

Efficiency evaluation for pooling resources in health care

Peter T. Vanberkel¹, Richard J. Boucherie², Erwin W. Hans³
Johann L. Hurink², Nelly Litvak²

University of Twente, Enschede, The Netherlands

^{1,2} Department of Applied Mathematics,
Faculty of Electrical Engineering, Mathematics, and Computer Science

^{1,3} Operational Methods for Production and Logistics,
School of Management and Governance

Abstract

Hospitals traditionally segregate resources into centralized functional departments such as diagnostic departments, ambulatory care centres, and nursing wards. In recent years this organizational model has been challenged by the idea that higher quality of care and efficiency in service delivery can be achieved when services are organized around patient groups. Examples include specialized clinics for breast cancer patients and clinical pathways for diabetes patients. Hospitals are struggling with the question of whether to become more centralized to achieve economies of scale or more decentralized to achieve economies of focus. Using quantitative Queueing Theory and Simulation models, we examine service and patient group characteristics to determine the conditions where a centralized model is more efficient and conversely where a decentralized model is more efficient. The results from the model measure the tradeoffs between economies of scale and economies of focus from which management guidelines are derived.

Keywords: Slotted Queueing Model, Simulation, Resource Pooling, Focused Factories, Health Care Modelling

1 Introduction

Health care facilities are under mounting pressure to both improve the quality of care and decrease costs by becoming more efficient. Efficiently organizing the delivery of care is one way to decrease cost and improve performance. At a national level this is achieved by aggregating services into large general hospitals in major urban centres, thereby gaining efficiencies through economies of scale (EOS). At the same time, some hospitals are becoming more specialized and

¹Corresponding Author: 7500AE Enschede, The Netherlands, p.t.vanberkel@utwente.nl, Phone: +31 53 489 5480, Fax: +31 53 489 2159

offer a limited range of services aiming to breed competence and improve service rates [15]. Such strategies aim to improve performance through focus.

At the hospital level, similar strategies to exploit focus are being considered [18, 20]. Rather than organizing departments around function (e.g. radiology, phlebotomy, etc.), departments dedicated to treating a particular patient population are being created. Examples included focused departments for back patients [22], cancer patients [14, 21], outpatients [16], trauma patients [11] and inpatients [10, 23]. In these studies the benefits of increased focus have shown mixed results, leading hospital managers to struggle with the choice to become more centralized to achieve EOS or more decentralized to achieve economies of focus (EOF). In this paper we examine service and patient population characteristics to determine under which circumstances the functional department, and conversely the patient focused department, is more efficient.

We derive an analytic approximation measuring EOS losses associated with unpooling resources. This approximate along with simulations of typical clinic environments provides the insight from which we develop general management guidelines and reference tables. The reference tables allow managers to “look up” specific results for 80 different clinic environments. Furthermore, the model relies only on typically available data and can easily be used to analyze specific clinic environments. The model and framework can represent any hospital department where the service time is less than one day and where the system empties between days. This includes outpatient clinics, diagnostic clinics and operating theaters. To our knowledge such a robust model for measuring the effects of pooling and unpooling has not been developed before.

The paper is organized as follows. Section 2 introduces the pooling principle and the debate between centralized and decentralized departments. Section 3 introduces the model used to measure the EOS lost in an unpooled system. Section 4 provides results and analysis for a series of numeric experiments of typical clinic environments. Section 5 summarizes the computational results and provides guidelines for hospital managers. Section 6 briefly discusses potential future research.

2 The Pooling Principle

In this section we summarize the pooling principle described in [4] as, “pooling of customer demands, along with pooling of the resources used to fill those demands” in order to “yield operational improvements.” This implies that a centralized (pooled) clinic that serves all customer types may achieve shorter waiting times than a number of decentralized (unpooled) clinics focusing on a more limited range of customer types. The intuition for this principle is as follows. Consider the situation in the unpooled setting, when a customer is waiting in one queue while a server for a different queue is free. Had the system been pooled in this situation, the waiting customer could have been served by the idle server, and thus experience a shorter waiting time. The gain in efficiency is a form of EOS.

Statistically, the advantage of pooling is credited to the reduction in variability due to the portfolio effect [9]. This is easily demonstrated for cases where the characteristics of the unpooled services are identical. For this discussion see [2, 7, 12]. However, pooling is not always of benefit. There may be situations where the pooling of customers actually adds variability to the system thus offsetting any efficiency gains, see [6]. Furthermore when the target performances of

customer types differ it may be more efficient to use dedicated capacity (i.e. unpooled capacity), see [3, 12]. And finally, in the pooled case all servers must be able to accommodate all demand. This flexibility may be expensive and, as is more directly related to this paper, may actually cause inefficiencies as servers are no longer able to focus on a single customer type.

It is clear that pooling is offered as a potential method to improve a system’s performance without adding additional resources. Interestingly, the principle of focus which “advocates for hospitals to abandon functional, discipline-focused departments (e.g., radiology, nursing, etc.) in favor of a design organized around patients and their diagnoses” [11, 13, 17], implies the same. In this paper we aim to enhance understanding of these seemingly contradictory view points.

3 Model

A discrete time slotted queueing model is used to evaluate the tradeoff between EOS and EOF. More specifically, the access time for a centralized ambulatory clinic serving all patient types is compared to the access time of decentralized clinics, focusing on a more limited range of patient types. Generally speaking the decentralized method results in longer access time, due to the loss in EOS. The model quantifies this loss and computes the improvement in service time required in the decentralized departments in order to achieve the equivalent access time as in the centralized department. This improved service time represents the amount of improvement due to focus (or EOF) necessary to offset the losses of EOS.

