Chapter 2
Capacity Planning

Martin Utley and Dave Worthington

Abstract In this chapter, we discuss some of the modeling methods available for use by health care organizations in determining the level of resources to make available for the delivery of a service or a set of services. We focus mainly on the insights available from different forms of queuing model but discuss also the role of simulation modeling. We then outline the challenges faced in populating capacity planning models with appropriate parameter estimates before closing with some remarks on the non-technical, cultural barriers to effective capacity planning.

2.1 Introduction

Scheduling in health care is often concerned with matching the demand for a service to the resources available to provide that service. At its simplest, this might involve scheduling for an outpatient clinic's patients who have identical and wholly predictable needs that can be met by a single clinician. Toward the other end of the spectrum is the composition of surgical theatre lists that balance the clinical urgency of the candidate patients and administrative wait time targets, account for heterogeneous, stochastic operating times and the prospect of emergency cases and enable targets to be met regarding theatre utilization, overtime and elective cancelation rates.

M. Utley (✉)
Clinical Operational Research Unit, University College London, London, UK
e-mail: m.utley@ucl.ac.uk

D. Worthington
Department of Management Science, Lancaster University Management School, Lancashire, UK
e-mail: d.worthington@lancaster.ac.uk

R. Hall (ed.), *Handbook of Healthcare System Scheduling*,
International Series in Operations Research & Management Science 168,
DOI: 10.1007/978-1-4614-1734-7_2, © Springer Science+Business Media, LLC 2012

Other chapters in this book set out Operations Research (OR) methods that can be used to foster efficient scheduling in health care, a major intellectual and organizational challenge and an area where OR has a considerable amount to offer. The pressure for efficient scheduling within health care stems from the fact that, in virtually all health care contexts, there are limits on the resources available to deliver services. Where resources are abundant to the point of excess, scheduling does not arise as a problem. Also, where resources are not at all adequate, the benefits of efficient scheduling can be marginal—system performance being determined by the lack of resources, however efficiently they are used. Given these considerations, it is fair to say that efficient and effective scheduling becomes a priority once there is a reasonable balance between the demands on a service and the resources available.

In this chapter, we focus on modeling methods for capacity planning, the process by which organizations determine the broad level of resources they make available for the delivery of a service or a set of services. Specifically we define capacity planning to be "deciding on the amount of beds, staff, consulting rooms, equipment, etc. sufficient to enable an organization to meet demand for one or more packages of care while achieving specified service standards".

Section 2.2 considers modeling approaches based on estimating the "unfettered demand" for resources associated with delivering a service, by which we mean the amount of each resource that would be used in a specified period of time if there were no constraints on any resource at any point in the system. Typically, this demand is the result of stochastic processes and, while the expected demand is a useful quantity for organizations to know, the intrinsic variability in this quantity is also of considerable importance when planning capacity.

Estimating unfettered demand is often amenable to intuitive, analytical techniques which can be quite insightful, and this aspect is emphasized. We introduce the notion of reserve capacity and the tension that variability introduces between the efficient use of resources and service standards based on accommodating a target proportion of requests for service. This leads to the central notion that, to estimate capacity requirements, you need to estimate or know demand, variability in demand and the desired service standards.

While modeling unfettered demand gives useful insights about capacity requirements, it cannot explicitly account for the impact that having finite resources available will have on the performance of a system. In Sect. 2.3 we introduce models designed to incorporate finite resource levels and their consequences more directly. Here there is scope for generating insights using analytical queueing models, but often simulation models are required for a detailed picture.

In Sect. 2.4 we discuss some of the issues that confront modelers and organizations when trying to populate capacity planning models with parameter estimates, with particular focus on the use of historical data. In Sect. 2.5 we give some closing remarks, touching on the potential role of simple "rule of thumb" approaches to capacity planning and some of the cultural and managerial challenges to ensuring that the benefits of careful capacity planning are realized. We do

not attempt to provide an expert view here, but rather to flag to modelers some of the non-technical issues to be aware of.

