01. Then, for a given λ, μ_L we can choose an appropriate s, using the asymptotic equation

$$\lambda = s\mu_L(1 - \beta/\sqrt{s}).$$

With this s and the known μ_H, we can compute γ by using

$$\mu_H = \mu_L(1 + \gamma/\sqrt{s}).$$

Using this β and γ we can compute the lower bound on $P(W > 0)$.

5.9 Semi-Markov Process: Background

Let $\{X(t), t \geq 0\}$ be a stochastic process on state space $\{0, 1, 2 \cdots\}$ with piecewise constant right continuous sample paths. Let S_n be the time of the nth jump in the process. $\{X(t), t \geq 0\}$ is called a semi-Markov process (SMP) if

$$P(X(S_{n+1+} = j, S_{n+1} - S_n \leq y | X(S_n+) = i, X(t), 0 \leq t \leq S_n) = p_{ij}Q_{ij}(y), \quad i, j \geq 0, y \geq 0.$$

In other words, when an SMP enters state i, it jumps to state j with probability p_{ij}, and given that the next jump is to state j, the sojourn time in state i has cdf $Q_{ij}(\cdot)$. An SMP has Markov property at each jump point S_n. Let

$$\tau_i = E(S_1 - S_0 | X(S_0) = i)$$

be the expected sojourn time in state i. Let $X_n = X(S_n+)$. Then $\{X_n, n \geq 0\}$ is a DTMC with transition probability matrix $P = [p_{ij}]$. It called the embedded DTMC in the SMP. Suppose it is recurrent with a positive vector $\pi = [\pi_0, \pi_1, \cdots]$ satisfying

$$\pi = \pi P.$$

The SMP $\{X(t), t \geq 0\}$ has a proper limiting distribution if

$$\sum_{i=0}^{\infty} \pi_i \tau_i < \infty.$$

(We also need the SMP to be aperiodic, but that condition is satisfied if the sojourn times are continuous random variables.) The limiting distribution is given by

$$p_j = \lim_{t \to \infty} P(X(t) = j) = \frac{\pi_j \tau_j}{\sum_{i=0}^{\infty} \pi_i \tau_i}, \quad j \geq 0.$$

5.10 Semi-Markov Model

The models so far model a hospital as a single queue. Next we present a model inspired by Hershey et al [61] that removes this restriction. (See also the papers by Kao [71], and [70].) Suppose a hospital consists of n units. Each unit can handle unlimited number of patients. Patients arrive at unit i according to $PP(\lambda_i)$ per day. A patient arriving at unit i stays there for an average of τ_i days and then moves to unit j with probability p_{ij}, or leaves the hospital with probability r_i, where

$$\sum_{j=1}^{n} p_{ij} + r_i = 1, \quad 1 \leq i \leq n.$$

Let $P = [P_{ij}]$ be the routing matrix. Assume that $I - P$ is invertible. This ensures that every patient arriving at the hospital eventually leaves. Assume that patient movements are independent of each other.

Let a_i be the rate at which patients arrive at unit i, externally or internally. The effective rates a_j's are given by the unique solution to

$$a_j = \lambda_j + \sum_{i=1}^{n} a_i p_{ij}, \quad 1 \leq j \leq n.$$

Letting $a = [1_1, \cdots, a_n]$ and $\lambda = [\lambda_1, \cdots, \lambda_n]$, we get

$$a = \lambda + aP.$$

Since $I - P$ is invertible, we get

$$a = \lambda (I - P)^{-1}.$$

Since the sojourn time in unit i is τ_i per visit, we can use Little's law to compute L_i, the expected number of patients in unit i in steady state as :

$$L_i = a_i \tau_i, \;\; 1 \leq i \leq n.$$

What happens when the units have finite capacity? Hershey et al consider the case when unit 1 has a finite capacity c. Assume that when a patient arriving at unit 1 finds it full, she is accommodated elsewhere in the hospital, but the future behavior (routing) of the patient remains the same. Let α be the probability that unit 1 is full. One can use Erlang loss formula to compute this as follows:

$$\alpha = \frac{\rho^c / c!}{\sum_{j=0}^{c} \rho^j / j!}$$

where $\rho = a_1/\tau_1$. This formula is exact in the following cases:

1. The sojourn times in unit 1 are general, and the patient in unit one never returns to unit one after leaving it.

2. The sojourn times in all units are exponential, and the patient in unit one can return after leaving it.

If one of the above condition is not satisfied, the Erlang loss formula represents an approximation. In any case, L_i^*, the expected number of patients in each unit can now be computed as

$$L_1^* = (1-\alpha)L_1, \;\; L_i^* = L_i, \;\; 2 \leq i \leq n.$$

Also, the expected number of overflow patients from unit 1 in steady state is given by αL_1.

5.11 Open Jackson Networks: Background

An open Jackson network consists of N service stations, each with infinite waiting room. The ith station has s_i servers and iid Exp(μ_i) service times. Customers arrive from out side to node i according to a PP(λ_i). After completing service at node i a customer joins the queue at station j with probability r_{ij}, or leaves the network with probability r_i. Clearly,

$$\sum_{j=1}^{N} r_{ij} + r_i = 1, \quad 1 \leq i \leq N.$$

Let $R = [r_{ij}]$ be the routing matrix and assume that $I - R$ is invertible. This ensure no customer stays in the network forever. Let a_j be the effective arrival rate (internal plus external) at node j. We have

$$a_j = \lambda_j + \sum_{i=1}^{N} a_i r_{ij}, \quad 1 \leq j \leq N.$$

Thus

$$a\lambda(I-R)^{-1},$$

where

$$a=[a_1,a_2,\cdots,a_N],\ \ \lambda=[\lambda_1,\lambda_2,\cdots,\lambda_N].$$

The network is stable id

$$a_i<s_i\mu_i,\quad 1\le i\le N.$$

$X_i(t)$ be the number of customers at node i at time t. The main result of Jackson networks states that the limiting distribution of $X_i(t)$ is the same as that in an $M/M/s_i$ queue with arrival rate a_i and service rate μ_i, and in steady state all queue lengths are independent of each other. Using these results it is easy to compute L_i, the expected number of customers at station i in steady state. Then the expected waiting time in steady state in station i is given by $W_i=L_i/a_i$.

5.12 Queueing Network Model

Queueing networks is an ideal model for describing patient movements among various units in a hospital. We first describe a model by Koizumi et al [76]. Note that the open Jackson network model assumes that there is infinite waiting space at each station. Koizume et al suggest an approximation when there is no waiting room at the nodes. This is the case when the servers at node i represents the number of beds at unit i. The approximation proceeds as follows:

1. Assume infinite waiting room and compute W_i's as explained in the previous subsection,
2. Compute the effective service time at station i by

$$1/\tilde{\mu}_i=1/\mu_i+\sum_{j=1}^{N}r_{ij}W_j,$$

3. Compute the new W_i using these modified service times.

The approximation works especially well if the Jackson network is acyclic, that a patient cannot visit the same station more than once. Koizume et al [76] apply this method to a mental health system consisting of three nodes and validate it with real data.

An alternate approximation would be as follows:

1. For station i with capacity c_i, compute

$$p_{ij}=\frac{a_i^j/j!}{\sum_{k=0}^{c_i}a_i^k/k!},\quad 0\le j\le c_i$$

2. Compute the effective service time at station i as

$$1/\tilde{\mu}_i = 1/\mu_i + \sum_{j=1}^{N} r_{ij} \frac{p_{jc_j}}{c_j \mu_j}.$$

3. Analyze the network assuming these new effective service rates.

If the network has cycles, the steps 1 and 2 of the above process need to be carried out repeatedly until the effective service rates converge.

Xie et al [145] model the total service time of a patient as a Coxian distribution with M phases, and derive detailed performance measures of number of patients in various phases. They explain it as a Jackson network, but it really is an $M/G/c/c$ queue.

Recommended for Presentation: Jiang and Giachetti [68], Creemers and Lambrecht [24], [54].

5.13 Time Varying Arrivals

A common feature at many health care facilities is time varying arrival rate. The rates vary by time of day, day of the weak and season of the year. They also vary based on whether the day is before or after a holiday. All the queueing models discussed above assume constant arrival rates. In this section we discuss how this gap can be bridged.

5.13.1 $M(t)|G|\infty$ Queue

A tractable model of an arrival process with time varying arrival rate is to assume that it is a non-homogenous Poisson process (NHPP) with rate function $\{\lambda(t), -\infty < t < \infty\}$. Assume that

$$\Lambda(t) = \int_{-\infty}^{t} \lambda(u)du < \infty, \quad -\infty < t < \infty.$$

Consider a queueing system with infinite number of servers, general service times with distribution $G(\cdot)$, and a NHPP arrival process. Such a queueing system is called an $M(t)|G|\infty$ queue. See Eick et al [36]. Let S be a generic random variable with

$$G(x) = P(S \leq x),$$

and S_e be a random variable representing the remaining service time, which has distribution

$$G_e(x) = P(S_e \leq x) = \frac{1}{E(S)} \int_0^x (1 - G(u))du.$$

Let $X(t)$ be the number of customers in such a system at time t. Then it is known that $X(t)$ is a Poisson random variable with parameter

$$m(t) = E\int_{-\infty}^{t} \lambda(u)(1 - G(t-u))du = E(\lambda(t - S_e))E(S). \tag{5.7}$$

If the service time is deterministic with mean τ, we have

$$m(t) = \int_{t-\tau}^{t} \lambda(u)du.$$

Thus we can can think of $m(t)/\tau$ as the moving average of the arrival rate over the last τ time units. As another example, suppose the arrival rate function is given by

$$\lambda(t) = \lambda(1 + \alpha\sin(\beta t)), \;\; -\infty < t < \infty,$$

where $\alpha \in [-1, 1]$ is a fixed constant. Thus the arrival rate fluctuates between $\lambda(1-\alpha)$ and $\lambda(1+\alpha)$, and attains its maximum at times $t = 2n\pi$ and minimum at $t = (2n+1)\pi$ for integer n's. We can show that

$$m(t) = \lambda\tau[1 + \alpha(\sin(\beta t)E(cos(\beta S_e)) - cos(\beta t)E(\sin(\beta S_e)))].$$

This achieves its extrema at any t satisfying

$$\tan(\beta t) = \frac{E(\cos(\beta S_e))}{E(\sin(\beta S_e))}.$$

A general observation is that the maximum in $m(\cdot)$ occurs after the maximum in $\lambda(\cdot)$. This lagged behavior is intuitive in retrospect, and is useful in designing service system.

5.13.2 Time Varying Staffing

The material in this subsection is based on the paper by Green et al [52]. They model a service system as $M(t)|G|s(t)+GI$ queue, where arrival occur according a NHPP($\lambda(\cdot)$), service times are iid with mean τ, cdf $G(\cdot)$, number of servers at time t is $s(t)$, and the patient impatience times are iid. (A patient leaves the system without service if his queueing time is longer than his impatience time.) The aim is to devise an optimal staffing policy, that is, the function $s(\cdot)$, so that the total staff hours are minimized, subject to a quality level constraint, such as: 80 percent of all service requests must begin service without delay, or the abandonment rate must be less than 5 percent, etc.

PSA: Point-wise Stationary Approximation: At each time t compute the performance measure (such as the delay probability) assuming a stationary system with arrival $\lambda = \lambda(t)$, and compute the needed staffing level $s(t)$. One can use the Halfin-Whitt square-root formula if appropriate. This is a reasonable approximation when the services are short and quality requirements are high.

Segmented PSA: Many times, changing staffing levels continuously with time is not feasible, and we are forced to keep them, constant over intervals of

length T, called the staffing intervals. Over the nth staffing interval $[nT, (n+1)T)$, we could compute

$$s_n = \max\{s(t) : nT \leq t < (n+1)T\}$$

and then set the staffing level over the nth interval as s_n. This is called the segmented PSA approach.

