

# Designing for Economies of Scale vs. Economies of Focus in Hospital Departments

Peter T. Vanberkel<sup>1</sup>, Richard J. Boucherie<sup>2</sup>, Erwin W. Hans<sup>3</sup>,  
Johann L. Hurink<sup>4</sup>, Nelly Litvak<sup>5</sup>

<sup>1</sup> University of Twente, School of Management and Governance, Operational Methods for Production and Logistics, P.O. Box 217, 7500AE, Enschede, p.t.vanberkel@utwente.nl

<sup>2</sup> University of Twente, Faculty of Electrical Engineering, Mathematics, and Computer Science, Stochastic Operations Research, P.O. Box 217, 7500AE, Enschede, r.j.boucherie@utwente.nl

<sup>3</sup> University of Twente, School of Management and Governance, Operational Methods for Production and Logistics, P.O. Box 217, 7500AE, Enschede, e.w.hans@utwente.nl

<sup>4</sup> University of Twente, Faculty of Electrical Engineering, Mathematics, and Computer Science, Discrete Mathematics and Mathematical Programming, P.O. Box 217, 7500AE, Enschede, j.l.hurink@utwente.nl

<sup>5</sup> University of Twente, Faculty of Electrical Engineering, Mathematics, and Computer Science, Stochastic Operations Research, P.O. Box 217, 7500AE, Enschede, n.litvak@utwente.nl

## ABSTRACT

### Subject/Research problem

Hospitals traditionally segregated 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 are specialized clinics for breast cancer patients and clinical pathways for diabetes patients. Hospitals are grappling more and more with the question, should we become more centralized to achieve economies of scale or more decentralized to achieve economies of focus. In this paper service and patient group characteristics are examined to determine conditions where a centralized model is more efficient and conversely where a decentralized model is more efficient.

### Research Question

When organizing hospital capacity what service and patient group characteristics indicate that efficiency can be gained through economies of scale vs. economies of focus?

### Approach

Using quantitative models from the Queueing Theory and Simulation disciplines the performance of centralized and decentralized hospital clinics are compared. This is done for a variety of services and patient groups.

### Result

The study results in a model measuring the tradeoffs between economies of scale and economies of focus. From this model "rules of thumb" for managers are derived.

### Application

The general results support strategic planning for a new facility at the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital. A model developed during this study is also applied in the Chemotherapy Department of the same hospital.

## 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, thus gaining efficiencies with economies of scale (EOS). On the other hand, some hospitals are becoming more specialized by offering a more limited range of services aiming to breed competence and improve service rates (Leung, 2000). Similar

strategies are also being considered within the organizational level of hospitals, where management grapples with the decision to become more centralized to achieve EOS or more decentralized to achieve economies of focus (EOF). In this paper service and patient group characteristics are examined to determine which model is more efficient. The majority of the algebraic computations is excluded from the text but is available in an extended version of this paper (Vanberkel *et al.*, 2009a).

## **2. THE POOLING PRINCIPLE**

The pooling principle is described in (Cattani and Schmidt, 2005) as “pooling of customer demands, along with pooling of the resources used to fill those demands, may yield operational improvements.” Indeed in the unpooled setting, a customer might be 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. This gain in efficiency is a form of economy of scale. In health care this implies that a centralized (pooled) clinic that serves all patient types may achieve shorter waiting times than a number of decentralized (unpooled) clinics focusing on a more limited range of patient types.

Statistically, the advantage of pooling is credited to the reduction in variability due to the portfolio effect (Hopp and Spearman, 2001). This is easily demonstrated for cases where the characteristics of the unpooled services are identical, see (Joustra *et al.*, 2009, Ata and van Mieghem, 2009) However, pooling is not always of benefit. There can be situations where the pooling of customers actually adds variability to the system thus offsetting any efficiency gains, see (van Dijk and van der Sluis, 2009). Further when the target performances of the customer types differ, than it may be more efficient to use dedicated capacity (i.e. unpooled capacity), see (Joustra *et al.*, 2009). And finally, in the pooled case all servers must be able to accommodate all demand. As a result the service can become more expensive and less efficient as it can no longer focus on a single customer type.