2.2 Estimating Unfettered Demand Using Analytical Models

In Sect. 2.2.2, we consider capacity requirements in terms of the number of staffed beds that should be made available for the delivery of an inpatient service. Of course for some health services, capacity might be considered primarily in terms of the theatre time allocated to a specialty, the number of appointments a clinic offers or the number of CT-scanners available. Modeling the requirements for these and other key resources can be approached by applying similar techniques to those outlined here. Also, in many cases, the models outlined in Sect. 2.2.2 can be augmented to give a way of simultaneously estimating the capacity requirements for a range of different resources associated with the same service. First, we consider briefly the problem of estimating the overall number of patients to anticipate.

2.2.1 Estimating the Number of Patients to Cater For

How to model the number of patients that a service should be designed to cater for depends on the nature of the service and the nature of the health economy. For conditions such as cancer or heart disease and in societies that aspire to universal access to care for such conditions, models to forecast future patient numbers may be based on demographic projections for a particular catchment area, allied with fixed or projected age-specific incidence and prevalence rates. The number of patients to cater for may also be based on considerations of what makes a financially viable or successful service and also, in some settings, considerations of market share. Whatever benefits accrue when different institutions compete (either explicitly or implicitly) for patients, competition introduces additional uncertainty in terms of patient numbers. In general, additional uncertainty increases capacity requirements across a health economy, particularly in systems looking to guarantee patient choice over provider. This is not a problem, so long as it is recognized that many health care costs are driven by having the capacity in place to treat patients (staff, equipment, etc.) rather than by actually treating them.

In a context where plans are being made to increase access to a service that is not widely available (as has been the case for, say, the expansion of bariatric surgery services in some countries) it may not be feasible to attempt to meet the full demand for the service. Here, the number of patients to cater for may be driven and limited by the number of specialist staff available. In this context, capacity planning may be concerned with identifying the level of other resources to provide in order to ensure that efficient use can be made of scarce specialist skills while meeting service standards for those patients who do gain access to the service.

2.2.2 Demand Associated with Known Level of Patient Numbers

Having made an attempt to estimate the overall number of patients per year (for example) that might be expected to access or seek to access a service, there comes the task of translating this into the day-to-day or shift-to-shift demand for key resources. In this section we illustrate the extent to which variability in arrivals and variability in the resources used in the management of individual patients influence the capacity required to cater for a given number of patients, ensuring that a given proportion of requests for service can be met (other service standards are considered in later sections).

Starting from the simplest case, we present models that incorporate different sources of uncertainty and variability, using the illustrative context of an intensive care unit that admits patients following surgery. These models are discrete in time in that arrivals to the unit all happen at integer time units (assumed here to be days but potentially staff-shifts or hours) and patients stay on the unit for an integer number of time units prior to discharge.

Model 1. Suppose first that, each and every day, N patients are operated upon and each requires a bed in the intensive care unit for exactly L days. The demand for beds in this instance will be constant and equal to $N \times L$. Supplying $N \times L$ staffed beds will allow all demand to be met with certainty. Supplying less than $N \times L$ staffed beds would render the planned surgical program infeasible.

Model 2. If, as is typical, length of stay is variable, there will be inevitable variation in demand. Gallivan et al. (2002) developed a simple model to illustrate the variation in demand introduced by variation in length of stay. Suppose that N patients are admitted on a particular day and that the probability of a patient remaining on the unit for at least i days is p_i. The distribution of bed demand from this tranche of arrivals, j days after they arrived, is given by the binomial distribution with N trials each of probability p_i. Extending this, if N such patients are admitted each and every day the probability of n beds being required on a given day is given by the coefficient of s^n in the polynomial.

$$Q(s) = \prod_{i=0}^{\infty} [(1 - p_i) + p_i s]^N \tag{2.1}$$

where s is a dummy variable between 0 and 1. The product in Eq. 2.1 is over a series of binomial expansions, each of which represents the bed demand associated with a single day's admissions.