SIPP: Statinary Independent Period by Period: For the interval $[nT, (n+1)T)$ consider a stationary queueing system with arrival rate

$$\lambda_n = \frac{1}{T}\int_0^T \lambda(nT+u)du$$

and assume that the system is in steady state over the entire interval. Using this, compute the staffing level s_n. Then use the staffing level function $s(t) = s_n$ for $nT \leq t < (n+1)T$. (See Green et al [48]). This works well for short service times, short staffing intervals, and slowly varying arrival rates.

SPHA: Simple Peak Hour Approximation: This method is used for short service times and long staffing intervals (say 10 minutes service times, and 8 hour staffing interval). Here we compute the peak arrival rate

$$\lambda_n = \max\{\lambda(t) : nT \leq t < (n+1)T\}$$

and compute an s_n assuming a stationary system with arrival rate λ_n.

OLM: Offered Load Models: When service times are short, we turn to the infinite server models with NHPP arrivals as studied before and compute the $m(t)$ from Equation 5.7. We then use $\lambda(t) = m(t)/E(S)$ and then use the PSA or SIPP. This induces a lag in the arrival rate. An even simpler method to induce time lag is to use

$$\hat{\lambda}(t) = \lambda(t - E(S^2)/2E(S)).$$

Green et al [48] suggest several other approximations that induce such a lag. Also see .

SDPP: Stationary Dependent Period by Period: Yu et al [149] suggest another approximation that seems to be appropriate when the service times are comparable or longer than the staffing intervals. Let X_n be the number of patients in the system at the end of the nth staffing interval. Then they estimate the arrival rate for the next staffing interval $[nT, (n+1)T)$ by

$$\lambda_n = \frac{X_n + \int_{nT}^{(n+1)T} \lambda(u)du}{T + E(S)}.$$

Then the staffing level for the nth interval is computed assuming the arrival process is Poisson with this rate and it is in steady state over the whole interval. The term X_n in the numerator makes behaviour over the nth interval dependent on what happened in the previous interval. Hence the name SDPP.

There are many papers related to time varying systems. See Madelbaum et al [90] and Massey and Whitt [92].

Chapter 6

Appointment Scheduling

6.1 Background

Patients face two types of delays in accessing a health care facility: the indirect delay and the direct delay. Suppose a patient calls for an appointment at time t_c and is given the appointment at time $t_a \geq t_c$. Then the indirect delay is given by $t_a - t_c$. It is generally measured in days. Suppose the customer keeps the appointment and punctually arrives at the service facility at time t_a and is then seen by the doctor at time t_d. The direct delay is given by $t_d - t_a$. It is generally measured in minutes.

Thus we can think of a health care facility as two queues in series: the appointment queue followed by the clinic queue. A patient arrives at the appointment queue when he requests an appointment. He stays (virtually, not physically) in the appointment queue until the appointment time and then leaves the appointment queue and joins the clinic queue. He waits for service in the clinic queue, and departs after service is complete. The time spent in the appointment queue is the indirect delay, while the time spent waiting in the clinic queue is the direct delay.

The length of the appointment slot is generally fixed, although the actual time spent with the doctor is random. If we choose short slot duration, the indirect delay will go down (since we give out more appointments) but the direct delay will go up (since the actual service time with the doctor is more likely to exceed the slot length). The reverse happens if we use longer slot durations.

The indirect delay affects the probability of no-shows and cancellation. The longer the indirect delay, the higher the probability that customer may not keep the appointment, or will cancel it. This patient behavior in health care systems has an important effect on their efficiency.

Many models have been proposed and analyzed for these two queues separately, and a few for the two queues together. We shall study them in this chapter.

In the early 2000's (see Murray and Tantau [96]) a system called *Open Access* was highly recommended. Its basic premise was "do today's work today". In effect it advocated doing away with appointments and just treat everybody as a walk-in patient. However, this leaves no control over day to day variation in demand. The main aim of an appointment schedule is to smooth out the demand.

6.2 Direct Waiting: Transient Models

We begin with the single-server queueing models of direct waiting. Let S_n be the service time of the n-th patient, A_n be the arrival time of the n-th patient, and W_n be the queueing time of the nth patient (not including time in service). Then we have the following recursion:

$$W_1 = 0, \quad W_{n+1} = \max\{W_n + S_n - (A_{n+1} - A_n), 0\}, \quad n \geq 1.$$

This is called Lindley's Equation, and can, in theory be used to compute the distribution of the queueing time of the nth patient. In practice this is highly intractable, and simulation is the best approach.

The above equation assumes that the doctor is available at time 0. What happens if the doctor arrives at time $T > 0$. Then the queuing time of the first patient becomes $W_1 = T$, and the above equation remains valid with this modified initial condition.

The idle time of the doctor (if any) after the nth patient finishes service is given by

$$I_n = \max\{A_{n+1} - (A_n + W_n + S_n), 0\}, \quad n \geq 1.$$

We can control the appointment times A_n. If the appointment times are close to each other, the idle time will reduce, but the waiting time will increase. We can also control N, the the number of patients with appointments. The doctor will be done seeing the last patient at time $A_N + W_N + S_N$. If the doctor's shift ends at B, then we can think of $\max\{A_N + W_N + S_N - B, 0\}$ as the overtime, and $\max\{B - (A_N + W_N + S_N), 0\}$ as the last idle time before the shift ends. The larger the N, the larger the overtime, and smaller the last idle time.

Let a be the cost of keeping a patient waiting for a unit time, b the cost of keeping the doctor on call for a unit of time, and c the additional cost of one unit of overtime for the doctor. The total expected queueing cost of patients + idle time cost of the doctor + overtime cost of the doctor is thus given by

$$\text{Min} \;\; a\sum_{i=1}^{N} \mathsf{E}(W_i) + b\mathsf{E}(A_N + W_N + S_N) + c\mathsf{E}(\max\{A_N + W_N + S_N - B, 0\}).$$

For a given N, the aim is to find the optimal scheduled arrival times $\{A_n, 1 \leq n \leq N\}$ to minimize the above cost. Or, for a given shift length B, the aim is to find the optimal N and the scheduled arrival times to minimize the above cost.

6.2.1 Pegden-Roesnshine Model

Pegden and Rosenshine [103] consider the case with iid exp(μ) service times and no over time cost. Stein and Cote [126]) stream line these results. Assume that the patients arrive punctually at the appointed times. Then it is optimal to set $A_1 = 0$. The optimization problem reduces to

$$\text{Min } a\sum_{i=1}^{N} \mathsf{E}(W_i) + b\mathsf{E}(A_N + W_N + \frac{1}{\mu}).$$

When $N = 2$, the optimal appointment times are:

$$A_1 = 0; \;\; A_2 = \frac{1}{\mu}\ln(1/\gamma),$$

where $\gamma = b/(a+b)$. Let $X(t)$ be the number of patients in the system at time t and $X_i = X(A_i-)$, for $i = 1, 2, ..., n$. Then, $X_0 = 0$, $0 \le X_i \le i-1$ and $\{X_i, i \ge 0\}$ is a time-nonhomogenous DTMC with transition probabilities given below. Let

$$a_i(x) = e^{-\mu x}\frac{(\mu x)^i}{i!}, \quad b_i(x) = 1 - \sum_{k=0}^{i} a_i(x).$$

$$A(x) = \begin{bmatrix} b_0(x) & a_0(x) & 0 & 0 & \cdots & 0 \\ b_1(x) & a_1(x) & a_0(x) & 0 & \cdots & 0 \\ b_2(x) & a_2(x) & a_1(x) & a_0(x) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \end{bmatrix}.$$

The transition probabilities are given by

$$P(X_{i+1} = k | X_i = j) = A_{j,k}(x_i), \;\; x_i = A_{i+1} - A_i.$$

Let $A_N(x)$ be the $(N+1)$ by $(N+1)$ major submatrix of $A(x)$ (obtained by taking the first $N+1$ rows and columns of $A(x)$). Also let e be the row vector $[1, 0, \cdots, 0]$ of length $N+1$ and f be the column vector $[0, 1, 2, \cdots, N]$. Then we can write

$$\mathsf{E}(W_i) = \frac{1}{\mu} e \prod_{j=1}^{i-1} A_N(x_j) f, \quad i \ge 2.$$

Stein and Cote [126] show numerically that the optimal inter-appointment times exhibit a dome shaped structure, with the appointments in the beginning and towards the end being clustered close to each other and the appointments in the middle being more or less evenly spaced. When n is large, the middle appointments are mostly evenly spaced. Inspired by this observation they consider a model that assumes that appointments are equally spaced, that is $x_i = x$ for $i = 1, 2, \cdots, n-1$. Then $\{X_i, i \ge 0\}$ is a $D/M/1$ queue. It is stable if $\mu x > 1$. The steady state waiting time is

$$\mathsf{E}(W_i) = \frac{\beta}{\mu(1-\beta)} = w(x),$$

where β is the unique solution in $(0,1)$ to

$$\beta = e^{-\mu(1-\beta)x}.$$

The cost function can then be written as

$$w(x)(1+(1-\gamma)(N-2)) + \gamma(N-1)x.$$

This can be shown to be convex and the optimal x can be obtained numerically. Stein and Cote show that the optimal x is close to the average of the x_i's in the Pegden-Rosenshine model. For large N, Wang [134] suggests the following approximation:

$$x = \frac{1}{\mu}\left(1+\sqrt{\frac{1-\gamma}{2\gamma}}\right).$$

This model was further extended to allow no-shows by Hassin and Mendel [59]. They assume that each customer is liable to be a no-show with probability p. To compute the objective function we first define

$$B(x) = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ b_0(x) & a_0(x) & 0 & 0 & \cdots & 0 \\ b_1(x) & a_1(x) & a_0(x) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \end{bmatrix}.$$

The transition probabilities are given by

$$P(X_{i+1} = k | X_i = j) = (1-p)A_{j,k}(x_i) + pB_{j,k}(x_i).$$

Then we can write

$$\mathsf{E}(W_i) = \frac{1}{\mu} e \prod_{j=1}^{i-1} ((1-p)A_N(x_j) + pB_N(x_j))f, \quad i \geq 2.$$

This is the expected queueing time of the ith customer if he arrives. If he is a no-show, the expected waiting time is zero. Hence the objective function is

$$a(1-p)\sum_{i=1}^{N} \mathsf{E}(W_i) + b\mathsf{E}(A_N + W_N) + b(1-p)\frac{1}{\mu}.$$

The optimal appointment times have to be computed numerically. The authors confirm (numerically) that the appoints have a dome shaped structure. That is, the inter-appointment times of the initial and later patients are small while those of the patients in the middle are long.

Wang [136] extends the above model to allow phase type service times for the patients. The main contribution is in developing an algorithm to efficiently compute $\mathsf{E}(W_i)$.

Wang [137] considers the same model but under the assumption that the patient service times are not identically distributed. In particular, he assumes that the service times are independent, and the the service time of patient i is an $\exp(\mu_i)$ random variable. Now there are two problems: find the optimal order in which to serve the customers, and for that order, find the optimal appointment times. It is shown that it optimal to serve patients in the order of shortest service time variance first. In the exponential case, this is same as ordering according to increasing expected service times. That is, if $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_N$, then the optimal order of service is $1, 2, \cdots, N$. It is important that the waiting costs are the same for all patients.

Denton and Gupta [32] and Gupta and Denton [55] use 2-stage stochastic Linear Programming methods to solve scheduling problems. We do not go into the details here.

6.2.2 Service Interruptions

In the models above we assume that the doctor is always available. However, in practice, the doctor may have to interrupt the service to attend to emergencies. Wang [135] considers one such model. Here we describe a special case of a model considered by Luo et al [85]. Specifically, we consider the pre-emptive resume case, which assumes that a patient's service can be interrupted any time by an emergency call, and is simply resumed after the emergency visit is over.

Let $Z(t)$ be 1 if the doctor is available at time t, and 0 if he is not. Assume that $\{Z(t), t \geq 0\}$ is a two state CTMC with rates $q_{0,1} = \theta$ and $q_{1,0} = \eta$. This implies that emergency calls arrive according to PP(η), but have no effect if the doctor is already busy serving an emergency call. The emergency calls requires an $\exp(\theta)$ amount of time to serve.