We describe the queueing model using language from an ambulatory clinic setting. For example, referrals for appointments are considered new arrivals, appointment length is the service time, the number of consultation rooms reflects the number of servers and finally, the time a patient must wait for a clinic appointment (often referred to as access time in health care literature) is the waiting time in the queue. The model can be used for any hospital department where the service time is less than one day and where the system empties between days (e.g. operating room or diagnostic clinics). In this paper the following notation is used:

- λ = Average demand for appointments per day
- D = Average appointment length in minutes
- V = Variance of the appointment length
- C = Coefficient of Variance for the appointment length $\left(C = \sqrt{V/D^2}\right)$
- M = Number of rooms
- ρ = Utilization of the rooms
- t = Working minutes per day
- W = Expected Waiting Time in days

A subscript “AB” corresponds to the pooled case and a subscript “A” or “B” corresponds to the unpooled case for patient groups “A” or “B” respectively. The schemes of the pooled and unpooled systems are shown in Figure 1.

When combined, the parameters of the unpooled system must equal the parameters of the pooled system. The parameters for two patient groups describe the patient mix. How the pa-

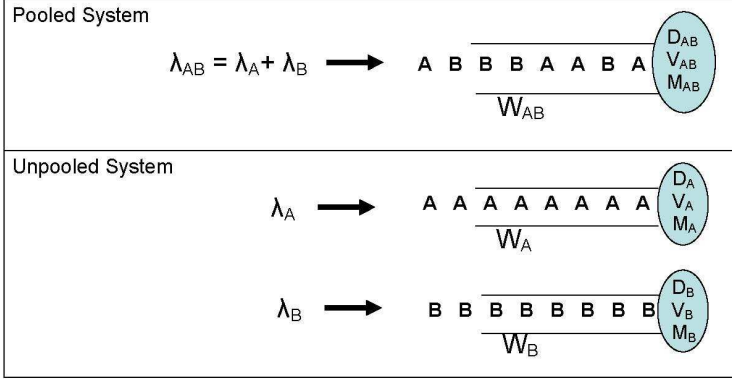


Figure 1: Scheme of the Pooled and Unpooled Systems

tient mix parameters in the unpooled system relate to the parameters in the pooled system is described below. These division “rules” imply that no additional resources become available in the unpooled setting and that patients are strictly divided into one or the other group.

$$M_{AB} = M_A + M_B \quad (1)$$

$$\lambda_{AB} = \lambda_A + \lambda_B \quad (2)$$

$$D_{AB} = qD_A + (1 - q)D_B \quad (3)$$

$$V_{AB} = q(V_A + D_A^2) + (1 - q)(V_B + D_B^2) - D_{AB}^2 \quad (4)$$

where $q = \lambda_A/\lambda_{AB}$.

Initially the waiting time in the three queueing systems depicted in Figure 1 are evaluated separately. The characteristics of the three systems are the same and as such the same model is used to evaluate them (the input parameters are changed to reflect the pooled and unpooled systems). The model is described in the following Subsections where the subscripts “A”, “B” and “AB” are left out for clarity.

3.1 Modelling Arrivals and Services

The mean (D) and variance (V) of appointment lengths is readily available in most ambulatory clinics. Relying only on these data, we use renewal theory approximations to estimate the number of appointments completed during one clinic day. We assume that D is i.i.d. and that $D \ll t$. $N(t)$ is defined as the number of appointments completed in one room between $[0, t]$. Under these assumptions, from renewal theory [19] (pg 315) we find

$$E[N(t)] \approx \frac{t}{D} + \frac{1}{2}(C^2 - 1). \quad (5)$$

Let M be the number of rooms, $N_i(t)$ the number of completed appointment in room $i = 1, \dots, M$. We assume that $N_i(t)$ s are independent and let S be the total number of completed

appointments per clinic day given a clinic has M rooms. Then

$$S = \sum_{i=1}^M N_i(t) \quad E[S] \approx ME[N(t)] \approx \frac{Mt}{D} + \frac{M}{2}(C^2 - 1). \quad (6)$$

Note that renewal theory approximation implies that $E[S]$ increases as C increases. Although perhaps counter-intuitive, this means that as the variance in the clinic increases, so too do the number of completed cases per day.

Let $V_{N(t)}$ and V_S be the variance of $N(t)$ and S respectively. Then the two-moment renewal theory approximation for $V_{N(t)}$ and V_S is as follows

$$V_{N(t)} \approx \frac{V^2 t}{D^3} = \frac{C^2 t}{D} \quad (7)$$

$$V_S \approx MV_{N(t)} = \frac{MC^2 t}{D}. \quad (8)$$

We note that (5), (6), (7) and (8) are based on the assumption $D \ll t$. In a contrary situation (e.g. chemotherapy, where appointments may last half the day) the influence of D , V , C on S is not so direct but the model is still valid [21].

In our model we assume the arrival process is Poisson. Let X be the arrivals per day and V_X and C_X be the variance and coefficient of variance of X respectively. Since X is distributed according to Poisson(λ) it follows that $E[X] = \lambda$, $V_X = \lambda$ and $C_X = 1/\lambda$.

3.2 Clinic Load

Workload in a clinic is measured by the utilization of its rooms. The standard measure of server utilization (ρ) is computed by $\rho = \lambda/(ME[N(t)])$. Using (6) we approximate ρ as follows

$$\begin{aligned} \rho &\approx \frac{\lambda}{\frac{Mt}{D} + \frac{M}{2}(C^2 - 1)} = \frac{\lambda D}{Mt} \frac{1}{1 + \frac{D}{2t}(C^2 - 1)} \\ &= \frac{\lambda D}{Mt} + \frac{\lambda D}{Mt} \left\{ \sum_{i=1}^{\infty} (-1)^i \left(\frac{D}{2t} (C^2 - 1) \right)^i \right\}. \end{aligned} \quad (9)$$

Where the last equality holds provided $|D/(2t)(C^2 - 1)| < 1$ (which is true in our cases since $D \ll t$). The second term in the last expression of (9) is of the order D/t and since we assume that $D \ll t$, it follows that it is small relative to the first term. From this observation we introduce ρ_0 as an estimate of ρ and define it as follows

$$\rho_0 = \frac{\lambda D}{Mt}. \quad (10)$$

In our simulation experiments of Section 4 we keep ρ_0 fixed for each setup. Because of the correction term in (9), actual ρ changes slightly depending on the patient mix parameters. For example if λ_A/λ_{AB} changes while C_A and C_B remain constant, than C_{AB} must change according to (4). This consequently causes slight changes in $E[S]$ and in turn in ρ .