Figure 2.1 shows the impact of a degree of variability in length of stay (top left hand corner of figure) on the distribution of demand (bottom right). Despite the fact that, in this example, over 80% of patients stay just one day, the long tail on the length of stay distribution results in demand occasionally being three or more staffed beds above the average demand of 8–9. In this instance, the model can be used to estimate that, in order for capacity to be sufficient to meet demand on 95%

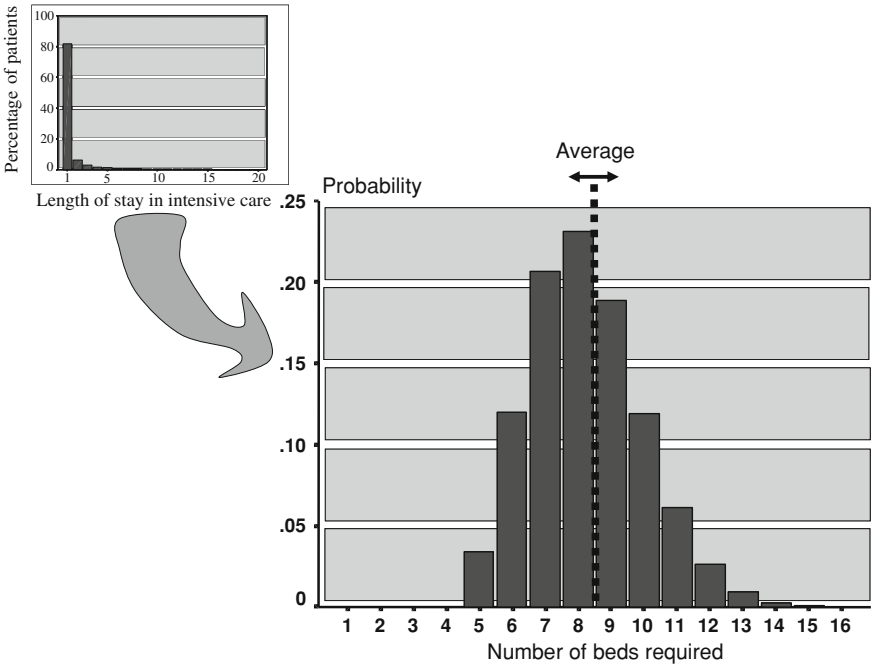

Fig. 2.1 The impact of variability in length of stay on bed demand

of days, the capacity provided should be about 30% above the expected daily demand. This illustrates the main use of models of unfettered demand: that of gauging the proportion of time that a system would be overloaded if capacity was in fact finite and set to a particular value.

Model 3. The model described above can be extended to incorporate numerous additional sources of variability and complexity (Utley et al. 2003). Suppose that any given surgical patient does not attend for surgery with probability v . Suppose further that the number of planned surgical patients is not constant but, on the dth day of the period of interest, is given by N_d. Suppose further still that, in addition to the post-operative patients, there is a probability q_{ed} of there being e emergency admissions to the unit on the dth day of the period of interest.

In this case, the probability of n beds being required on the dth day of operation is given by the coefficient of s^n in the polynomial.

$$Q_d(s) = \prod_{i=0}^{d} \left[[v + (1-v)((1-p_i) + p_i s)]^{N_{(d-i)}} \sum_{e=0}^{\infty} q_{e(d-i)} ((1-p_i) + p_i s)^e \right] \tag{2.2}$$

Again, simple binomial expansions form the building blocks of this expression. Equation 2.2 gives the transient result for the full distribution of demand for a unit

that is empty prior to the first day of operation. With the focus here on strategic capacity planning, it is worth considering demand after some period M that represents a maximum length of stay, before which any steady-state behavior will not emerge. Also, it can be useful to work with the expected demand and its variance rather than the full distribution. Obtaining expressions for these quantities is relatively straightforward, involving the application of standard results concerning generating functions. If the expectation and variance for the number of emergency admissions on day d are given by $E(e_d)$ and $\mathrm{Var}(e_d)$, the expectation and variance of the total demand on day d, t_d say, are given by

$$E(t_d) = \sum_{i=0}^{M} (N_{(d-i)}(1-v) + E(e_{(d-i)}))p_i \quad d \geq \mathrm{M} \tag{2.3}$$

and

$$\begin{aligned} \mathrm{Var}(t_d) = \sum_{i=0}^{M} \big[& (N_{(d-i)}(1-v) + E(e_{(d-i)}))p_i(1-p_i) + p_i^2(N_{(d-i)}v(1-v) \\ & + \mathrm{Var}(e_{(d-i)})) \big] \quad d \geq \mathrm{M}. \end{aligned} \tag{2.4}$$

These expressions provide a means of exploring the impact of different sources of variability on the variations in day-to-day demand and, hence, the capacity required to meet demand to a certain standard.