Suppose the regular patients' service times are iid $\exp(\mu)$. Assume that $Z(0) = 1, A_1 = 0$. Let $X(t)$ be the number of customers in the system (in queue and in service) at time t. Then we see that $\{(X(t), Z(t)), t \geq 0\}$ is a Markov Regenerative process that behaves like a CTMC over the intervals $[A_n, A_{n+1})$ for $n = 0, 1, \cdots, N$. ($A_{N+1} = \infty$). Also, for $1 \leq n \leq N$,

$$Z(A_0+) = 1; \;\; X(A_0+) = 1; \;\; Z(A_n+) = Z(A_n-); \quad X(A_n+) = X(A_n-) + 1.$$

The above equations assume that all patients show up with probability 1. It is easy to account for no-show probabilities. Now consider a CTMC (X', Z') with state-space $\{(i, j) : 0 \leq i \leq N, j = 0, 1\}$ and non-zero transition rates given by

$$q_{(i,1),(i,0)} = \eta, \;\; q_{(i,1),(i-1,0)} = \mu, \;\; q_{(i,0),(i,1)} = \theta, \;\; 1 \leq i \leq N,$$

$$q_{(0,1),(0,0)} = \eta, \;\; q_{(0,0),(0,1)} = \theta.$$

Using this characterization, it is possible to compute

$$p'_{(i,j),(k,m)}(t) = P(X'(t) = k, Z'(t) = m | X'(0) = i, Z'(0) = j),$$

for $0 \leq k \leq i \leq N$, $j, m = 0, 1$, $t \geq 0$. Then we can compute the transient distribution of $(X(t), Z(t))$ as follows: For $0 = A_1 \leq t < A_2$,

$$p_{(1,1),(k,m)}(t) = P(X(t) = k, Z(t) = m) = p'_{(1,1),(k,m)}(t), \quad k, m = 0, 1,$$

and for $2 \leq n \leq N$, and $A_n \leq t < A_{n+1}$, $0 \leq k \leq n$, $m = 0, 1$,

$$p_{(1,1),(k,m)}(t) = P(X(t) = k, Z(t) = m) = \sum_{i=0}^{n} \sum_{j=0}^{1} p_{(1,1),(i,j)}(A_n -) p'_{(i+1,j),(k,m)}(t - A_n).$$

Now, the cost rates are given by

$$c(k, m) = a * k + m.$$

Then the total expected cost is given by

$$f(A_2, A_2, \cdots, A_N) = \sum_k \sum_m c(k, m) \int_0^\infty P(X(t) = k, Z(t) = m | X(0) = 1, Z(0) = 1).$$

Then one can find the decision variables $A_2, A_3, \cdots, A_N$ numerically that minimize the above function. Note that function f above is non-convex, and can have multiple local minima. Luo et al show by numerical experimentation that ignoring interruptions can lead to severly unoptimal schedules.

6.2.3 Slotted Schedules

In the previous subesections we studied continuous time models, where the appointment times were continuous variables, and service times were exponential random variables. Several authors have considered a slotted version of the problem, where the appointment horizon is divided into a slots of equal length and the appointments are given for the beginning of each slot. The service times maybe random.

We begin with the model by Kaandorp and Koole [69]. They assume T slots of length d each, and a total of N patients have to be scheduled over these slots. Let x_t be the number of patients scheduled to arrive at the beginning of slot t, $1 \leq t \leq T$. Clearly we must have $x_1 + \cdots + x_T = N$. Let $x = (x_1, \cdots, x_T)$ represent the schedule. They assume that there are no no-shows and all patients arrive on time. The service times are iid $\exp(\mu)$ and there is a single server.

Let Q_t be the number of patients in the system just before the patients scheduled at the beginning of the t-th slot arrive, and D_t be the number of departures during the t-th slot. Then we have

$$Q_1 = 0, \quad Q_{t+1} = Q_t + x_t - D_t, \quad t \geq 1.$$

Conditioned on $Q_t + x_t = j$, $D_t \sim \min(Y_t, j)$, where $\{Y_t, t \geq 1\}$ are iid $\mathrm{P}(\mu d)$ random variables. Thus the probability mass function of Q_t can be computed

recursively for increasing values of t. Then the total expected queueing time of all the patients is given by

$$W(x) = \frac{1}{\mu} \sum_{t=1}^{T} \sum_{i=1}^{x_t} \sum_{j=0}^{N} P(Q_t = j)(j + i - 1).$$

The expected over time is given by

$$L(x) = \frac{1}{\mu} \sum_{j=1}^{N} P(Q_{T+1} = j)j.$$

Suppose the doctor is paid a dollars per slot for the T slots, whether he is busy or idle. He is paid $b > a$ dollars per slot afterwards until all the scheduled patients are served. The waiting time of patients costs c dollars per slot per patient. Then the cost function is

$$C(x) = aT + bL(x) + cW(x)$$

This has to be minimized over all feasible integer-vectors x satisfying

$$x_i \geq 0, x_1 + \cdots + x_T = N.$$

This minimization has to be done numerically. The authors also consider the extension where patients are allowed to be be no-shows with a given probability in an independent fashion.

Muthuraman and Lawley [97] consider a similar model, with one major difference. They assume that the patients call one by one at the beginning of the day (before the first slot) and are assigned to one of the T slots, or are rejected, in a sequential fashion. They show that the optimal value function is in fact a unimodal function of number of patients that are scheduled, and hence once the local minima is reached, no more patients will be scheduled. They develop an algorithm to do the scheduling. Also see Turcan et al [132] and Chakraborty et al [17] for further refinements of this model.

Finally we present a model considered by Zacharias and Pindeo [150]. They assume that there are T regular appointment slots and $N \geq T$ patients requesting appointments. Thus $N - T$ is the level of overbooking. Customer j shows up with probability r_j. All patients who show up are punctual. The waiting cost of customer j is w_j per slot. Each customer requires exactly one time slot for service. The cost of service is a per slot in regular slots and b per slot in the over time slot. Let C_1 be the set of customers j with $r_j = 1$, and C_2 be the rest of the customers. Assume that $|C_1| < T$ so the overbooking model is non-trivial. Under these assumptions they derive the following rather obvious results: An optimal schedule

1. assigns the customers in C_1 to the first $|C_1|$ slots,

2. these slots have no overbooking,

3. it has no holes in it, that is there is no slot in the middle of the schedule with no patient scheduled for it.

Zacharias and Pinedo show that all schedules with the above three properties have the same service time cost. Thus we need to find a schedule that minimizes the waiting cost among all the schedules that satisfy the above properties. They show that such a schedule has the following structure: Assign $N - T + 1$ customers from C_2 to slot number C_1+1, to be served in the order of decreasing waiting costs. Order the remaining $N - |C_1| - 1$ customers in C_2 in increasing order of $w_j r_j/(1 - r_j)$ and assign them to slots $|C_1| + 1, \cdots, N$ in that order.

Note that which subset of customers is assigned to slot $|C_1|+1$ is left unspecified. This ambiguity can be resolved explicitly if all the waiting costs are the same, that is, $w_j = w$. In that case the subset that gets assigned to slot $|C_1| + 1$ is the customers in C_2 with the highest no show probabilities, the remaining ones are assigned in order of decreasing non-show probabilities.

Zacharias and Pinedo consider several extensions of the basic model.

Recommended for presentation: Zacharias and Pindeo citeZP14, Turcan et al [132], Chakraborty et al [17], LaGanga and Lawrence [81].

6.3 Direct Waiting: Steady State Models

In this section we study the steady- state queueing models of direct waiting. When the patients have appointments (like in many clinics) it is common to assume that arrivals occur at time nd, $n \geq 0$, where d is a fixed constant. The actual service times may be random. This implies that $D/M/1$ queue is a reasonable model of such systems. One can further modify these models to account for patient lateness, or no-shows. In other settings, such as emergency rooms, arrival process is random, and time varying. In such cases $M/G/1$ or $M(t)/G/\infty$ models may be appropriate. We shall study some of them here.

6.3.1 Winsten Model

Winsten [142] studies a $D/M/1$ queue as a model of a queue with scheduled arrivals of patients. He assumes that the inter-arrival times are exactly one unit. The service time are iid $\exp(\mu)$, with $\mu > 1$. Let $X(t)$ be the number of patients waiting in the clinic at time t, and $X_n = X(n-)$. He then computes

$$\pi_j = \lim_{n\to\infty} P(X_n = j), \quad j \geq 0. \tag{6.1}$$

as

$$\pi_j = (1 - \pi)\pi^j, \quad j \geq 0, \tag{6.2}$$

where π is the unique solution in $(0, 1)$ to

$$\pi = e^{-\mu(1-\pi)}. \tag{6.3}$$

This model assumes that all patients are punctual. Winsten then extends the model to the case where the nth patient, who is scheduled to arrive at time n,

actually arrives at time $n + \tau_n$, where the $\{\tau_n\}$ are iid random variables with support in $(0, 1)$, with cdf $F(\cdot)$. Thus the patients still arrive in the order of their appointments, but can be late. He then shows that

$$\begin{aligned}
\pi_0 &= 1 - \pi \frac{\int_0^1 e^{\mu x} dF(x)}{\int_0^1 e^{\mu \pi x} dF(x)}, \\
\pi_1 &= \pi(1-\pi) \frac{\int_0^1 e^{\mu x} dF(x)}{\int_0^1 e^{\mu \pi x} dF(x)}, \\
\pi_j &= \pi_1 \pi^{j-1}, \quad j \geq 2.
\end{aligned}$$

He connects this to the distribution at the actual time of arrival. He also extends the analysis to the case where $F(0) = 0$ and $F(2) = 1$, so some patients may arrive out of order. It is possible to construct and embedded DTMC for this model as well, but the analysis gets complicated.

Remark: I think this analysis also holds when $F(-1) = 0$ and $F(1) = 1$, that is, the patient may up to one unit of time early, or up to one unit of time late.

Mercer [94] and [95] extend the above analysis to the case of phase type service times, general lateness distribution, and positive probability that a patient may never show up.

6.3.2 Panel Size Model

Consider a single doctor clinic that serves a total of N patients (called the panel size). Each patient stays healthy for a random amount of time with mean $1/\lambda$ and then gets sick and visits the doctor. The sick patients queue up for service at the doctor, and after completing service return to the healthy patient pool. If N is large and λ is sufficiently small, we can approximate the overall arrival process to clinic as a PP(λN). Suppose all service times are of fixed length T. Then the clinic queue can be modeled as a simple $M|D|1$ queue. Using the standard theory of $M|G|1$ queue, we see that the queue is stable if

$$\rho = \lambda N T < 1.$$

The expected queueing time in steady state in the system is given by

$$\frac{\lambda N T^2}{2(1-\rho)} = \frac{\rho T}{2(1-\rho)}.$$

Now suppose we want to ensure that a patient's queueing time should not exceed cT, for a fixed $c > 0$. Thus we must have

$$\frac{\rho}{2(1-\rho)} \leq c,$$

or

$$\rho \leq 2c/(1+2c).$$

Thus the panel size should be bounded above by

$$N \leq \frac{2c}{(1+2c)\lambda T}.$$

Thus if $\lambda = 1$ visit per year, and $T = 20$ minutes (=1/(3*250*8) years), using $c = 1$, we get

$$N = (2/3) * (2 * 250 * 8) = 2667.$$

Green and Savin [51] extend this simple model assuming that the clinic is an $M|D|1|K$ queue, and that the patients exhibit a now show behavior that is dependent on the time spent in the clinic. Suppose the patient that sees k persons ahead of him will actually be a no show with probability $\gamma(k)$. It is intractable to model this no-show behavior precisely. Hence they assume that a patient whose service is about to start is no-show with probability $\gamma(k)$ if there are k patients waiting behind him when he is about to start service. They assume that the service slot of a no show patient is wasted. At the end of this wasted service slot the no-show patient rejoins at the end of the queue with probability r, or is lost with probability $1 - r$. Using more sophisticated analysis, they recommend a panel size of about 2500.