3.3 Waiting Times

With these input parameters the expected queue length is computed using Lindley's Recursion [5]. Consider subsequent days $1, 2, \dots$, and let L_n be the queue length at the beginning of day n . Further, let X_n be the number of arrivals on day n , and S_n the number of services that can possibly be completed on day n . We assume that X_n and S_n , $n > 1$, are independent and distributed as described above. The number of appointment requests on day n is then $L_n + X_n$, and the dynamics of the queue length process is given by

$$L_{n+1} = (L_n + X_n - S_n)^+; \quad n > 1 \quad (11)$$

where $x^+ = x$ if $x \geq 0$ and $x^+ = 0$ otherwise.

If $n \rightarrow \infty$ then the expectation of L_n converges to its equivalent value L .

To compute the expected waiting time W we use Little's Law ($W = L/\lambda$). A related model described in [21] explains how to compute the waiting time distribution through a similar recursion. In general, equation (11) is hard to solve analytically. A variety of techniques, such as Wiener-Hopf factorization, have been developed but they usually lead to explicit solutions only in special cases. In the simulation experiments of Section 4 we solve (11) numerically.

The average queue length (L) in our slotted queueing model is analogous to the average waiting time of a GI/GI/1 queue because both are measured by Lindley's Recursion. The waiting time of a GI/GI/1 queue can be approximated with Allen-Cunneen approximation [1] thus leading to an approximation for L in our slotted model. Using (6) and (8) and the assumption that $D \ll t$, we write the approximation formula as

$$\begin{aligned} L &\approx \lambda \frac{\rho}{1-\rho} \frac{C_S^2 + (1/\lambda)^2}{2} = \lambda \frac{\rho}{2(1-\rho)} \left(\frac{1}{\lambda} + \frac{MC^2t}{D} \frac{1}{M^2 \left(\frac{t}{D} + \frac{1}{2}(C^2 - 1) \right)^2} \right) \\ &\approx \frac{\rho}{2(1-\rho)} \left(1 + \frac{C^2}{\rho_0} \right). \end{aligned} \quad (12)$$

Using Little's Law and (12) we approximate the expected waiting by

$$W \approx \frac{\rho}{2(1-\rho)\lambda} \left(1 + \frac{C^2}{\rho_0} \right). \quad (13)$$

3.4 Required Change in Service Time

To compare the performance of the pooled and unpooled systems, W is computed for the three queueing systems depicted in Figure 1. The objective of the model is to determine a new appointment length (D'_A) required to make $W_A = W_{AB}$. As a standard measure we define Z_A as the proportional difference between D_A and D'_A (likewise for D'_B and Z_B). Ignoring the subscripts "A" and "B" we formally define Z as follows

$$Z = \frac{D'}{D} - 1. \quad (14)$$

Z essentially measures the EOF needed to make the access time in the pooled and unpooled systems equal. Z can be both negative and positive. When Z is negative it represents the amount the appointment length must decrease (attributed to the increased focus on a single patient group) in order to overcome any EOS losses resulting from unpooling. When Z is positive it indicates that the appointment length can increase and still maintain the same service level as in the pooled system. This happens when the number of rooms assigned to one of the patient classes is large. Although practically less relevant, the positive Z value does help illustrate how the tradeoff between EOS and EOF is influenced by the distribution of rooms.

In the simulation experiments of Section 4, Z_A is computed by incrementally decreasing [or increasing] D_A by Z_A , until $W_A \leq W_{AB}$ [$W_A \geq W_{AB}$]. The percentage change (Z_B) for patient group B is computed in the same manner. These computations are automated with Microsoft Visual Basic.

Using our estimation (13) for W , we show how the Z values can also be estimated. First we assume $\rho_0 \approx \rho$ and define ρ'_0 as the load in unpooled clinic A with appointment length D'_A .

$$\rho'_0 = \frac{\lambda_A D'_A}{M_A t}$$

Next we set the waiting time approximations (13) for the pooled and unpooled system A equal to each other.

$$\frac{\rho'_0}{2(1 - \rho'_0)\lambda_A} \left(1 + \frac{C_A^2}{\rho'_0}\right) = \frac{\rho_0}{2(1 - \rho_0)\lambda_{AB}} \left(1 + \frac{C_{AB}^2}{\rho_0}\right) \quad (15)$$

We also assume the servers are divided between the pooled and unpooled clinics in such a way that the clinic load remains the same. From this it follows

$$\rho_0 = \frac{D_{AB}\lambda_{AB}}{M_{AB}t} \approx \frac{D_A\lambda_A}{M_A t}.$$

Finally, with algebra and by ignoring second order and higher terms of $(1 - \rho_0)$ we solve (15) for D'_A/D_A to obtain

$$Z_A = \frac{D'_A}{D_A} - 1 \approx \left(1 - \frac{1 + C_A^2}{1 + C_{AB}^2} \frac{\lambda_{AB}}{\lambda_A}\right) (1 - \rho_0). \quad (16)$$

Similarly (16) can be rewritten to obtain $Z_B = D'_B/D_B - 1$. From 4 it can be shown that either Z_A or Z_B in (16) is negative.

We note that while deriving formula (16) we made a number of simplifying assumptions and ignored second order and higher terms of $(1 - \rho_0)$. Thus, one can expect that (16) gives an accurate approximation for Z_A only in some special cases, e.g., when ρ_0 is close to one. The main goal of deriving this formula however, is to reveal the main parameters that influence Z_A and to identify the importance of these parameters in reasonable hospital settings. To this end, our calculations show that ρ_0 , λ_A/λ_{AB} , and $(1 + C_A^2)/(1 + C_{AB}^2)$ are the most influential factors. Furthermore, (16) also indicates which factors can be ignored. The absence of M_{AB} and D_{AB} implies that their influence is minimal. This was also confirmed by simulations that we omit in this paper for brevity. Thus, in the rest of the paper we focus on the most influential factors appearing in (16).