It is clear that pooling is offered as a potential method to improve a systems performance without adding additional resources. Interestingly, the principle of focus in hospitals implies the same (Hyer *et al.*, 2008). In this paper we aim to enhance understanding of these seemingly contradictory view points.

## **3. MODEL**

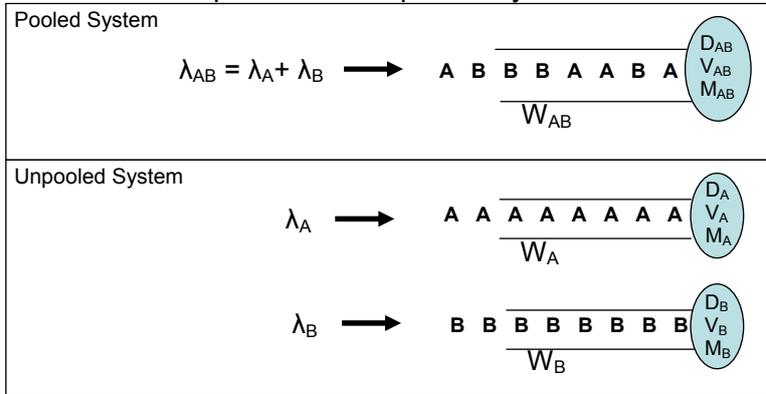
A discrete time slotted queueing model is used to evaluate the tradeoff between EOS and EOF. More specifically, the access time for a centralized ambulatory clinic serving all patient types is compared to the access time of decentralized clinics, focusing on a more limited range of patient types. Generally speaking the decentralized method results in longer access time, due to a loss in EOS. The model quantifies this loss and computes the improvement in service time required in the decentralized method in order to achieve the equivalent access time as in the centralized method. 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 an appointment 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 (e.g. operating room or diagnostic clinics) where

the service time is less than 1 day and where the system empties between days. In this paper the following notation is used:

- $\lambda$  = 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 ( $C= (V/D^2)^{1/2}$ )
- $M$  = Number of rooms
- $\rho$  = The 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 show in *Figure 1*.



*Figure 1: Scheme of the Pooled and Unpooled Systems*

The parameters for two patient groups describe the patient mix. How the patient mix parameters in the unpooled system relate to the parameters in the pooled system, is described below. These division "rules" imply for the unpooled case that no additional resources become available and 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 Subsections 3.1, 3.2 and 3.3 where the subscripts "A", "B" and "AB" are left out for clarity.

### 3.1 Modelling Arrivals and Services

In our model we assume the arrival process is Poisson( $\lambda$ ). If  $X$  denotes the arrivals per day, then  $E[X]=\lambda$ ,  $Var(X)=\lambda$  and  $C_X=1/\lambda$ . Let  $N(t)$  denote the number of appointments completed in one room between  $[0, t]$ . We assume the values of  $N_i(t)$ , where  $i=1, \dots, M$  are independent. Let  $S$  be the total number of completed appointments per clinic day given a clinic has  $M$  rooms, i.e.  $S = N_1(t) + \dots + N_M(t)$ .

Relying only on the mean ( $D$ ) and variance ( $V$ ) of appointment lengths, we use renewal theory to approximate the number of appointments completed in one clinic day. We assume that appointment lengths are i.i.d. and that  $D \ll t$ . In the contrary situation (e.g. chemotherapy, where appointments can last half the day), the following approximations can not be used but the basic approach remains valid, see (Vanberkel *et al.*, 2009b). Under the mentioned assumptions and from renewal theory (Tijms, 2003) we find

$$E[N(t)] \approx \frac{t}{D} + \frac{1}{2}(C^2 - 1) \quad E[S] = ME[N(t)] \approx \frac{Mt}{D} + \frac{M}{2}(C^2 - 1). \quad (5, 6)$$