6.4 Joint Model of Indirect and Direct Wait

This section is based on Luo et al [86] where they consider a tandem queue model of an appointment system. The customers first call an appointment nurse who gives them the first available appointment. They wait (virtually) in the appointment queue. At their appointment time they leave the appointment queue. They arrive at the clinic with a given probability, or become a no-show. We first discuss the appointment queue in the next subsection, and the tandem queue in the following one.

6.4.1 $M/D/1$-type Model of an Appointment Queue.

We assume that appointment requests arrive according to a Poisson process with rate λ and appointments can only be scheduled at equidistant time epochs with distance d, which is controlled by the system designer. For instance, suppose the service provider decides to schedule appointments every half an hour, such as 8:00, 8:30, 9:00, etc, and a customer calls at 10:20 to request an appointment, then he will be scheduled at 10:30 if there is no scheduled appointment in the system, or 30 minutes later than the last appointment time that has already been assigned. When a customer's appointment time is due, that customer is removed from the appointment list.

Let $X(t)$ be the number of customers waiting in the appointment queue at time t. Define

$$X_n = X(nd^-),\ n = 0, 1, 2, \ldots.$$

One can show that $\{X_n, n \geq 0\}$ is a DTMC with an $M/G/1$ type transition probability matrix:

$$P(X_{n+1} = j | X_n = i) = T_{i,j} = a_{j-i+1},$$

where

$$a_k = e^{-\lambda d} \frac{(\lambda d)^k}{k!}, \quad k \geq 0.$$

Define (assuming the limits exist)

$$\pi_i^A = \lim_{n \to \infty} P(X_n = i), \ i = 0, 1, 2, \ldots$$

and

$$p_i^A = \lim_{t \to \infty} P(X(t) = i), \ i = 0, 1, 2, \ldots.$$

Also define the generating function

$$\phi(z) = \sum_{j=0}^{\infty} \pi_j^A z^j.$$

Define

$$\rho = \lambda d,$$

and

$$\psi(z) = e^{\rho(z-1)}. \tag{6.4}$$

The appointment queue is stable if $\rho < 1$. Assuming stability, we have

$$\phi(z) = \sum_{j=0}^{\infty} \pi_j^A z^j = (1 - \rho) \frac{(1 - z)\psi(z)}{\psi(z) - z}.$$

Furthermore,

$$p_i^A = \frac{\pi_{i+1}^A}{\rho}, \ i = 0, 1, 2, \ldots. \tag{6.5}$$

The limiting expected number of appointments is given by

$$L_A = \lim_{t \to \infty} E(X(t)) = \frac{\rho}{2(1 - \rho)},$$

and the limiting expected indirect delay is given by

$$W_A = \frac{d}{2(1 - \rho)}. \tag{6.6}$$

6.4.2 Tandem Queue Model

The customer departing from the appointment queue enters the clinic queue with probability α^k if there are k patients in the appointment queue when this patient leaves the appointment queue. (This idea was first suggested by Green and Savin [51].) One can show that, in steady state, the probability that a departure from the appointment queue joins the clinic queue is

$$(1-\rho)\frac{1-e^{\rho(\alpha-1)}}{e^{\rho(\alpha-1)-\alpha}}.$$

Note that the no-show events of successive patients are not independent, which makes the analysis of clinic queue mush harder.

Let $Y(t)$ be the number of customers in the clinic queue at time t. We model the clinic queue as a single server system with iid exp(μ) service times. Let $Y_n = Y(nd-)$. Let X_n be as in the previous subsection. Let

$$b_k = e^{-\mu d}\frac{(\mu d)^k}{k!}, \;\; k \geq 0.$$

Then one can show that $\{(X_n, Y_n), n \geq 0\}$ is a DTMC with transition probabilities given by

$$P(Y_{n+1} = m, X_{n+1} = k | Y_n = j, X_n = i)$$

$$= \begin{cases} \alpha^{i-1}a_{k-i+1}b_{j-m+1} + (1-\alpha^{i-1})a_{k-i+1}b_{j-m} & i > 0 \\ a_k b_{j-m} & i = 0. \end{cases}$$

This DTMC is positive recurrent if $\lambda < \mu$ and $\rho = \lambda d < 1$. Let $Q = \text{diag}(0, \alpha^0, \alpha^1, \alpha^2, \cdots)$, $M = (I-Q)T$, $N = QT$, where T is from the previous subsection,

$$A_0 = b_0 N, \; A_k = b_{k-1}M + b_k N, \; k \geq 1,$$

$$B_k = \sum_{i=k}^{\infty} A_{i+1}.$$

The DTMC has a bock-$G|M|1$ structure, that is, its transition probability matrix is given by

$$P = \begin{bmatrix} B_0 & A_0 & & \\ B_1 & A_1 & A_0 & \\ B_2 & A_2 & A_1 & A_0 \\ \vdots & \cdots & \vdots & \ddots \end{bmatrix}.$$

One can (in theory) compute the limiting distribution

$$\pi_{k,j} = \lim_{n\to\infty} P(Y_n = k, X_n = j)$$

using this structure as follows. Let R be a solution to

$$R = Re^{\mu d(R-I)}M + e^{\mu d(R-I)}N$$

and the vector η be a solution to

$$\eta = \eta(I - R)^{-1}(T - R), \;\; \eta(I - R)^{-1}e = 1.$$

Then the limiting distribution

$$\pi_k = [\pi_{k,0}, \pi_{k,1}, \cdots]$$

is given by

$$\pi = \eta R^k, \;\; k \geq 0.$$

This produces an efficient method of computing the distribution if the state space of the appointment queue is finite, say K. Then the block matrices are $K + 1$ by $K + 1$ and the R matrix can be computed by simple iteration. The required performance measures can be computed using this distribution. For example, the steady state expected direct waiting time is

$$W = \frac{\eta}{\lambda\mu d}(I - R)^{-2}Qe,$$

and the steady state idle time in the clinic queue is

$$p_0 = \frac{\eta}{\lambda\mu d}(I - R)^{-1}Qe.$$

One can use these to find an optimal d that minimizes the relevant costs. We refer the readers to Luo et al [86] for further results. They also suggest several approximations to compute various performance measures.

6.5 Dynamic Appointment Scheduling

In this section we consider dynamic scheduling of appointment for patients, where the appointment given to a new patient depends on the current appointment schedule.

6.5.1 Markov Decision Processes and Index Policies

Here we collect the background information on Markov Decision Processes (MDP) and the index policies.

MDP is a tool to study sequential decision making problem. [108] is an excellent reference on this subject. An MDP has five elements: the decision epochs, the state space, the action space, the transition probabilities and the costs. The decision epoch is the point of time when a decision is made. Let X_n be the system state at the decision epoch n. Suppose $X_n \in S$ for all $n \geq 0$. We call S the state space. Let A_n be the action taken at the decision epoch n. Suppose $A_n \in A$ for all $n \geq 0$. We call A the action space. The process $\{(X_n, A_n), n \geq 0\}$ is called an MDP if

$$P(X_{n+1} = j | X_n = i, A_n = a, X_{n-1}, ..., X_1, X_0, A_{n-1}, ..., A_1, A_0) = p_{ij}(a),$$

for all $n \geq 0, i, j \in S, a \in A$. We call $p_{ij}(a)$ the transition probabilities. Let $c(i, a)$ be the expected cost incurred if action a is chosen in state i at any time $n \geq 0$.

A policy is a description of actions taken at each decision epoch. Let $\pi : S \to A$ be a policy. We choose action $\pi(i)$ whenever the system is in state i under policy π. Such policies are called stationary deterministic Markovian (SDM) policies. For a given SDM policy π, define (assuming the limit exists)

$$g^{\pi}(i) = \lim_{N\to\infty} \frac{1}{N+1} \mathrm{E}^{\pi} \left[\sum_{n=0}^{N} c(X_n, A_n) \middle| X_0 = i \right].$$

We call $g^{\pi}(i)$ the long-run average cost of following the policy π. Let

$$g^*(i) = \inf_{\pi} g^{\pi}(i), \quad \forall i \in S,$$

where the infimum is taken over all SDM policies. If there is a policy π^* that achieves this infimum, it is called the (average-cost) optimal policy. Thus an optimal policy (if it exists) satisfies

$$g^{\pi^*}(i) = g^*(i), \quad \forall i \in S.$$

Now we discuss when such an optimal policy exists and how to compute it. Define

$$v_{n+1}(i) = \min_{a\in A}\{c(i, a) + \sum_{j\in S} p_{ij}(a) v_n(j)\},$$

for all $i \in S$ and $n \geq 0$, where $v_0(i) = 0$ all $i \in S$. We can interpret $v_n(i)$ as the optimal total expected cost incurred over the n days starting from state i. It is known (see [131]) that $v_n(i)$ is asymptotically linear in n with slope g and intercept $h(i)$. We can write

$$v_n(i) = ng + h(i) + o(n),$$

where $\frac{o(n)}{n} \to 0$ as $n \to \infty$. The slope g is the optimal long-run average cost. The intercept $h(\cdot)$ is called the bias function. It is known (see [131]) that g and $h(\cdot)$ satisfy the Bellman equation

$$h(i) + g = \min_{a\in A}\{c(i, a) + \sum_{j\in S} p_{ij}(a) h(j)\},$$

It is also known (see [131]) that if the above equation has a solution, then we can use it to compute the optimal decision as follows. Define

$$a(i) = \arg\min_{a\in A}\{c(i, a) + \sum_{j\in S} p_{ij}(a) h(j)\}, \quad \forall i \in S.$$

The standard theory of dynamic programming shows that the Markovian policy that chooses action $a(i)$ in state i is optimal. It is known (see [131]) that the

optimality equation has a solution if the MDP is unichain, that is, for each stationary policy the associated Markov chain has no two disjoint closed sets.

If the optimality equation has a solution, we can solve it by an iterative method. We restate Theorem 6.6.1 of [131] in the theorem below, which allows us to use the iterative method to compute g and $h(\cdot)$.

Theorem 6.1 *For any state i, we have*

$$h(i) - h(0) = \lim_{n\to\infty}[v_n(i) - v_n(0)],$$

and

$$g = \lim_{n\to\infty} \frac{v_n(i)}{n}.$$

Furthermore,

$$\min_{i\in S}\{v_n(i) - v_{n-1}(i)\} \le g \le \max_{i\in S}\{v_n(i) - v_{n-1}(i)\}$$

However, solving the optimality equation is intractable when the state space and action space are large. Hence, we develop heuristic policies which perform well. One such policy is called the index policy, which we define below.

Suppose the MDP is unichain. Let π be a given initial policy. Then there exists a constant g^π and a bias function h^π that satisfy

$$h^\pi(i) + g^\pi = \min_{a\in A}\{c(i,a) + \sum_{j\in S} p_{ij}(a)h^\pi(j)\}.$$

Now consider a policy $\hat{\pi}$ that chooses the action $\hat{a}(i)$ in state i, where

$$\hat{a}(i) \in \arg\min_a\{c(i,a) + \sum_{j\in S} p_{ij}(a)h^\pi(j)\}.$$

The next theorem shows the importance of this construction.

Theorem 6.2

$$g^{\hat{\pi}} \le g^\pi.$$

and if $g^{\hat{\pi}} = g^\pi$, then $\hat{\pi}$ is the average-cost optimal policy.

Suppose we can construct a function $f : S \times A \to R$ such that

$$\arg\min_a f(i,a) \subset \arg\min_a\{c(i,a) + \sum_{j\in S} p_{ij}(a)h^\pi(j)\}.$$

The function f is called an index function and the policy $\hat{\pi}$ is called an index policy using the index function f. It has been observed that the index policy $\hat{\pi}$ provides a tractable heuristic policy, especially if the initial policy π is chosen wisely so that the f function is easy to evaluate.

We apply this methodology to appointment scheduling.