Table 1: Relative importance of Factors Influencing Z_A , according to (16)

| # | Clinic Description | ρ_0 | $\frac{\lambda_A}{\lambda_{AB}}$ | $\frac{1+C_A^2}{1+C_{AB}^2}$ | Z_A |
|----|--|----------|----------------------------------|------------------------------|-------|
| 1 | Busy Clinic, $\lambda_A \gg \lambda_B, V_A \ll V_B$ | 0.99 | 0.7 | 0.32 | 0 |
| 2 | Busy Clinic, $\lambda_A \gg \lambda_B, V_A = V_B$ | 0.99 | 0.7 | 1 | -0.01 |
| 3 | Busy Clinic, $\lambda_A \gg \lambda_B, V_A \gg V_B$ | 0.99 | 0.7 | 1.36 | -0.01 |
| 4 | Busy Clinic, $\lambda_A \ll \lambda_B, V_A \ll V_B$ | 0.99 | 0.3 | 0.17 | 0 |
| 5 | Busy Clinic, $\lambda_A \ll \lambda_B, V_A = V_B$ | 0.99 | 0.3 | 1 | -0.03 |
| 6 | Busy Clinic, $\lambda_A \ll \lambda_B, V_A \gg V_B$ | 0.99 | 0.3 | 2.58 | -0.08 |
| 7 | Quite Clinic, $\lambda_A \gg \lambda_B, V_A \ll V_B$ | 0.7 | 0.7 | 0.32 | 0.16 |
| 8 | Quite Clinic, $\lambda_A \gg \lambda_B, V_A = V_B$ | 0.7 | 0.7 | 1 | -0.13 |
| 9 | Quite Clinic, $\lambda_A \gg \lambda_B, V_A \gg V_B$ | 0.7 | 0.7 | 1.36 | -0.29 |
| 10 | Quite Clinic, $\lambda_A \ll \lambda_B, V_A \ll V_B$ | 0.7 | 0.3 | 0.17 | 0.13 |
| 11 | Quite Clinic, $\lambda_A \ll \lambda_B, V_A = V_B$ | 0.7 | 0.3 | 1 | -0.7 |
| 12 | Quite Clinic, $\lambda_A \ll \lambda_B, V_A \gg V_B$ | 0.7 | 0.3 | 2.58 | -2.28 |

Table 2: Percentage by which Z_A is overestimated by (16)

| $\frac{\lambda_A}{\lambda_{AB}}$ | $\rho_0 = 0.79$ | $\rho_0 = 0.88$ | $\rho_0 = 0.97$ |
|----------------------------------|-----------------|-----------------|-----------------|
| 0.3 | 40.6% | 18.1% | 4.1% |
| 0.4 | 22.1% | 9.8% | 1.5% |
| 0.5 | 13.1% | 6.3% | 1.0% |
| 0.6 | 10.4% | 3.1% | 0.0% |
| 0.7 | 5.2% | 1.1% | 0.0% |

To illustrate the relative importance of terms ρ_0 , λ_A/λ_{AB} , and $(1 + C_A^2)/(1 + C_{AB}^2)$ in (16), consider the following typical ranges for each of them: $\rho_0 \in [0.7, 0.99]$; $\lambda_A/\lambda_{AB} \in [0.3, 0.7]$, as having values outside of this range implies a very small unpooled department which would be impractical [21]; $C_A^2, C_B^2 \in [0.5, 3]$. Note also that $(1 + C_A^2)/(1 + C_{AB}^2)$ depends on λ_A/λ_{AB} through (4). Table 1 shows twelve scenarios reflecting the border values of the three influential factors. We clearly observe that when ρ_0 is large it dominates Z_A and appears to be the most influential factor. It is also observable that the busier the clinic is, the smaller the loss in EOS. This is consistent with [7], who states that ‘‘pooling is not so much about pooling capacity but about pooling idleness’’ implying that unpooled systems with less idleness can expect less EOS gains when pooled. Next consider that a high value of λ_A/λ_{AB} forces $(1 + C_A^2)/(1 + C_{AB}^2)$ close to 1 diminishing the affect of $(1 + C_A^2)/(1 + C_{AB}^2)$ on Z_A . However, for the corresponding smaller group, this factor becomes increasingly important (see rows 9 and 10 from Table 1).

Finally, Table 2 illustrates the accuracy of approximation (16) by showing the percent by which (16) overestimates Z_A compared with simulated results. Here the simulation results are obtained as described in Section 4 below. As expected, (16) is quite accurate for larger values of ρ_0 and λ_A/λ_{AB} , while for other cases the approximation is poor. Thus, in the next section we obtain an accurate approximation for Z_A in a wide range of realistic scenarios, using computer simulations.

4 Simulation Experiments

To gain further perspective on the factors that influence the loss in EOS and to validate the inferences drawn from (16) a number of numeric experiments are completed.

4.1 Simulation Description

Service Rate Distributions: We model the appointment length as random variables with phase-type distributions [8, 19] where expectation and variance are fitted in the data. We opt for a two moment approximation, instead of a more involved distribution fit (e.g. empirical distribution), because mean and variance data for appointment lengths are typically available. As such it is easily transferable to other settings and the likelihood of implementation is increased [21].