Although somewhat counterintuitive, note that (6) implies that  $E[S]$  increases as  $C$  increases. 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

$$V_{N(t)} \approx \frac{V^2 t}{D^3} = \frac{C^2 t}{D} \quad V_S \approx MV_{N(t)} = \frac{MC^2 t}{D}. \quad (7, 8)$$

### 3.2 Clinic Load

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

$$\rho \approx \frac{\lambda}{\frac{Mt}{D} + \frac{M}{2}(C^2 - 1)} = \frac{\lambda D}{Mt} \frac{1}{1 + \frac{D}{2t}(C^2 - 1)} \approx \frac{\lambda D}{Mt} = \rho_0. \quad (9)$$

Indeed, the difference between  $\rho$  and  $\rho_0$  is small because (9) is of the order  $D/t$  and we assume  $D \ll t$ . In our simulation experiments of Section 4 we keep  $\rho_0$  fixed for each setup. From (9) we see the actual  $\rho$  changes slightly with the patient mix parameters. For example if  $\lambda_A/\lambda_{AB}$  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  $\rho$ .

### 3.3 Waiting Times

With the above input parameters the expected queue length is computed using Lindley's recursion (Cohen, 1982). Consider subsequent days 1, 2, . . . , 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 be completed on day  $n$ . We assume that  $X_n$  and  $S_n$ ,  $n \geq 1$ , 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

$$L_{n+1} = (L_n + X_n - S_n)^+; \quad n \geq 1 \quad (10)$$

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, namely,  $W=L/\lambda$ . In (Vanberkel *et al.*, 2009b) we explain how to compute the waiting time distribution through a similar recursion. Equation (10) only has an explicit solution in special cases. Therefore in the simulation experiments we solve (10) 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 described by Lindley's recursion. The waiting time of a GI/GI/1 queue can be approximated with Allen-Cunneen approximation (Allen, 1990) thus leading to an approximation for  $L$  in our slotted model. Under the assumption that  $D \ll t$  and by using Little's Law, Allen-Cunneen approximation, (6) and (8) we obtain

$$W = \frac{L}{\lambda} \approx \frac{1}{\lambda} \left\{ \lambda \frac{\rho}{1-\rho} \frac{\left(\frac{1}{\lambda}\right)^2 + C^2}{2} \right\} = \frac{\rho}{2(1-\rho)\lambda} \left( 1 + \frac{C^2}{\rho_0} \right). \quad (11)$$

### 3.4 Required Improvement 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 goal is to determine a new appointment length ( $D_A'$ ) which makes  $W_A = W_{AB}$ . As a standard measure we define  $Z_A$  as the proportional difference between  $D_A$  and  $D_A'$  (likewise for  $Z_B$ ). Ignoring the subscripts "A" and "B" we formally define  $Z$  as follows

$$Z = \begin{cases} -\left(1 - D'/D\right) & \text{when } D'/D < 1 \\ D'/D - 1 & \text{when } D'/D \geq 1. \end{cases} \quad (12)$$

$Z$  essentially measures the EOF needed to make the access time in the pooled and unpooled systems equal. When  $Z$  is negative it represents the amount the appointment length must decrease due to EOF in order to overcome 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 to help illustrate how the tradeoff between EOS and EOF is influenced by the distribution of rooms.

In the simulation experiments,  $Z_A$  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 a same manner. These computations are executed using Microsoft Visual Basic.

Using our estimation (11) for  $W$ , we show how the  $Z$  values can also be estimated. First we assume  $\rho_0 \approx \rho$  and define  $\rho_0'$  as the load in the unpooled clinic A with appointment length  $D_A'$ . It follows:  $\rho_0' = D_A' \lambda_A / M_A t$ . Next we set the waiting time approximations (11) for the pooled and unpooled system A equal to each other

$$\frac{\rho_0'}{2(1-\rho_0')\lambda_A} \left( 1 + \frac{C_A^2}{\rho_0'} \right) = \frac{\rho_0}{2(1-\rho_0)\lambda_{AB}} \left( 1 + \frac{C_{AB}^2}{\rho_0} \right). \quad (13)$$

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 (13) for  $D_A'/D_A$  to obtain

$$\frac{D_A'}{D_A} = 1 - (1-\rho_0) \left( \frac{1+C_A^2}{1+C_{AB}^2} \frac{\lambda_{AB}}{\lambda_A} - 1 \right). \quad (14)$$

Similarly (14) can be rewritten to obtain  $D_B'/D_B$ . From (4) it can be shown that either  $D_A'/D_A$  or  $D_B'/D_B$  in (14) is smaller than 1.