6.5.2 Patient No-shows and Cancellations

We describe a discrete time model considered by Liu et al [84]. At the beginning of the t-th day, a random number A^t patients call in and request appointments over days $t, t+1, \cdots, t+T$, where T is the scheduling horizon. We assume that $\{A^t, t \geq 0\}$ is a sequence of iid random variables. For $1 \leq i \leq T$ and $0 \leq j \leq T-i$ let $X^t_{i,j}$ be the number of patients who called on day $t-i$ and are given appointment $t+j$, and who have not cancelled their appointment so far. The state of the schedule on day t is given by vector X^t of these elements. For example, for $T = 3$

$$X^t = [X^t_{1,0}, X^t_{1,1}, X^t_{1,2}, X^t_{2,0}, X^t_{2,1}, X^t_{3,0}].$$

Note that the total number of appointments for today is given by $\sum_{i=1}^{T} X^t_{i,0}$. The state of the system at the beginning of day n is (X^t, A^t). Based on this state we have decide how to schedule the appointment days of the A^t new requests. Let Y^t_j represent the number of patients who make their requests on day t and are given appointments for day $t+j$, $0 \leq j \leq T$. Clearly, we must have

$$A^t = \sum_{j=0}^{T} Y^t_j.$$

Thus the decision on day t is the vector $Y^t = [Y^t_0, Y^t_1, \cdots, Y^t_T]$ satisfying the above constraint. Now suppose that a patient who arrived on day $t-i$ does not cancel his appointment by day t with probability β_i and actually keeps his appointment with probability α_i if his appointment is on day t. Liu et al show how these probabilities can be estimated from data. Now, given the system state (X^t, A^t) and action Y^t, we have

$$X^{t+1}_{ij} = \begin{cases} B(Y^t_{j+1}, \beta_0) & \text{for } i = 1, 0 \leq j \leq T-1, \\ B(X^t_{i-1,j+1}, \beta_{i-1}) & \text{for } 2 \leq i \leq T, 0 \leq j \leq T-i. \end{cases} \tag{6.7}$$

Here $B(n,p)$ stands for a Binomial random variable with parameters n and p. Now, the number of patients scheduled to arrive on day t is given by

$$S^t = \sum_{i=1}^{T} X^t_{i,0} + Y^t_0,$$

and the actual number of patients who actually keep their appointments on day t is given by

$$U^t = \sum_{i=1}^{T} B(X^t_{i,0}, \alpha_i) + B(Y^t_0, \alpha_0).$$

Next we describe the revenue function. Let $r(x,z)$ be the net revenue if the z patients were scheduled on day t, but only x of them actually turn up (due to cancellation and no-shows). Thus the revenue at time t is $r(S^t, U^t)$ which is a function of $((X^t, A^t), Y^t)$. This shows that $\{((X^t, A^t), Y^t), t \geq 0\}$ is an MDP.

We want to derive a policy that maximizes the long run average reward. However, the state-space is too large to to do this exactly. Hence Liu et al develop an index policy. Here we explain the simplest one obtained by improving the open access policy (OAP). That is, suppose the initial policy is the open access policy that schedules every new request at the beginning of day t on day t itself. Under this policy, starting from day state (x, a) on day one, we have

$$S^0 = \sum_{i=1}^{T} x_{i,0} + a, \;\; U^0 = \sum_{i=1}^{T} Bin(x_{i,0}\alpha_i) + Bin(a, \alpha_0),$$

$$S^i = x_{i,0} + A^i, \;\; U^i = Bin(x_{i,0}, \beta_i) + Bin(A^i, \alpha_0), \;\; 1 \leq i \leq T,$$

and

$$S^t = A^t, \;\; U^t = Bin(A^t, \alpha_0), \;\; t > T.$$

Hence the long run average reward per day is given by

$$g^{OAP} = E(r(A^t, Bin(A^t, \alpha_0))),$$

and the bias function is

$$h^{OAP}(x, a) = \sum_{i=0}^{T} r(S^i, U^i). \tag{6.8}$$

Now the policy improvement step yields

$$g+h(x,a) = \max_y \{r(\sum_{i=1}^{T} x_{i,0}+y_0, \sum_{i=1}^{T} Bin(x_{i,0}\alpha_i)+Bin(y_0,\alpha_0))+E(h^{OAP}(X', A'))\},$$

X' is the state on the next day after decision y is taken on day 1, and A' is the new random number of arrivals on day 2 (see Eq. 6.7.) That is,

$$X'_{ij} = \begin{cases} B(y_{j+1}, \beta_0) & \text{for } i = 1, 0 \leq j \leq T-1, \\ B(x_{i-1,j+1}, \beta_{i-1}) & \text{for } 2 \leq i \leq T, 0 \leq j \leq T-i. \end{cases} \tag{6.9}$$

The maximization is done over all vectors y such that

$$\sum_{j=0}^{T} y_j = a.$$

Using Eq. 6.8, we see that this is a separable maximization problem as follows:

$$\text{Maximize} \;\; \sum_{j=0}^{T} f_j(y_j, x)$$

$$\text{Subject To:} \;\; \sum_{j=0}^{T} y_j = a.$$

(The f_j do depend on x. See Liu et al for detailed expression for the f_j's.) If we further assume that $r(\cdot,\cdot)$ is jointly concave, we can show that the f_j's are concave functions. Hence this problem can be solved in a greedy fashion. This yields the decision y in state (x,a). This policy is guaranteed to be better than the OAP.

It can also be used to dynamically schedule patients as follows. Suppose the state of the schedule at the beginning of the day is x. Set the initial vector $y = [0, 0, \cdots, 0]$. Let

$$j^* \in \text{argmax } \{f_j(y_j + 1, x) - f_j(y_j, x),\ 0 \le j \le T\}$$

and assign the incoming patient to day j^*, and update $y_{j^*} = y_{j^*} + 1$. Repeat this for all arrivals. Thus we can think of

$$I_j(y, x) = f_j(y + 1, x) - f_j(y, x)$$

as the index for the state x when y new patients have already arrived for appointments at the beginning of the current day.

We present some special cases below:

Linear Costs and Rewards: Suppose

$$r(x, z) = \tau x - \eta z$$

where $\tau > 0$ is the revenue from serving one patient, and $\eta > 0$ is cost per scheduled appointment. In this case the index functions simplify to

$$I_j(y, x) = \tau\alpha_j - \eta\beta_j$$

which are independent of both y and x. Thus an incoming patient is assigned to day j^* where

$$j^* \in \text{argmax}\{\tau\alpha_j - \eta\beta_j : 0 \le j \le T\}.$$

This is the improved policy corresponding to OAP in the case of linear costs and rewards. In fact it can be shown to be the optimal policy.

Linear Rewards and Quadratic Costs: Suppose

$$r(x, z) = \tau x - \eta z^2.$$

In this case the index functions simplify to

$$I_j(y, x) = \begin{cases} \tau\alpha_0 - \nu\beta_0(1 + 2\beta_0 y_0 + 2\sum_{i=1}^{T} x_{i0}\alpha_i) & \text{if } j = 0 \\ \tau\alpha_j - \nu\beta_j(1 + 2\beta_j y_j + 2(\sum_{i=1}^{T} x_{ij}\alpha_{i+j} + \mu)) & \text{if } 1 \le j \le T. \end{cases}$$

Liu et al show that index policy performs quite well and compare its performance to several other heuristic policies. It is an easy policy to implement, and uses all the parameters of the problem.

6.5.3 Patient Preferences

We now consider a general version of the discrete time model studied by Zhang and Kulkarni [156]. The appointment horizon is T. There are no same day appointments. A new request for appointment arriving on day n may be rejected or accommodated on one of the days $n+j$, $1 \leq j \leq T$. The patients belong to K classes, and the class k patient preference is given by a subset $\mathcal{A}_k \subseteq \{1, 2, \cdots, T\}$. A patient of type k arriving on day n is willing to accept an appointment on any day $n + j$ as long as $j \in \mathcal{A}_k$. (He is indifferent among these appointments.)

Let A_n be the number of arrival on day n. Assume that $\{A_n, n \geq 1\}$ is a sequence of iid G(α) random variables, that is

$$P(A_n = i) = \alpha^i(1 - \alpha), \;\; i \geq 0.$$

Let p_k be the probability that an arriving patient belongs to type k, $1 \leq k \leq K$. Clearly $p_k > 0$ and $\sum p_k = 1$. We observe the system after every arrival and every change of day. The new event is an arrival with probability α and a day-change with probability $1 - \alpha$. The state of the system is given by a vector $x = (x_1, x_2, \cdots, x_T)$, where x_j is the number of appointments currently on the book for day j from today. If we observe the system after an arrival, and we schedule this patient j days in to the future, the state increases in the j-th coordinate by 1. If we reject this patient, the state remains unchanged. If it is a change of day event, we have x_1 customers scheduled for service on the new day, but each of them keeps the appointment with probability θ, so the actual number of patients who arrive for service is Bin(x_1, θ). This costs $c(x_1)$ dollars. The new state is given by $x' = (x_2, x_3, \cdots, x_T, 0)$.

Now let $v_n(x)$ be the minimum total expected cost incurred over the first n events. Then we have

$$v_0(x) = 0,$$

$$v_{n+1}(x) = \alpha \sum_{k=1}^{K} p_k \min\{\min_{j \in \mathcal{A}_k} \{v_n(x + e_j)\}, v_n(x)\} + (1 - \alpha)(c(x_1) + v_n(x')).$$

The dynamic programming equation for the average cost case is given by

$$g + h(x) = \alpha \sum_{k=1}^{K} p_k \min\{\min_{j \in \mathcal{A}_k} \{h(x + e_j)\}, h(x)\} + (1 - \alpha)(c(x_1) + h(x')).$$

One can show that this MDP is unichain and the average cost optimal policy exists. The bias function h and the optimal long run average cost can be computed as described earlier.

The optimal policy can be then derived from the h function. Now suppose $c(\cdot)$ is a convex function that achieves its minimum at m. The authors show that the optimal policy does not schedule more that m patients on any day. If a patient of type k arrives and all days in $\mathcal{A}_k$ have m patients scheduled on them, then it is optimal to reject this patient. The actual scheduling policy is rather complicated and hard to compute and implement because of the huge state space. Hence they construct an index policy as follows.

Suppose the initial policy assigns an incoming patient of type k to any day in $\mathcal{A}_k$ with equal probability. Call this policy UAP (Uniform Assignment Policy). Compute its bias function $h^{UAP}(x)$. This is a bit messy but doable. Next we improve this policy. When a type k patient arrives we reject him if every day in $\mathcal{A}_k$ already has m patients scheduled on it. Else, we solve the optimization problem

$$\min_{j \in \mathcal{A}_k : x_j < m} \{h^{UOP}(x + e_j)\}$$

and choose the decision that minimizes the function. Thus we can think of $h(x+e_j) - h(x)$ as the index for day j, and assign the patient to that day in $\mathcal{A}_k$ which has less than m patients assigned to it and has the smallest index. Zhang and Kulkarni [156] give expression for this index for the case $T = 2$. Here we give the general expression: Let $a_k = |\mathcal{A}_k|$. Let A_t be the number of patients that are assigned to day t in one day under this policy. Then we can show that

$$A_t \sim G\left(\frac{\alpha\rho_t}{1 - \alpha(1 - \rho_t)}\right), \quad t = 1, ..., T$$

where

$$\rho_t = \sum_{k : t \in \mathcal{A}_k} p_k / a_k.$$

Define

$$B_k = \sum_{t=1}^{k} A_t,$$

where the A's on the right hand side are independent random variables. Then the index simplifies to

$$I_k(y) = E[c(y + 1 + B_k)] - E[c(y + B_k)], 1 \leq k \leq T.$$

The index policy in state $x = (x_1, x_2, \cdots, x_T)$ on day n when a patient of type k arrives is to assign her to day $n + k^*$ where

$$k^* = \text{argmin}_{t \in \mathcal{A}_k : x_t < m} \{I_k(x_t)\}.$$

If $x_t = m$ for all $t \in \mathcal{A}_k$, we reject the patient.

Recommended for Presentation: Balasubramanian et al. [9], Gupta and Wang [56], Wang and Gupta [138], and Feldman et al. [38].