If the appointment length duration has $C \leq 1$ then the appointment length is assumed to follow an Erlang(k, μ) distribution where $\mu = k/D$ and k is the best integer solution to $k = D^2/V$. The completed patients per day (S) is computed by considering that an Erlang(k, μ) distribution is equal to a sum of k independent exponential random variables (phases) with parameter μ and the number of such phases completed in t time units is Poisson with mean μt . It follows that $N(t) = \lfloor \text{Poisson}(\mu t)/k \rfloor$. If $C > 1$ the appointment length is assumed to follow a hyperexponential phase type distribution. The appointment length is distributed according to $p\text{Expo}(\mu_1) + (1-p)\text{Expo}(\mu_2)$ and the total number of complete patients per day (S) is computed by Monte Carlo Simulation where

$$p = \frac{1}{2} \left(1 + \sqrt{\frac{C^2 - 1}{C^2 + 1}} \right), \quad \mu_1 = \frac{2p}{D}, \quad \mu_2 = \frac{2(1-p)}{D}.$$

Patient Mix: The patient mix is described by two factors: λ_A/λ_{AB} , and D_A/D_{AB} . The values for λ_A/λ_{AB} are 0.3, 0.4, 0.5, 0.6, and 0.7. This represents the range of situations where patient group A is 30% [group B is 70%] of the pooled group up to the situation where group A is 70% [group B is 30%] of the pooled group. The values for D_A/D_{AB} are 0.5, 1, 1.5 and 2 representing situations where the appointment length for Group A is half that of the pooled group, and up to and including the case, where it is two-and-a-half times longer. The appointment length of Group B can be computed easily from (3).

Server Allotment: Initially we do not impose restrictions on how to divide the servers between the two unpooled systems as the optimal division follows from the model. To keep the experiments more manageable, results are limited to only “reasonable” room allotments where $|Z_A|$ and $|Z_B| \leq 0.25$. Practically this means we excluded situations where more than a 25% change in appointment length is required to make the performance of the unpooled system equal the performance of the pooled system.

4.2 Results

The results in this section are organized as follows. Initially a Base Clinic is defined and analyzed for the various patient mixes and room allotments. Next the parameters for the pooled clinic

Table 3: Parameters for different Clinic Environment Scenarios

| Clinic Environments | M_{AB} | D_{AB} | λ_{AB} | ρ_0 | C_A, C_B |
|---------------------------------------|-----------|-----------|----------------|-------------|-----------------|
| Base Clinic | 20 | 30 | 282 | 0.88 | 0.5, 0.5 |
| Busier Clinic | 20 | 30 | 310 | 0.97 | 0.5, 0.5 |
| Smaller Clinic | 10 | 30 | 141 | 0.88 | 0.5, 0.5 |
| Shorter Appointment Lengths | 20 | 15 | 564 | 0.88 | 0.5, 0.5 |
| Higher Appointment Length Variability | 20 | 30 | 282 | 0.88 | 2.0, 2.0 |
| Different Coefficient of Variance | 20 | 30 | 282 | 0.88 | 2.0, 0.5 |

are changed representing different clinic environments, e.g. busier clinics, smaller clinics, etc. The results for these different environments are compared to the Base Clinic. The scenarios considered in this section (as listed in Table 3) are meant to encompass a wide range of typical clinic environments. The bold values of Table 3 indicate the parameters which are changed relative to the Base Clinic.

Initial results for managers may come from the clinic environment that most closely reflects their clinic’s make-up. For more specific results, the described simulation (which only requires the mean and variance data) should be used. General management guidelines follow in Section 5.

4.2.1 Base Clinic

The parameters and results for the initial Base Clinic environment are shown in Table 4. The patient mix factors λ_A/λ_{AB} , and D_A/D_{AB} represent the rows and columns respectively. In each table cell, multiple room allotments (represented by the number in parenthesis) and the corresponding Z values are given. The results are in the following format: $Z_A (M_A), Z_B (M_B)$. This represents the amount of change (Z_A) in D_A necessary, when the unpooled clinic is allotted M_A rooms (likewise for patient group B). As an example consider when $\lambda_A/\lambda_{AB} = 0.3$ and $D_A/D_{AB} = 0.5$. The value in the corresponding cell is “-10%(3), -4%(17)”. The result represents the case where 3 rooms are allotted to Group A and 17 to Group B, as noted by the numbers in parentheses. In this case, for the unpooled systems to perform equally as well as the pooled systems, Groups A and B are required to change their appointment length by $Z_A = -10\%$ and $Z_B = -4\%$ respectively. The blank cells in the table are a consequence of excluding room divisions which result in a $|Z|$ value greater than 25%.

From Table 4 and as identified in (16), Z depends on the ratio λ_A/λ_{AB} . When Group A is smaller than Group B (i.e. $\lambda_A/\lambda_{AB} < 0.5$), Group A requires less rooms but a greater decrease in service time. The counter situation (i.e. $\lambda_A/\lambda_{AB} > 0.5$) holds for Group B. It follows that larger patient groups retain EOS and require less EOF to compensate. Furthermore the smallest total loss in EOS (i.e. $Z_A + Z_B$) occurs when the two unpooled departments are the same size. Practically this implies that making a small department to serve a small patient population is not a good idea. This influence of λ_A/λ_{AB} is observable in all tables in this section.

Although not identified by (16), from Table 4 it appears that Z depends on the ratio D_A/D_B . This dependency is not easily characterized as it appears dependent on λ_A/λ_{AB} . Within the range of values tested, the influence of D_A/D_B is small relative to that of λ_A/λ_{AB} . This is observable in all tables in this section except Table 5 where the factor ρ_0 dominates.

Table 4: Base Clinic Results ($M_{AB} = 20$, $D_{AB} = 30$, $\lambda_{AB} = 282$, $C_A = C_B = 0.5$)

| $\frac{\lambda_A}{\lambda_{AB}}$ | $D_A/D_{AB} = 0.5$ | $D_A/D_{AB} = 1.0$ | $D_A/D_{AB} = 1.5$ | $D_A/D_{AB} = 2.0$ |
|----------------------------------|---|---------------------------|--------------------|---------------------|
| 0.3 | -10% (3), -4% (17) | 20% (8), -18% (12) | 10% (11), -21% (9) | |
| | | -12% (6), -4% (14) | 5% (7), -11% (13) | -2% (10), -12% (10) |
| 0.4 | 19% (5), -12% (15) -7% (4), -5% (16) | 16% (10), -21% (10) | 0% (13), -15% (7) | 6% (17), -22% (3) |
| | | -9% (8), -5% (12) | -20% (7), 5% (13) | -9% (12), -4% (8) |
| 0.5 | 17% (6), -12% (14) -4% (5), -7% (15) | 4% (11), -16% (9) | -7% (15), -4% (5) | |
| | | -6% (10), -6% (10) | -16% (9), 5% (11) | -13% (14), 16% (6) |
| 0.6 | 15% (7), -15% (13) -3% (6), -9% (14) -19% (5), -3% (15) | 5% (13), -20% (7) | -5% (18), -6% (2) | |
| | | -5% (12), -8% (8) | -13% (11), 5% (9) | |
| 0.7 | 14% (8), -19% (12) -2% (7), -13% (13) -16% (6), -6% (14) | -4% (14), -11% (6) | | |
| | | -10% (13), 5% (7) | -18% (12), 19% (8) | |