Although several assumptions have been made in the derivation of (14) it does provide insight into the factors effecting  $Z_A$ . The least influential factor is the ratio of the coefficient of variance of the pooled group and the coefficient of variance of the

unpooled group. As the discrepancy between the  $C_A$  and  $C_{AB}$  grows the loss in EOS increases slightly. The smaller  $\lambda_A$  is, relative to  $\lambda_{AB}$ , the greater the loss in EOS. This is demonstrated in a case study in (Vanberkel *et al.*, 2009b). The most influential factor is the clinic's load. The busier the clinic is, the smaller the loss in EOS is. This is consistent with (van Dijk and van der Sluis, 2009), who states "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.

#### 4. SIMULATION EXPERIMENTS

To gain further perspective on the factors that influence the loss in EOS and to validate the inferences drawn from (14) 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 (Tijms, 2003 and Fackrell, 2008) where expectation and variance are fitted in the data. If the appointment length duration has  $C \leq 1$  then the appointment length is assumed to follow an Erlang( $k, \mu$ ) 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( $k, \mu$ ) distribution is equal to a sum of  $k$  independent exponential random variables (phases) with parameter  $\mu$  and the number of such phases completed in  $t$  time units is Poisson with mean  $\mu t$ . 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  $p \cdot \text{Expo}(\mu_1) + (1-p) \cdot \text{Expo}(\mu_2)$  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:  $\lambda_A/\lambda_{AB}$ , and  $D_A/D_{AB}$ . The values for  $\lambda_A/\lambda_{AB}$  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, 2 and 2.5 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 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 are listed in *Table 1*. The shaded cells highlight the parameters which are changed relative to the Base Clinic.

Table 1: Parameters for different Clinic Environment Scenarios

Clinic Environments	$M_{AB}$	$D_{AB}$	$\lambda_{AB}$	$\rho_0$	$C_A, C_B$
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 Appointments Length Variability	20	30	282	0.88	2.0, 2.0
Different Coefficient of Variance	20	30	282	0.88	2.0, 0.5

Initial results for managers can come from for the clinic environment that most closely reflects their clinic's make-up. For more accurate results, the described simulation (which only requires the mean and variance data) should be used. General guidelines follow.

#### 4.2.1 Base Clinic

For initial perspective a Base Clinic environment is evaluated. The parameters and results shown in Table 2. The patient mix factors,  $\lambda_A/\lambda_{AB}$  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, when  $\lambda_A/\lambda_{AB} = 0.3$  and  $D_A/D_{AB} = 0.5$  the corresponding cell contains 1 result. 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, group A and group B are required to change their appointment length by  $Z_A=-10\%$  and  $Z_B=-4\%$  respectively.

Table 2: Base Clinic Results ( $M_{AB} = 20, D_{AB} = 30, \lambda_{AB} = 282, C_A = C_B = 0.5$ )