6.5.4 Series Patients

Here we consider a discrete-time model of scheduling patients who need more than one visit to the clinic. Such patients are called series patients. Examples of the series patients occur in physical therapy, psychotherapy, chemotherapy, etc. A series patient is described by three attributes: number of visits needed, frequency of visits, and the duration of the visits. Thus a physical therapy patient may need 10 weekly visits of one hour each. The material presented

here is based on Yu et al [148]. They consider both staffing and scheduling. Here we concentrate only on the scheduling aspect.

Let B_n be the number of new arrivals on day n. We assume that the i-th patient needs a total V_i hour-long visits, one per T days. $\{B_n, n \geq 1\}$ and $\{V_i, i \geq 1\}$ are two independent sequences of iid non-negative integer valued random variables. When the i-th patient calls and requests k visits, we need to schedule her first visit on day $m \in \{n+1, n+2, \cdots, n+T\}$. The rest of the visits are scheduled on days $m+T, m+2T, \cdots, m+(k-1)T$. Thus there are no same-day appointment, no walk-ins. They also assume no cancellations or no-shows. The rule tells us how to schedule the first appointment is called the scheduling policy.

Suppose there are q slots available on each day. This is called the staffing level. We assume that this is given and fixed. Let $\phi(d)$ be the cost of serving d patients on a given day. We assume that it is an increasing concave function of d. Also, $c_d(t)$ is the cost of scheduling the first appointment t after the arrival day of the patient, and assume that it increases with t. $(1 \leq t \leq T)$. This creates an incentive to accommodate a new patient as quickly as possible in the schedule.

The state of the appointment schedule at the beginning of day $n \geq 0$ is described by a matrix $X_n = [X_n(k,t)]$, where $X_n(k,t)$ is the number of patients who have appointments (not necessarily their first) on day $n+t$, and have *at least* k visits remaining.

Let $A_n = [A_{n,k}]$ be the arrival vector at the beginning of day n, where $A_{n,k}$ is the number of new arrivals on day n that need at least k visits to the clinic. Note that in practice, the scheduling decisions are made for each new patient at the time of her arrival, depending on the state of the appointment schedule at that time. This, unfortunately, leads to more complicated models. The state of the system at the beginning of day n is (X_n, A_n). We need to decide $Y_{n,k}$, the number of $A_{n,k}$ new patients to be given their first appointments on day $n+t$, for $1 \leq t \leq T$. The dynamics of the system can be represented as

$$X_{n+1} = T(X_n + Y_n),$$

for an operator T. We see that $\{((X_n, A_n), Y_n), n \geq 0\}$ is an MDP.

The total number of patients who are scheduled for service on day n is given by

$$D_n = X_{n-1}(1,1) + Y_{n-1}(1,1).$$

Hence the cost incurred from on day n in state $(X_n, A_n) = (x, a)$ under decision $Y_{=}y$ is given by

$$r(x,y) = \phi(x(1,1) + y(1,1)) - \sum_{t=1}^{T} c_d(t) y(1,t).$$

The MDP equation can now be written as

$$v(x,a) + g = \max_y \{r(x,y) + E(v(T(x+y), A))\}.$$

The authors show that this MDP is unichain and the avg cost optimal policy exists. It is clear that it is not possible to actually compute the optimal policy for this problem, since the state space is just too large. Hence the authors construct an index policy by improving upon an initial policy that assigns every new patient to the next day ($t = 1$). We give the main result below in a special case when the number of new arrivals is a P(λ) random variable. The cost function $\phi(d)$ is derived by assuming that there are q regular slots, and each regular slot costs c dollars, whether it is filled or not, each filled slot above q (the overtime slots) costs c', each filled slot up to $(1+\gamma)q$ produces a revenue of p, but any slot after that produces a revenue of c' (so net revenue is zero). Let $f_j = P(V_i \geq j)$ and define

$$\lambda_{k,t} = \begin{cases} 0 & \text{if } k = 1,\ t = 1; \\ \lambda \sum_{j=1}^{k-1} f_j, & \text{if } k \geq 2,\ t = 1; \\ \lambda \sum_{j=1}^{k} f_j, & \text{if } k \geq 1,\ t \geq 2; \end{cases}$$

Let $F_{j,t}(\cdot)$ be the cdf of a P($\lambda_{j,t}$) random variable. Define the index functions I_k as follows

$$I_k(x,t) = \sum_{j=1}^{k} (p - c')F_{j,t}(\gamma q - x(j,t) - 1) + c'F_{j,t}(q - x(j,t) - 1) - c_d(t)$$

The authors show that the policy obtained by using one step of the policy improvement step starting with the next-day policy yields the following decision rule: when a new patient needing k visits arrives on day n, we first observe the state of the schedule x, and compute

$$t^* \in \text{argmin}\{I_k(x,t) : 1 \leq t \leq T\},$$

and schedule the first appointment of this patient on day $n + t^*$ and update the schedule accordingly.

The authors consider many other heuristic policies and compare their performances. They find that the above index policy performs better than them in most cases. They also study the joint problem of staffing and scheduling. We refer the reader to their paper for further details.

6.5.5 Walk-in Patients

The material in this section is based on Bolia and Kulkarni [12]. They model an MRI-clinic with N appointment slots of equal length during each working day. At the beginning of the day some of these slots are filled with appointments, and some are empty. If the nth slot is filled, there is a probability α_n that the patient will fail to show up for the appointment. With probability $1 - \alpha_n$, the patient shows up on time. Let W_n be the number of walk in patients during the nth time slot. If the nth slot is not filled by an appointment, we set $\alpha_n = 1$. Assume that $\{W_n, 1 \leq n \leq N\}$ are independent random variables with pmf

$$w_n(j) = P(W_n = j), \quad j \geq 0.$$

The scheduled patients have priority over the walk-in patients, and all of them need exactly one slot for service. Every admitted patient must be served during the day, using over-time slots if needed.

Each walk-in patient produces a net revenue of $r > 0$ if served in one of the N slots, and a net revenue of $r' < 0$ if served in an overtime slot. At the beginning of the nth time slot (after the service in that slot, if any, has started) we know X_n, the number of walk-ins already waiting for service. Note that any customer, walk-in or otherwise, who starts service in the nth slot is not counted in X_n. At this time we have to decide A_n, the number of walk-ins to admit during the nth time slot, so that $X_n + \min(A_n, W_n)$ will be available for service for the $(n+1)$st time slot. If no scheduled patient turns up for the $(n+1)$st time slot, and $X_n + \min(A_n, W_n) \geq 1$, one of these walk-ins will be served, and will produce a revenue of r. It is clear that $A_N = 0$ is optimal, that is, it is optimal not admit any walk-ins during the last slot, since they have to be served in overtime and hence only produce negative revenue. Hence, the number of walk-ins left unserved at the end of slot N is $X_{N+1} = X_N$, and each one of them produces a net revenue of r'. No more walk-ins are admitted during the overtime period. There is no explicit cost of rejecting a walk-in customer, other than the potential loss of revenue. The aim is to derive an optimal admission policy that maximizes the expected revenue from the walk-in customers. The expected revenue from the customers with appointments is $\sum_{n=1}^{N}(1-\alpha_n)r$, and is unaffected by the admission policy for the walk-in customers. We do so by formulating the problem as a Markov decision process $\{(X_n, A_n), 1 \leq n \leq N\}$ as follows.

We have

$$X_{n+1} = \begin{cases} X_n + \min(A_n, W_n) & \text{scheduled customer shows up in slot } n+1 \\ X_n + \{\min(A_n, W_n) - 1\}^+ & \text{scheduled cust fails to show up in slot } n+1. \end{cases}$$

Using this we can compute

$$p_{ij}^{(n)}(a) = P(X_{n+1} = j | X_n = i, A_n = a), \quad i, j, a \geq 0, \;\; 1 \leq n < N.$$

We aim to maximize the expected net revenue from the walk-in patients. Let $r_n(i, a)$ be the one step expected reward earned from the walk-in customers in slot i if there are i patients waiting at the beginning of slot n, and we decide to admit a more walk-in during the nth slot. We have

$$r_N(i, a) = r'i, \;\; i \geq 0, a \geq 0,$$

and, for $1 \leq n < N$

$$r_n(0, 0) = 0,$$
$$r_n(0, a) = \alpha_{n+1}(1 - w_n(0))r, \;\; a > 0$$
$$r_n(i, a) = \alpha_{n+1}r, \;\; i > 0, a \geq 0.$$

Let $V_n(j)$ be the expected total net revenue from the walk-in customers over time slots $\{n+1, n+2, \cdots, N\}$, if there are j walk-ins waiting at the beginning

of the nth time slot. Then the optimality equations are given by

$$\begin{aligned} V_N(j) &= r'j, \\ V_n(j) &= r\alpha_{n+1} + \max_{a\geq 0}\{\alpha_{n+1}E(V_{n+1}(j+\min(W_n,a)-1)) \\ &\quad +(1-\alpha_{n+1})E(V_{n+1}(j+\min(W_n,a)))\}, \;\; j>0, 1\leq n\leq N-1, \\ V_n(0) &= \max_{a\geq 0}\{r\alpha_{n+1}(1-w_n(0))1_{\{a>0\}} + \alpha_{n+1}E(V_{n+1}((\min(W_n,a)-1))^+) \\ &\quad +(1-\alpha_{n+1})E(V_{n+1}(\min(W_n,a)))\}. \end{aligned}$$

Let $a_n^*(j)$ be the value of a that maximizes the right hand side of (6.10) for $j>0$ and (6.10) for $j=0$. Then, the optimal policy is to admit as many as $a_n^*(j)$ walk-in customers to the clinic during the nth slot. Using this model they show that there exist critical numbers b_n^* and c_n* such that is is optimal to admit upto b_n^* walk-in patients in slot n if $X_n>0$ and up to c_n^* if $X_n=0$. They also show that

$$b_1^* \geq b_2^* \geq \cdots \geq b_{N-1}^* = 0,$$

$$c_{N-1}^* = \begin{cases} 0 & \text{if } K(1-\alpha_N)\geq r \\ 1 & \text{if } K(1-\alpha_N)<r \end{cases}$$

and

$$b_n^*>0 \rightarrow c_n^* = b_n^*.$$

Computing these critical numbers involves a lot of work, and hence they develop a heuristic policy that is simpler to compute and implement. The main result is as follows: Let B_m be a Bernoulli rv with parameter α_m, and G_n be the cdf of $B_{n+1}+\cdots+B_N$. Define

$$a_n^* = G_n^{-1}(r/(r-r')).$$

The heuristic policy admits up to a_n^* walk-in patients in slot n. They show numerically that this policy works quite well.

Bibliography

[1] Alagoz, O., L. M. Maillart, A. J. Schaefer and M. S. Roberts, (2004). The Optimal Timing of Living-Donor Liver Transplantation. *Management Science*, 50, 1420-1430.

[2] Ahn, J.H. and J.C. Hornberger (1996). Involving patients in the cadaveric kidney transplant allocation process: A decision-theoretic perspective. *Management Science*, 42, 629-641.

[3] Albright, S. C. (1974). Optimal Sequential Assignments with Random Arrival Times. *Management Science*, 21, 60-67.

[4] Albright, S.C. and C. Derman (1972). Asymptotic optimal policies for the stochastic sequential assignment problem. *Management Science*, 19, 46-51.

[5] Armitage P., Doll R. (1954). The age distribution of cancer and a multistage theory of carcinogenesis. *British Journal of Cancer*, 8, 112.

[6] Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin and G. B. Yom-Tov (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stoch. Sys.*, 5. 1-49.

[7] Bailey, N. T. J. (1952). A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical So- ciety*, Series B (Methodological), 185-199.

[8] Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications.* London: Griffin.

[9] Balasubramanian, H., S. Biehl, L. Dai, and A. Muriel (2014). Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments, *Health care management science*, 17,31-48.

[10] Ball, F. and Lyne, O. (2002). Optimal vaccination policies for stochastic epidemics among a population of households. *Mathematical Biosciences*, 177, 333354

[11] Behncke, H. (2000). Optimal control of deterministic epidemics. *Optimal Control Applications and Methods,* 21, 269-285.