The room allotment which represents the smallest loss in EOS occurs when the difference between ρ_{AB} , ρ_A and ρ_B is minimized. For ease of comparison, the results for these *proportional room distributions* are bolded. For such allotments $\rho_{0,AB} = \rho_{0,A}$ which implies

$$\frac{\lambda_{AB} D_{AB}}{t M_{AB}} = \frac{\lambda_A D_A}{t M_A}$$

$$M_A = \frac{\lambda_A}{\lambda_{AB}} \frac{D_A}{D_{AB}} M_{AB} \quad , \quad M_B = M_{AB} - M_A. \quad (17)$$

Practically speaking this division represents the most equitable way to divide the rooms such that the difference in workload for staff in the two unpooled clinics is minimized. For cases where $C_A = C_B$, it also represents the most equitable way to divide the rooms such that the difference in waiting time for both patient groups is minimized. The high degree by which Z depends on the room division is observable in all the tables in this section.

4.2.2 Busier Clinic

To determine how Z_A and Z_B are influenced by how busy a clinic is, the demand for appointments is increased to $\lambda_{AB} = 310$. Comparing Table 4 with Table 5 it is clear that $|Z_A| + |Z_B|$ is decreasing as the clinic load increases. This means, that the EOS loss of unpooling is smaller for clinics of higher load. This is consistent with the findings from (16). In the remaining scenarios ρ_0 is kept constant with the Base Case.

Table 5: Busier Clinic Results ($M_{AB} = 20$, $D_{AB} = 30$, $\lambda_{AB} = 310$, $C_A = C_B = 0.5$)

| $\frac{\lambda_A}{\lambda_{AB}}$ | $D_A/D_{AB} = 0.5$ | $D_A/D_{AB} = 1.0$ | $D_A/D_{AB} = 1.5$ | $D_A/D_{AB} = 2.0$ |
|----------------------------------|--|--|---|---|
| 0.3 | -4% (3), -3% (17) | 15% (7), -9% (13) -3% (6), -2% (14) -19% (5), 7% (15) | 17% (11), -20% (9) 7% (10), -11% (10) -6% (9), -2% (11) -16% (8), 9% (12) | 1% (13), -15% (7) -8% (12), -3% (8) -15% (11), 12% (9) |
| 0.4 | -3% (4), -3% (16) | 11% (9), -10% (11) -3% (8), -2% (12) -15% (7), 8% (13) | 5% (13), -14% (7) -5% (12), -2% (8) -13% (11), 12% (9) | 2% (16), 6% (4) |
| 0.5 | 18% (6), -12% (14) -3% (5), -6% (15) | 19% (12), -22% (8) 10% (11), -12% (9) -2% (10), -2% (10) -12% (9), 9% (11) -22% (8), 19% (12) | -5% (15), -3% (5) -12% (14), 18% (6) | |
| 0.6 | 16% (7), -13% (13) -3% (6), -6% (14) -19% (5), 2% (15) | 8% (13), -15% (7) -2% (12), -3% (8) -10% (11), 11% (9) | -5% (18), -3% (2) | |
| 0.7 | 14% (8), -15% (12) -2% (7), -9% (13) -16% (6), -2% (14) | 7% (15), -19% (5) -2% (14), -3% (6) -9% (13), 14% (7) | | |

4.2.3 Smaller Clinic and Clinics with Shorter Appointment Lengths

As expected from (16), the results for the clinic with fewer rooms showed only modest changes in Z_A and Z_B and are therefore excluded from the text. However, it is important to note that in smaller clinics, it is more likely that (17) results in a noninteger solution, hence there is a discretization effect. In (16) we assume $\rho_{0,AB} = \rho_{0,A}$ and overlook this influence. The results for a clinic with shorter appointments found Z_A and Z_B to also be insensitive to D_{AB} which is again what is expected from (16).

4.2.4 Higher Appointments Length Variability

Results for a clinic with Higher Appointments Length Variability are available in Table 6. Relative to the Base Case, C_A and C_B were both increased from 0.5 to 2. Contrasting Table 4 and Table 6 it is clear that $|Z_A| + |Z_B|$ has increased considerably with C_A and C_B . Although an increase was expected from (16) the extent of the increase is greater than anticipated. This leads to the conclusion that changes in C_A and C_B have a greater impact than (16) indicates. This is most easily illustrated by considering the patient mix when $\lambda_A/\lambda_{AB} = 0.5$ and $D_A/D_{AB} = 1$ which represents the case where both patient groups have equal service rate and arrival rate parameters. Furthermore, the aggregate service rate for the pooled group also has the same parameters, see (3) and (4). As such, with this patient mix, C_{AB} always equals C_A and likewise C_B . In the simulation experiment for this patient mix, $|Z_A|$ increased by 4% when C_A and C_B were increased from 0.5 to 2. Evaluating (16) for the same situations shows no change in $|Z_A|$, illustrating that (16) does not fully capture the impact of C_A on $|Z_A|$.