		$D_A/D_{AB}$				
		0.5	1	1.5	2	2.5
$\lambda_A/\lambda_{AB}$	0.3	-10% (3), -4% (17)	20% (8), -18% (12) 5% (7), -11% (13) <b>-12% (6), -4% (14)</b>	10% (11), -21% (9) -2% (10), -12% (10) <b>-12% (9), -3% (11)</b> -22% (8), 8% (12)	-5% (13), -14% (7) <b>-12% (12), -2% (8)</b> -20% (11), 12% (9)	3% (16), -17% (4) <b>-5% (15), 6% (5)</b>
	0.4	19% (5), -12% (15) <b>-7% (4), -5% (16)</b>	16% (10), -21% (10) 5% (9), -13% (11) <b>-9% (8), -5% (12)</b> -20% (7), 5% (13)	0% (13), -15% (7) <b>-9% (12), -4% (8)</b> -16% (11), 10% (9)	6% (17), -22% (3) <b>-2% (16), 6% (4)</b>	
	0.5	17% (6), -12% (14) <b>-4% (5), -7% (15)</b>	4% (11), -16% (9) <b>-6% (10), -6% (10)</b> -16% (9), 5% (11)	<b>-7% (15), -4% (5)</b> -13% (14), 16% (6)		
	0.6	15% (7), -15% (13) <b>-3% (6), -9% (14)</b> -19% (5), -3% (15)	5% (13), -20% (7) <b>-5% (12), -8% (8)</b> -13% (11), 5% (9) -21% (10), 15% (10)	-5% (18), -6% (2)		
	0.7	14% (8), -19% (12) <b>-2% (7), -13% (13)</b> -16% (6), -6% (14)	<b>-4% (14), -11% (6)</b> -10% (13), 5% (7) -18% (12), 19% (8)			

It is clear from the table that the smallest total loss of EOS ( $|Z_A|+|Z_B|$ ) happens when  $Z_A < 0$  and  $Z_B < 0$ . More specifically, the room allotment which represents the smallest

loss in EOS, occurs when the difference between  $\rho_{AB}$ ,  $\rho_A$  and  $\rho_B$  is minimized. For ease of comparison, the results for these *proportional room distributions* are bolded. For such allotments we have  $\rho_{0,AB} = \rho_{0,A}$  which implies

$$\frac{\lambda_{AB} D_{AB}}{tM_{AB}} = \frac{\lambda_A D_A}{tM_A}$$

$$M_A = \frac{\lambda_A D_A}{\lambda_{AB} D_{AB}} M_{AB}, \quad M_B = M_{AB} - M_A \quad (15)$$

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.

#### 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 2* with *Table 3* 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 (14). In the remaining scenarios  $\rho_0$  is kept constant with the Base Case.

Table 3: Busier Clinic Results ( $M_{AB} = 20$ ,  $D_{AB} = 30$ ,  $\lambda_{AB} = 310$ ,  $C_A = C_B = 0.5$ )

		$D_A/D_{AB}$				
		0.5	1	1.5	2	2.5
$\lambda_A/\lambda_{AB}$	0.3	<b>-4% (3), -3% (17)</b>	15% (7), -9% (13) <b>-3% (6), -2% (14)</b> -19% (5), 7% (15)	17% (11), -20% (9) 7% (10), -11% (10) <b>-6% (9), -2% (11)</b> -16% (8), 9% (12)	1% (13), -15% (7) <b>-8% (12), -3% (8)</b> -15% (11), 12% (9)	5% (16), -18% (4) <b>-2% (15), 5% (5)</b>
	0.4	<b>-3% (4), -3% (16)</b>	11% (9), -10% (11) <b>-3% (8), -2% (12)</b> -15% (7), 8% (13)	5% (13), -14% (7) <b>-5% (12), -2% (8)</b> -13% (11), 12% (9)	<b>2% (16), 6% (4)</b>	
	0.5	18% (6), -12% (14) <b>-3% (5), -6% (15)</b>	19% (12), -22% (8) 10% (11), -12% (9) <b>-2% (10), -2% (10)</b> -12% (9), 9% (11) -22% (8), 19% (12)	<b>-5% (15), -3% (5)</b> -12% (14), 18% (6)		
	0.6	16% (7), -13% (13) <b>-3% (6), -6% (14)</b> -19% (5), 2% (15)	8% (13), -15% (7) <b>-2% (12), -3% (8)</b> -10% (11), 11% (9)	<b>-5% (18), -3% (2)</b>		
	0.7	14% (8), -15% (12) <b>-2% (7), -9% (13)</b> -16% (6), -2% (14)	7% (15), -19% (5) <b>-2% (14), -3% (6)</b> -9% (13), 14% (7)			