[12] Bolia, N. and V. G. Kulkarni (2015). Admission Policies for Walk-in Patients in Diagnostic Clinics, Working Paper.

[13] Boxma, O. J., Israel, D., Perry, D. and Stadje W. (2011). A new look at organ transplantation models and double matching queues, *Probability in the Engineering and Informational Sciences*, 25, 135-155.

[14] Brahimi, M., D.J Worthington, (1991). Queueing models for out-patient appointment systemsa case study, *Journal of the Operational Research Society*, 42, 733746.

[15] Brandeau, M. L., F. Sainfort and W. P. Pierskalla, (2004). *Operations Research and Health Care: A Handbook of Methods and Applications*, Kluwer International Series.

[16] Brodheim, E., C. Derman and G. Prastacos, (1975). On the Evaluation of a Class of Inventory Policies for Perishable Products Such As Blood. *Management Science,* 21, Theory Series, 1320-1325.

[17] Chakraborty, S., K. Muthuraman, and M. Lawley (2013). Sequential clinical scheduling with patient no-show: The impact of pre-defined slot structures. *Socio-Economic Planning Sciences,* 47, 205-219.

[18] Chazan, D. and S. Gal, (1977). A Markovian Model for a Perishable Product Inventory. *Management Science,* 23, 512-521.

[19] Cohen, M. A. (1976). Analysis of Single Critical Number Ordering Policies for Perishable Inventories, *Operations Research*, 24, 726-741.

[20] Cohen, M. A. and D. Pekelman, (1979). Lifo Inventory Systems, it Management Science, 24, 1150-1162.

[21] Cohen, M.A., P. Pierskalla and H. Yen, (1981). Analysis of Ordering and Allocation Policies for Multi-Echelon, Age- Differentiated Inventory Systems, *TIMS Studies in the Management Sciences*, 16, 353-378.

[22] Cote, M. J. and W. E. Stein, (2000). An Erlang Based Stochastic Model for Patient Flow, *Omega, Int. J. Mgmt. Sci.*, 28, 347-359.

[23] Cox, T. F., J.R Birchall, H Wong, (1985). Optimising the queuing system for an ear, nose, and throat outpatient clinic, *Journal of Applied Statistics*, 12, 113126.

[24] Creemers, S. and M. R. Lambrecht (2007). Modeling a Healthcare System as a Queueing Network: The Case of a Belgian Hospital. Available at SSRN: http://ssrn.com/abstract=1093618 or http://dx.doi.org/10.2139/ssrn.1093618

[25] Daniels, H. E. (1967). The distribution of the total size of an epidemic. Proc. 5th Berkeley Symp. Math. Statist. Prob. 4, 281-293.

[26] Daley, D. J. and J. Gani (1999). *Epidemic Modelling: An Introduction*, Cambridge Studies in mathematical Biology: 15, Cambridge University Press, Cambridge, UK.

[27] David, I. and U. Yechiali (1985). A Time-Dependent Stopping Problem with Application to Live Organ Transplants, *Operations Research,* 33, 491-504.

[28] David, I. and Yechiali, U. (1990). Sequential assignment match processes with arrival of candidates and offers. Probability in the Engineering and Informational Sciences 4: 413430.

[29] David, I. and U. Yechiali (1995). One-attribute sequential assignment match processes in discrete time. *Operations Research*, 43, 879-884.

[30] Davies, R. and HTO Davies, (1994). Modelling Patient Flows and Resource Provision in Health Systems, *Omega, Int. J. Mgmt., Sci.*, 22, 123-131.

[31] Davies, R., D. Johnson and S. Farrow, (1975). Planning Patient Care with a Markov Model, *Operational Research Quarterly*, 26, 599-607.

[32] Denton, B. and Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling, *IIE Transactions*, 35, 1003-1016.

[33] Derman, C., G. J. Lieberman, and S. M. Ross (1972). A Sequential Stochastic Assignment Problem, *Management Science*, 18, 349-355.

[34] Downton, F. (1968). The ultimate size of carrier-borne epidemics. Biometrika 55, 277-289.

[35] Duncan, I. B. and R. N. Curnow, (1978). Operational Research in the Health and Social, *Journal of the Royal Statistical Society*, Series A (General), 141, 153-194.

[36] Eick, S. G., W. A. Massey and W. Whitt (1993). The Physics of the $Mt/G/\infty$ Queue. *Operations Research*, 41, 731-742.

[37] Esogbue, A. O. and A. J. Singh, (1976). A Stochastic Model for an Optimal Priority Bed Distribution Problem in a Hospital Ward. *Operations Research,* 24, 884-898.

[38] Feldman, J., N. Liu, H. Topaloglu, and S. Ziya (2014). Appointment Scheduling Under Patient Preference and No- Show Behavior. *Operations Research*, 62, 794-811.

[39] Fomundam, S. and J. Hermann (2007). A Survey of Queueing Theory Applications in Healthcare, ISR Technical Report 2007-24.

[40] Frank, S. A. (2007). *Dynamics of Cancer: Incidence, Inheritance, and Evolution.* Princeton University Press, NJ.

[41] Frank S. A. (2004). A multistage theory of age-specific acceleration in human mortality. *BMC Biology Online Journal.* 2, 16.

[42] Fries, B., (1975). Optimal Ordering Policy for a Perishable Commodity with Fixed Lifetime, *Operations Research*, 23, 46-61.

[43] Gani, J, S. Yakowitz and M. Blount (1997). The Spread and Quarantine of HIV Infection In a Prison system. *SIAM Journal of Applied Mathematics*, 57, 1510-1530.

[44] Gentleman R. C., J. F. Lawless, J. C. Lindsey and P. Yan (1994). Multistate Markov-models for analyzing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13, 805–821.

[45] Gorunescu, F., S. I. McClean and P. H. Millard (2002). A Queueing Model for Bed-Occupancy Management and Planning of Hospitals. *The Journal of the Operational Research Society,* 53, 19-24.

[46] Graves, S. C., H.S. Leff, J Natkins, and M Senger, (1983). A simple stochastic model for facility planning in a mental health care system, *Interfaces,* 13, 101110.

[47] Green, L.V. and Nguyen, V., (2001). Strategies for cutting hospital beds: the impact on patient service. *Health Services Research,* 36, 421-442.

[48] Green, L. V., P. J. Kolesar, J. Soares, (2001). Improving the SIPP approach for staffing service systems that have syclic demands. *Operations Research,* 49, 549564.

[49] Green, L. (2004). Capacity Planning and Management in Hospitals, *Operations Research and Health Care*, International Series in Operations Research and Management Science, 2005, 70, Part 1, 15-41.

[50] Green, L. (2006). Queueing Analysis in Healthcare, in *Patient Flows: Reducing Delay in Healthcare,* edited by R. Hall, Springer, 281-307.

[51] Green, L. V. and Savin, S. (2008). Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56, 1526-1538.

[52] Green, L., P. Kolesar, and W. Whitt (2007). Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System, *Production and Operations Management*, 16, 1339.

[53] Greenwood, P. and L. F. Gordillo (2009). Stochastic Epidemic Modeling, book chapter in *Mathematical and Statistical Estimation Approaches in Epidemiology*, Ed Chowell et al, Springer Science+Business Media B. V.

[54] Griffiths, J. D., N. Price-Lloyd, M. Smithies and J. Williams, (2007). A queueing model of activities in an intensive care unit. *IMA Journal of Management Mathematics,* 17, 277-288.

[55] Gupta, D. and Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities, *IIE Transactions*, 40, 800-819.

[56] Gupta, D. and L. Wang (2008). Revenue management for a primary-care clinic in the presence of patient choice, *Operations Research*, 56, 576-592.

[57] Hall, R. (2006) *Patient Flows: Reducing Delay in Healthcare,* Springer, NJ.

[58] Halfin, Shlomo; Whitt, Ward (1981). Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, 29, 567-588.

[59] Hassin, R. and S. Mendel, (2008). Scheduling arrivals to queues: A single server model with no-shows. *Management Science,* 54, 565572.

[60] Hauskrecht, M. and H. Fraser (2000). Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artificial Intelligence in Medicine*, 18, 221-244.

[61] Hershey, J. C, E.N Weiss, M.A Cohen, (1981). A stochastic service network model with application to hospital facilities, *Operations Research*, 29, 122.

[62] Hethcote, H. W. and P. Waltman (1973). Optimal Vaccination Schedules in a Deterministic Epidemic Model, *Mathematical Biosciences*, 18, 365-381.

[63] Hethcote, H. H. (1989). Three Basic Epidemiological Models. In *Applied Mathematical Ecology*, Eds. Levin, Hallam, Gross. Springer, NY.

[64] Isham, V. (2004). Stochastic Models for Epidemics. Avavilable at http://www.ucl.ac.uk/statistics/research/pdfs/rr263.pdf.

[65] Jackson C. H. and L. D. Sharples (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, 21, 113–128.

[66] Jackson, C. H., L. D. Sharples, S. G. Thompson, S. W. Duffy, E. Couto(2003). Multistate Markov models for disease progression with classification error. *The Statistician,* 52, Part 2, 193209.

[67] Jagannathan, R. and T. Sen, (1991). Storing Crossmatched Blood: A Perishable Inventory Model with Prior Allocation. *Management Science,* 37, 251-266.

[68] Jiang, L. and R. E. Giachetti, (2008). A queueing network model to analyze the impact of parallelization of care on patient cycle time, *Health Care Manage Sci*, 11, 248261.

[69] Kaandorp, G. C. and G. Koole (2007). Optimal outpatient appointment scheduling, *Health Care Manage Sci.* 10, 217229

[70] Kao, E. P. C. (1972). A semi-Markov model to predict recovery progress of coronary patients. *Health Services Research,* 7, 191-208.

[71] Kao, E. P. C. (1974). Modeling the movement of coronary patients within a hospital by semi-Markov processes. *Operations Research*, 22, 683-699.

[72] Kay, R. (1986). A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. *Biometrics,* 42, 855-865.

[73] Keller, T.F., D.J Laughhunn. (1973). An application of queuing theory to a congestion problem in an outpatient clinic, *Decision Sciences*, 4, 379394.

[74] Klassen, K. J., T. R. Rohleder, (2004). Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment, *International Journal of Service Industry Management*, 15, 167 - 186.

[75] Klein, J. P., J. H. Klotz and M. R. Grever (1984). A Biological Marker Model for Predicting Disease Transitions. *Biometrics,* 40, 927-936.

[76] Koizumi, N., E. Kuno and T. E. Smith, (2005). Modeling patient flows using a queuing network with blocking. *Health Care Manage Sci*, 4960.

[77] Kermack, W. O. and McKendrick, A. G. (1927). Contributions to the mathematical theory of epidemics, part I. *Proceedings of the Royal Society of London*, A115, 700721.

[78] Kulkarni, V. G. (2010). *Modeling and Analysis of Stochastic Systems*, Second Edition, CRC Press, NY.

[79] Kurt, M., M. Roberts, A. Schaefer, and M. Utku, (2011). Valuing Prearranged Paired Kidney Exchanges: A Stochastic Game Approach , Working Paper, Boston College.

[80] Kuske, R., L. F. Gordillo, and P. Greenwood (2007). Sustained oscillations via coherence resonance in SIR, *Journal of Theoretical Biology,* 245, 459-469.

[81] LaGanga, L. R. and S. R. Lawrence (2012). Appointment overbooking in health care clinics to improve patient service and clinic performance. *Prod. Oper. Manag.*, 21, 874888.

[82] Lefevre, C. (1981). Optimal Control of a Birth and Death Epidemic Process. *Operations Research,* 29, 971-982.

[83] Longini, J. I. M., W. S. Clark, R. H. Byers, J. W. Ward, W. W. Darrow, G. F. Lemp, and H. W. Hethcote (1989). Statistical analysis of the stages of HIV infection using a Markov model, *Statistics in Medicine,* 8, 831-843.

[84] Liu, N., S. Ziya, V. G. Kulkarni (2010). Dynamic Scheduling of Outpatient Appointments under Patient No-shows and cancellations, *MSOM*, 12, 347-364.