Table 6: Higher Appointment Length Variability Results ($M_{AB} = 20$, $D_{AB} = 30$, $\lambda_{AB} = 282$, $C_A = C_B = 2$)

| $\frac{\lambda_A}{\lambda_{AB}}$ | $D_A/D_{AB} = 0.5$ | $D_A/D_{AB} = 1.0$ | $D_A/D_{AB} = 1.5$ | $D_A/D_{AB} = 2.0$ |
|----------------------------------|---|---|--|---|
| 0.3 | 8% (4), -11% (16) -22% (3), -5% (17) | 14% (8), -20% (12) -4% (7), -13% (13) -19% (6), -6% (14) | -6% (10), -17% (10) -17% (9), -7% (11) | -18% (12), -12% (8) |
| 0.4 | 5% (5), -14% (15) -18% (4), -8% (16) | -2% (9), -16% (11) -14% (8), -8% (12) | -13% (12), -11% (8) -21% (11), 3% (9) | -16% (16), -17% (4) -23% (15), 6% (5) |
| 0.5 | 5% (6), -17% (14) -15% (5), -11% (15) | 1% (11), -20% (9) -10% (10), -10% (10) -20% (9), 2% (11) | -11% (15), -15% (5) -16% (14), 5% (6) | |
| 0.6 | 2% (7), -20% (13) -14% (6), -14% (14) | -8% (12), -14% (8) -16% (11), -3% (9) | -9% (18), -22% (2) | |
| 0.7 | -13% (7), -19% (13) | -5% (14), -18% (6) -13% (13), -5% (7) -20% (12), 13% (8) | | |

4.2.5 Different Coefficient of Variance

Results for the scenario when $C_A = 0.5$ and $C_B = 2$ are shown in Table 7. Relative to the Base Case, Z_A decreased and, with few exceptions, Z_B sees almost no changes.

4.3 Conclusions

From the analytic approximation of Z we conclude that when contemplating dividing a pooled department, managers should consider ρ , λ_A/λ_{AB} , and $(1 + C_A^2)/(1 + C_{AB}^2)$. The importance of all three of these factors is confirmed by the simulation experiments, which also identified further factors for consideration. In the simulation experiments we find that Z_A and Z_B values are influenced by C_A and C_B . Z_A and Z_B values also appear slightly sensitive to the ratio D_A/D_B , although characterizing this influence is not observable from the results. Furthermore, with the simulation we identified how the division of rooms between the unpooled departments is also an important decision factor. Finally the simulation also illustrated the discretization effect that occurs in smaller clinics. Both approaches used to quantify the factors impacting the unpooling decisions illustrated that there are numerous considerations necessary and many cannot be considered in isolation. In Table 8 we summarize these factors.

5 Implication for Practice

In general, managers should consider the following when approaching the decision to unpool a centralized department. Under most circumstances access time to clinics will increase unless the service time in the unpooled department is decreased, assuming that no additional resources are made available. The amount of service time decrease needed to compensate for this performance

Table 7: Different Coefficient of Variance Results ($M_{AB} = 20$, $D_{AB} = 30$, $\lambda_{AB} = 282$, $C_A = 0.5$ $C_B = 2$)

| $\frac{\lambda_A}{\lambda_{AB}}$ | $D_A/D_{AB} = 0.5$ | $D_A/D_{AB} = 1.0$ | $D_A/D_{AB} = 1.5$ | $D_A/D_{AB} = 2.0$ |
|----------------------------------|--|---|---|--|
| 0.3 | -5% (3), -5% (17) | 14% (7), -10% (13) | 19% (11), -21% (9) | 5% (13), -15% (7) |
| | | -4% (6), -3% (14) -20% (5), 6% (15) | -4% (9), -2% (11) -14% (8), 9% (12) | -5% (12), -2% (8) -13% (11), 12% (9) |
| 0.4 | -4% (4), -7% (16) | 12% (9), -13% (11) | 9% (13), -16% (7) | |
| | | -2% (8), -4% (12) -14% (7), 6% (13) | 1% (12), -3% (8) -10% (11), 11% (9) | -3% (16), -5% (4) -9% (15), 20% (5) |
| 0.5 | 20% (6), -16% (14) | 12% (11), -16% (9) | | |
| | -2% (5), -9% (15) -21% (4), -3% (16) | 2% (10), -6% (10) -10% (9), 6% (11) -20% (8), 16% (12) | 3% (15), -5% (5) -6% (14), 17% (6) | |
| 0.6 | 17% (7), -20% (13) | 12% (13), -20% (7) | -11% (17), 17% (3) | |
| | 1% (6), -13% (14) -18% (5), -6% (15) | 3% (12), -8% (8) -7% (11), 7% (9) -15% (10), 19% (10) | | |
| 0.7 | 1% (7), -19% (13) | 5% (14), -12% (6) | | |
| | -15% (6), -12% (14) | -5% (13), 6% (7) | | |

loss depends on the characteristics of the original pooled clinic and the characteristics of the newly created unpooled clinics. The main characteristics to consider are clinic load (ρ), number of rooms (N_{AB}), bed division and variability in appointment length. Table 8 summarizes all factors considered in this paper.

When looking at the original pooled clinic consider the following. Clinics under high load require less decrease in service time to compensate for unpooling losses. The number of rooms in a clinic does not greatly influence the needed service time change, however in smaller clinics it is more difficult to proportionally divide the rooms.

When deciding how to split the pooled clinic (which consequently defines the characteristics of the new unpooled clinics) consider the following. The smallest required decrease in service time occurs when the difference between the clinic load in the two unpooled clinics is minimized. To compute the resource allocation that corresponds to this bed division see (17). The smaller patient group resulting from the split will require a greater decrease in service time to compensate for unpooling losses. Finally, unpooling patient groups with highly variable appointment lengths also requires a greater decrease in service time to compensate.

For more specific results refer to the tables in Section 4 or apply the approach described in the same section. The approach used for developing these tables is versatile in terms of the application area and practical in that it requires only typical clinical data as input. The management guidelines are listed in Table 8.