#### 4.2.3 Smaller Clinic and Clinics with Shorter Appointment Lengths

The results for the clinic with fewer rooms showed only modest increases in  $|Z_A|+|Z_B|$  and are therefore excluded from the text. However, it is important to note that in smaller clinics, it is more likely that (15) results in a noninteger solution, hence there is a discretization effect. In (14) we assume  $\rho_{0,AB} = \rho_{0,A}$  and overlook this influence. However in practice, when rooms are distributed disproportionately one unpooled group receives extra capacity at the expense of the other. The results for a clinic with

shorter appointments proved  $Z_A$  and  $Z_B$  to be insensitive to  $D_{AB}$  which is also the case in (15). The results for both scenarios are available in (Vanberkel *et al*, 2009a).

#### 4.2.4 Higher Appointments Length Variability

Results for a clinic with Higher Appointments Length Variability are available in *Table 4*. Relative to the Base Case,  $C_A$  and  $C_B$  were increased from 0.5 to 2. Contrasting *Table 2* and *Table 4* it is clear that  $|Z_A|+|Z_B|$  has increased considerably with  $C_A$  and  $C_B$ . Although an increase was expected from (14) the extent of the increase is greater than anticipated. This leads us to the conclusion that changes in  $C_A$  and  $C_B$  have a greater impact than (14) 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 (14) for the same situations shows no change in  $|Z_A|$ , illustrating that (14) does not fully capture the impact of  $C_A$  on  $|Z_A|$ .

*Table 4: Higher Appointments Length Variability Results ( $M_{AB} = 20$ ,  $D_{AB} = 30$ ,  $\lambda_{AB} = 282$ ,  $C_A = C_B = 2$ )*

		$D_A/D_{AB}$				
		0.5	1	1.5	2	2.5
$\lambda_A/\lambda_{AB}$	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 5*. Relative to the Base Case,  $|Z_A|$  decreased and, with few exceptions,  $Z_B$  seen almost no change.

### 5. MANAGEMENT GUIDELINES

From the analytic approximation and the simulation experiments we find the most influential factors effecting efficiency loss due to unpooling are, 1) the clinics load ( $\rho_0$ ), 2) the proportional size of each group ( $\lambda_A/\lambda_{AB}$ ,  $\lambda_B/\lambda_{AB}$ ) and 3) the coefficient of variance ( $C_A$  and  $C_B$ ) for the patient groups. In *Table 6* the all factors considered are listed, and general rules of thumb are provided. Note, a smaller  $|Z_A|$  value means a smaller loss in EOS (likewise for  $|Z_B|$ ). Consequently this means only a small gain in EOF is required to make the unpooled and pooled system perform equivalently.

Table 5: Different Coefficient of Variance Results ( $M_{AB} = 20$ ,  $D_{AB} = 30$ ,  $\lambda_{AB} = 282$ ,  $C_A = 0.5$ ,  $C_B = 2$ )

		$D_A/D_{AB}$				
		0.5	1	1.5	2	2.5
$\lambda_A/\lambda_{AB}$	0.3	-5% (3), -5% (17)	14% (7), -10% (13) -4% (6), -3% (14) -20% (5), 6% (15)	19% (11), -21% (9) 8% (10), -11% (10) -4% (9), -2% (11) -14% (8), 9% (12)	5% (13), -15% (7) -5% (12), -2% (8) -13% (11), 12% (9)	-6% (15), -4% (5) -12% (14), 18% (6)
	0.4	-4% (4), -7% (16)	12% (9), -13% (11) -2% (8), -4% (12) -14% (7), 6% (13)	9% (13), -16% (7) 1% (12), -3% (8) -10% (11), 11% (9)	-3% (16), -5% (4) -9% (15), 20% (5)	
	0.5	20% (6), -16% (14) -2% (5), -9% (15) -21% (4), -3% (16)	12% (11), -16% (9) 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) 1% (6), -13% (14) -18% (5), -6% (15)	12% (13), -20% (7) 3% (12), -8% (8) -7% (11), 7% (9) -15% (10), 19% (10)	-11% (17), 17% (3)		
	0.7	1% (7), -19% (13) -15% (6), -12% (14)	5% (14), -12% (6) -5% (13), 6% (7)			