[85] Luo, J., V. G. Kulkarni, S. Ziya (2012). Appointment Scheduling under Patient No-Shows and Service Interruptions, *MSOM*, 14, 67-684.

[86] Luo, J., V. G. Kulkarni, S. Ziya (2015). A Tandem Queueing Model for an Appointment-based Service System, *QUESTA*, 79, 53-85.

[87] Marchetti, M., S. Quaglini, and G. Barosi (1998). Prophylactic splenectomy and cholecystectomy in mild hereditary spherocytosis: analyzing the decision in different clinical scenarios. *Journal of Internal Medicine*, 244, 217-226.

[88] Magni, P., S. Quaglini, M. Marchetti and G. Barosi (2000). Deciding when to intervene: a Markov decision process approach. *International Journal of Medical Informatics*, 60, 237-253.

[89] Maillart, L. M., J. S. Ivy, S. Ransom and K. Diehl (2008). Assessing Dynamic Breast Cancer Screening Policies, *Operations Research*, 56, 14111427.

[90] Mandelbaum, A., W. A. Massey, M. I. Reiman, R. Rider, (1999). Time-varying multiserver queues with abandonments and retrials. P. Key, D. Smith, eds. *Proceedings of the 16th International Teletraffic Congress*, 355364.

[91] Marshall, G. and R. H. Jones (1995). Multi-State Models and Diabetic Retinopathy. *Statistics in Medicine,* 14, 1975-1983.

[92] Massey, W. A., and W. Whitt (1999). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems,* 13, 183-250.

[93] McManus, M. L., N. C. Long, C. Michael, A. Cooper, E. Litvak, (2004). Queuing Theory Accurately Models the Need for Critical Care Resources, *Anesthesiology,* 100, 1271-1276.

[94] Mercer, A. (1960). A Queueing Problem in which the Arrival Times of the Customers are Scheduled *Journal of the Royal Statistical Society.* Series B (Methodological), 22, 108-113.

[95] Mercer, A. (1973). Queues with Scheduled Arrivals: A Correction, Simplification and Extension. *Journal of the Royal Statistical Society,* 35, 104116.

[96] Murray, M., C. Tantau, (2000). Same-day appointments: Exploding the access paradigm. *Family Practice Management,* 7, 45-50.

[97] Muthuraman, K. and M. Lawley (2008). A stochastic overbooking model for outpatient clinical scheduling with no-shows, *IIE Trans.,* 9, 820-837.

[98] Nasell, I. (2002). Stochastic models of some endemic infections. *Mathematical Biosciences*, 179, 119.

[99] Nahmias, S. and W. P. Pierskalla, (1973). Optimal Ordering Policies for a Product that Perishes in Two Periods Subject to Stochastic Demand. *Naval Res. Log. Quart.* 20, 207-229.

[100] Nahmias, S. (1975). Optimal Ordering Policies for Perishable Inventory-II, *Operations Research,* 23, 735-749.

[101] Nahmias, S. (1976). Myopic Approximations for the Perishable Inventory Problem, *Management Science,* 22, 1002-1008.

[102] Osorio, C., and M. Bierlaire, (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *European Journal of Operational Research,* 196, 996-1007.

[103] Pegden, C. D. and M. Rosenshine, (1990). Scheduling aarivals to queues. *Computers Opns Res.* 17, 343-348.

[104] Pegels, C. C. and Jelmert, A. E. (1970). An Evaluation of Blood-Inventory Policies: A Markov Chain Application. *Operations Research*, 18, 1087-1098.

[105] Perez-Ocon R, J. E. Ruiz-Castro, M. L. Gamiz-Perez. (1998). A multivariate model to measure the effect of treatments in survival to breast cancer. *Biometrica*, 40, 703-715.

[106] Perez-Ocon R, Ruiz-Castro JE, Gamiz-Perez ML. (2001). Non-homogeneous Markov models in the analysis of survival after breast cancer. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 50, 111–124.

[107] Pierskalla, P. and C. D. Roach, (1972). Optimal Issuing Policies for Perishable Inventory *Management Science,* 18, Theory Series, 603-614.

[108] Pierskalla, W. (2004). Supply chain management of blood banks.

[109] Picard, P. (1980). Applications of Martingale Theory to Some Epidemic Models, *Journal of Applied Probability,* 17, 583-599.

[110] Prastacos, G. P. (1978). Optimal myopic allocation of a product with fixed lifetime, *The Journal of the Operational Research Society,* 29, 905-913.

[111] Prastacos, G. P. (1979). LIFO Distribution Systems, *The Journal of the Operational Research Society,* 30, 539-546.

[112] Prastaos, G. P. (1984). Blood Inventory Management: An Overview of Theory and Practice *Management Science*, 30, 777-800.

[113] Reluga, T. C. (2010). Game Theory of Social Distancing in Response to an Epidemic. PLoS Comput Biol 6(5): e1000793. doi:10.1371/journal.pcbi.1000793

[114] Righter, R. L. (1987). The Stochastic Sequential Assignment Problem with Random Deadlines. *Prob. Engrg. Inform. Sci.* 1, 189-202.

[115] Righter, R.L. (1988). A resource allocation problem in a random environment. *Operations Research*, 37, 329-338.

[116] Righter, R. L. (2011). The Stochastic Sequential Assignment Problem with Arrivals. *Prob. Engrg. Inform. Sci.* (Upcoming).

[117] Rais, A. and A. Viana (2011). Operations Research in Healthcare: a survey *International Transactions in Operational Research*, 18, 131.

[118] Schaefer, A. J., M. D. Bailey, S. M. Shechter and M. S. Roberts (2005). Modeling medical treatment using Markov decision processes, *Operations Research and Health Care*, International Series in Operations Research and Management Science, 2005, Volume 70, Part 3, 597-616.

[119] Selen, J., I. J. B. F. Adan, V. G. Kulkarni and J. S. H. van Leeuwaarden (2016). The snowball effect of customer slowdown in critical many-server systems, *Stochastic Models*, 32, 366-391.

[120] Sethi, S. (1978). Optimal Quarantine Programmes for Controlling an Epidemic Spread. *The Journal of the Operational Research Society,* 29, 265-268.

[121] Sethi, S. P. and P. W. Staats (1978). Optimal Control of Some Simple Deterministic Epidemic Models. *The Journal of the Operational Research Society*, 29, 129-136.

[122] S. M. Shechter S. M., M. D. Bailey, A. J. Schaefer, and M. S. Roberts (2008). The Optimal Time to Initiate HIV Therapy Under Ordered Health States. *Operations research*, 56, 2033.

[123] Shonick, W. and J. R. Jackson, (1973). An Improved Stochastic Model for Occupancy-Related Random Variables in General-Acute Hospitals, *Operations Research,* 21, 952-965.

[124] Shwartz, M. 1978. A mathematical model used to analyze breast cancer screening strategies. *Operations Research*, 26, 937955.

[125] Stafford, E. F. Jr., S.C Aggarwal (1979). Managerial analysis and decision-making in outpatient health clinics, *Journal of the Operational Research Society*, 30, 905915.

[126] Stein, W.E. and M. J. Cote (1994). Scheduling arrivals to a queue. *Computers Ops Res.*, 21, 607-614.

[127] Su, X. and S. A. Zenios (2005). Patient Choice in Kidney Allocation: Role of the Queueing Discipline, *MS&OM*, 6, 280-301.

[128] Su, X. and S. A. Zenios (2005). Patient Choice in Kidney Allocation: A Sequential Stochastic Assignment Model, *Operations Research*, 53, 443-455.

[129] Su, X. and S. A. Zenios (2006). Recipient Choice Can Address the Efficiency-Equity Trade-off in Kidney Transplantation: A Mechanism Design Model, *Management Science,* 52, 1647-1660.

[130] Sweeting, M. J., V. T. Farewella and D. De Angelisa (2009). Multi-state Markov models for disease progression in the presence of informative examination times: An application to hepatitis C. *Statistics in Medicine,* 29, 1161-1174.

[131] Tijms, H. C. (2003). *A first course in stochastics models*, Wiley.

[132] A. Turkcan, B. Zeng, K. Muthuraman, M. Lawley (2011). Sequential clinical scheduling with service criteria, *EJOR,* 214, 780795.

[133] Vassilacopoulis, G. (1985). Allocating doctors to shifts in an accident and emergency department. *Journal of the Operational Research Society*, 36, 517-523.

[134] Wang, P. P. (1993). Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Res Logist Quart,* 40, 345-360.

[135] Wang, P. P. (1994). Releasing N jobs to an unreliable machine, *Computers and Industrial Engineering*, 26, 661-671.

[136] Wang, P. P. (1997). Optimally scheduling N customer arrival times for a single-server system, *Computers and Operations Research,* 24, 703-716.

[137] Wang, P. P. (1999). Sequencing and scheduling N customers for a stochastic server, *European Journal of Operational Research*, 119, 729-738.

[138] Wang, W. and D. Gupta (2011). Adaptive appointment systems with patient preferences, *Manufacturing & Service Operations Management*, 13, 373-389.

[139] Weiss, G. H. (1965). On the spread of epidemics by carriers. Biometrics 21, 481-90.

[140] Weiss, G. H. and M. Dishon (1971). On the asymptotic behavior of the stochastic and deterministic models of an epidemic. *Mathematical Biosciences,* 11, 261-265.

[141] Weiss, E. N. and J. O. McClain, (1987). Administrative Days in Acute Care Facilities: A Queueing-Analytic Approach. *Operations Research,* 35, 35-44.

[142] Winsten, C. B. (1959). Geometric distributions in the theory of queues, *J. R. Statist. Soc., B,* 21, 1-35.

[143] Wong J. B., F. A. Sonnenberg, D. N. Salem and S. G. Pauker (1990). Myocardial revascularization for chronic stable angina. *Annals of Internal Medicine*, 113, 852871.

[144] Worthington, D.J. (1987). Queueing models for hospital waiting lists. *Journal of the Operations Research Society,* 38, 413-422.

[145] Xie, H., T. Chaussalet and M. Rees, (2009). A Semi-open Queueing Network Approach to the Analysis of Patient Flow in Healthcare Systems, Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)

[146] Xu, S. and Zenios, S. (2004). Patient Choice in Kidney Allocation: The Role of the Queueing Discipline, *MS& OM*, 6, 280301.

[147] Yan, X., Y. Zhou and J. Li (2007). Optimal quarantine and isolation strategies in epidemics control. *World Journal of Modelling and Simulation*, 3, 202-211.

[148] Yu, S., V. G. Kulkarni, V. Deshpande (2016). Staffing and Scheduling for a Health Care Facility with Series Patients, Submitted to *MSOM.*

[149] Yu, S., N. Lee, V. G. Kulkarni, H. Sehn (2016). Server Allocation at Virtual Computing Labs via Queueing Models and Statistical Forecasting, Submitted to *IIE Transactions.*

[150] Zacharias, C., and M. Pinedo (2014). Appointment Scheduling with No-Shows and Overbooking, *Producton and Operations Management*, 23, 788801.

[151] Zelen, M. 1993. Optimal scheduling of examinations for the early detection of disease. *Biometrika*, 80, 279293.

[152] Zenios, S. (1999). Modeling the transplant waiting list: A queueing model with reneging, *Queueing Systems*, 31, 239251.

[153] Zenios, S. (2002). Optimal Control of a Paired-Kidney Exchange Program, *Management Science*, 48, 328-342.

[154] Zenios, S. (2005). Models for Kidney Allocation, *Operations Research and Health Care*, International Series in Operations Research and Management Science, 2005, Volume 70, Part 3, 537-554.

[155] Zenios, S., G. M. Chertow and L. M. Wein (2000). Dynamic Allocation of Kidneys to Candidates on the Transplant Waiting List. *Operations Research,* 48, 549-569.

[156] Zhang, Yu, and V. G. Kulkarni (2017). Two-Day Appointment Scheduling with Patient Preferences and Geometric Arrivals, to appear in *QUESTA.*

[157] van Zyl, G. (1964). Inventory Control for Perishable Commodities, Ph.D. Disserta- tion, University of North Carolina, Chapel Hill, N. C.