Table 8: Summary of Factors Effecting EOS losses due to Unpooling

| Factors | Change in Z_A | General Management Guidelines |
|---|--|---|
| Clinic Load (ρ_0) | Decreases as ρ_0 increases | Unpooling clinics with high load results in less EOS losses than clinics under lesser load. |
| Room Division | Disproportionate splits increase $ Z_A + Z_B $ | The room allotment representing the smallest loss in EOS occurs when the difference between ρ_{AB} , ρ_A and ρ_B is minimized, see (17). |
| Clinic Size (M_{AB}) | Increases (slightly) as M_{AB} decreases | EOS losses appear mostly insensitive to the size of the clinic. In smaller clinics it is more difficult to proportionally split servers. |
| Clinics with Short Appointment Lengths (D_{AB}) | Mostly insensitive to D_{AB} | EOS losses appear to be mostly insensitive to the length of the appointment. |
| Clinics with Highly Variable Appointment Lengths (C_A, C_B) | Increases as C_A, C_B increases | Unpooling patient groups with highly variable appointment lengths results in larger EOS losses. |
| Clinics with Different Coefficient of Variance for Patient Groups ($C_A < C_B$) | Decreases when $C_A < C_B$ | The patient group with the smaller C generally experiences a smaller loss in EOS as a result of unpooling. |
| Proportional Size of each group (λ_A/λ_{AB}) | Increases as λ_A/λ_{AB} decreases | Smaller patient groups experience a greater loss in EOS as a result of unpooling. |
| Appointment Length Proportion (D_A/D_{AB}) | Mostly insensitive to D_A/D_{AB} | EOS losses appear to be mostly insensitive to the ratio of appointment lengths. |

6 Future Research

The analytic approximation provided initial insight into the influence of the many factors causing losses in EOS, however since it is an approximation it does not fully account for them. The simulation provided more accurate results for a given range of circumstances, and the approach is demonstrated to be robust. However, due to the large number of factors and the complex relationships that exist between them, it proved difficult to use simulation to draw stringent general conclusions. Further research is required to determine how exactly these factors influence losses of EOS related to unpooling. With comprehensive descriptions of these relationships, operational researchers can further improve or even optimize the mix of the functional and patient focused departments within a hospital.

References

- [1] Allen, A.: Probability, Statistics and Queueing Theory. Academic Press, London (1990)
- [2] Ata, B., Van Mieghem, J.: The Value of Partial Resource Pooling: Should a Service Network Be Integrated or Product-Focused? *Management Science* **55**(1), 115 (2009)
- [3] Blake, J., Carter, M., Richardson, S.: An Analysis of Emergency Room Wait Time Issues via Computer Simulation. *INFOR* **34**, 263–273 (1996)

- [4] Cattani, K., Schmidt, G.: The pooling principle. *INFORMS Transactions on Education* **5**(2) (2005)
- [5] Cohen, J.W.: The single server queue, *North-Holland Series in Applied Mathematics and Mechanics*, vol. 8, second edn. North-Holland Publishing Co., Amsterdam (1982)
- [6] van Dijk, N.: On hybrid combination of queueing and simulation. In: *Proceedings of the 2000 Winter simulation Conference*, pp. 147–150 (2000)
- [7] van Dijk, N., van der Sluis, E.: Pooling is not the answer. *European Journal of Operational Research* **197**(1), Pages 415–421 (2009)
- [8] Fackrell, M.: Modelling healthcare systems with phase-type distributions. *Health Care Management Science* (to appear) **12**(1) (2009)
- [9] Hopp, W., Spearman, M.: *Factory physics: foundations of manufacturing management*. McGraw-Hill, Boston (2001)
- [10] Huckman, R., Zinner, D.: Does focus improve operational performance? Lessons from the management of clinical trials. *Strategic Management Journal* **29**(2) (2008)
- [11] Hyer, N., Wemmerlöv, U., Morris, J.: Performance analysis of a focused hospital unit: The case of an integrated trauma center. *Journal of Operations Management* **27**(3), 203–219 (2009)
- [12] Joustra, P.E., van der Sluis, E., van Dijk, N.: To pool or not to pool in hospitals: A theoretical and practical comparison for a radiotherapy outpatient department. In: *Proceedings of the 32nd Meeting of the European Working Group on Operational Research Applied to Health Services* (to appear)
- [13] Kremitske, D., West DJ, J.: Patient-focused primary care: a model. *Hospital Topics* **75**(4), 22 – 28 (1997)
- [14] Langabeer, J., Ozcan, Y.: The economics of cancer care: longitudinal changes in provider efficiency. *Health Care Management Science* (to appear)
- [15] Leung, G.: Hospitals must become Focused Factories. *BMJ: British Medical Journal* **320**(7239), 942 (2000)
- [16] McLaughlin, C., Yang, S., van Dierdonck, R.: Professional service organizations and focus. *Management Science* pp. 1185–1193 (1995)
- [17] Newman, K.: Towards a new health care paradigm. Patient-focused care. The case of Kingston Hospital Trust. *Journal of Management in Medicine* **11**(6), 357–371 (1997)
- [18] Schneider, J., Miller, T., Ohsfeldt, R., Morrissey, M., Zelner, B., Li, P.: The Economics of Specialty Hospitals. *Medical Care Research and Review* **65**(5), 531 (2008)
- [19] Tijms, H.C.: *A First Course in Stochastic Models*. John Wiley and Sons, New York (2003)

- [20] Tiwari, V., Heese, H.: Specialization and competition in healthcare delivery networks. *Health Care Management Science* (to appear)
- [21] Vanberkel, P.T., Boucherie, R.J., Hans, E.W., Hurink, J.L., Litvak, N.: Reallocating Resources to Focused Factories: A Case Study in Chemotherapy. In: J. Blake (ed.) *Proceedings of the 34th Meeting of the European Working Group on Operational Research Applied to Health Services* (to appear)
- [22] Wickramasinghe, N., Bloemendal, J., De Bruin, A., Krabbendam, J.: Enabling innovative healthcare delivery through the use of the focused factory model: the case of the spine clinic of the future. *International Journal of Innovation and Learning* **2**(1), 90–110 (2005)
- [23] Wolstenholme, E.: A patient flow perspective of UK health services: exploring the case for new intermediate care initiatives. *System Dynamics Review* **15**(3) (1999)