Table 6: Management Summary of Factors Effecting EOS losses due to Unpooling

	Factors	Change in $ Z_A $	Change in $ Z_B $	General Rules of Thumb
Clinic Environment	Clinic Load ( $\rho_0$ )	Decreases as $\rho_0$ increases	Decreases as $\rho_0$ increases	Unpooling clinics with high load results in less EOS losses than clinics under lesser load.
	Clinic Size ( $M_{AB}$ )	Increases (slightly) as $M_{AB}$ decreases	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 $D_{AB}$	Mostly insensitive to $D_{AB}$	EOS losses appear to be mostly insensitive to the length of the appointment.
	Clinics with Highly Variable Appointment Lengths ( $C_{AB}$ )	Increases as $C_{AB}$ increases	Increases as $C_{AB}$ 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 $C_A < C_B$	Mostly insensitive when $C_A < C_B$	The patient group with the smaller C generally experiences a smaller loss in EOS as a result of unpooling.
Patient Mix	Proportional Size of each group ( $\lambda_A/\lambda_{AB}$ )	Increases as $\lambda_A/\lambda_{AB}$ decreases	Decreases as $\lambda_A/\lambda_{AB}$ decreases	The smaller patient group generally experiences a greater loss in EOS as a result of unpooling.
	Appointment Length Proportion ( $D_A/D_{AB}$ )	Mostly insensitive to $D_A/D_{AB}$	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 it could not be fine-tuned enough to fully account for them. The simulation provided more accurate results but only for a limited range of circumstances. Furthermore 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 hone in on exactly how 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 centralized and decentralized hospital departments.

## REFERENCES

- Allen, A. (1990). *Probability, Statistics and Queueing Theory*. Academic Press, London.
- Ata, B. and Van Mieghem, J. (2009). The Value of Partial Resource Pooling: Should a Service Network Be Integrated or Product-Focused? *Management Science*, 55,1, 115.
- Cattani, K. and Schmidt, G. (2005). The pooling principle. *INFORMS Transactions on Education*, 5, 2.
- Cohen, J. W. (1982). The single server queue, volume 8 of North-Holland Series in Applied Mathematics and Mechanics. North-Holland Publishing Co., Amsterdam, second edition.
- Fackrell, M. (2008). Modelling healthcare systems with phase-type distributions. *Health Care Management Science* (to appear).
- Hopp, W. and Spearman, M. (2001). *Factory physics: foundations of manufacturing management*. McGraw-Hill, Boston.
- Hyer, N., Wemmerlöv, U., and Morris, J. (2008). Performance analysis of a focused hospital unit: The case of an integrated trauma center. *Journal of Operations Management*, 27, 3, 203-219.
- Joustra, P. E., van der Sluis, E., and van Dijk, N. (2009). 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*.
- Leung, G. (2000). Hospitals must become "focused factories". *British Medical Journal*, 320, 7239, 942.
- Tijms, H. C. (2003). *A First Course in Stochastic Models*. John Wiley and Sons, NY, New York.
- van Dijk, N. and van der Sluis, E. (2009). Pooling is not the answer. *European Journal of Operational Research*, 197, 1, 415-421.
- Vanberkel, P. T., Boucherie, R. J., Hans, E. W., Hurink, J. L., and Litvak, N. (2009a). Methods for Designing Hospital Departments to Achieve Economies of Scale vs. Economies of Focus. *Technical Report*, University of Twente.
- Vanberkel, P. T., Boucherie, R. J., Hans, E. W., Hurink, J. L., and Litvak, N. (2009b). Reallocating Resources to Focused Factories: A Case Study in Chemotherapy. In *Proceedings of the 34th Meeting of the European Working Group on Operational Research Applied to Health Services*.