It is possible within such a modeling framework to estimate capacity requirements for a system incorporating different streams of patients with different length of stay and attendance characteristics and cyclic or otherwise varying patterns of planned and emergency admissions (Utley et al. 2003; Gallivan and Utley 2005). As an aside, the linear form of Eq. 2.2 in terms of the number of patients planned for admission on a particular day introduces the potential for using integer programming techniques to determine schedules of cyclic planned admission that smooth out demand over the cycle (Gallivan and Utley 2005).

In terms of strategic capacity planning, the key insight obtained from such models is that the greater the variability in a system, the greater the capacity required to meet a given service standard on availability or timely access to the service. Another insight is that, given that the variance in demand is (given reasonable assumptions around the variance of emergency admissions) linear in terms of overall patient numbers, there are economies of scale in capacity planning. Essentially, the reserve capacity (as expressed as a proportion of expected demand) required to meet a service standard diminishes with increasing mean demand. In designing services, there can be a tension here between ring-fencing capacity for particular patient groups (to reduce variability) and the economies of scale suggested by having larger pools of capacity. These models can be used effectively to explore the potential trade-offs (in terms of capacity requirements) between smaller pools of capacity dedicated to more homogeneous patient groups versus larger pools of capacity with more heterogeneous case-mix (see for example Utley et al. 2008).

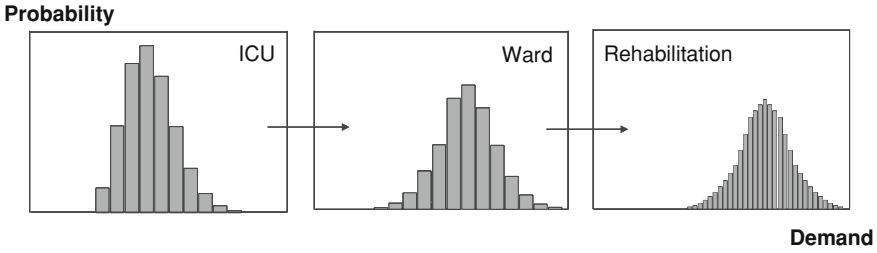

Fig. 2.2 Estimating demand across a number of clinical environments

Similar approaches can also be developed to estimate unfettered demand along a chain of clinical environments or, more generically, for the occupancy of states within any unfettered multi-state system that can be characterized with a simple tree structure (Utley et al. 2009). Such models can be used in capacity planning across environments (either within a single institution or across different institutions), for example to balance provision of acute and rehabilitation services as illustrated in Fig. 2.2.

The appeal of using estimates of unfettered demand in capacity planning lies in the simplicity of the approach and the ability to obtain an impression of the capacity required in order for a system to operate reasonably smoothly. In the next section we examine some of the intrinsic limitations that come with this approach and outline some of the other modeling approaches available.

2.3 Capacity Planning Using Queueing Models

2.3.1 Preamble

The models described in Sect. 2.2 can be used to estimate how often additional resources, above some notional level, would be called upon to meet all demand, or conversely how much resource to provide in order for provision to be adequate roughly 80, 90 or 95% of the time.

At a more detailed level of capacity planning we may need to consider how in practice any demand exceeding the agreed level of capacity would be managed. Is there some spare resource from a neighboring service or from a staffing agency that can be called upon when required? Does the excess demand simply get turned away, for instance referrals for intensive care diverted to another provider? If so, simple modifications of the previous models are possible to improve their accuracy. Does excess demand build up in a queue as with hospital waiting lists or outpatient waiting rooms? If so queueing models which reflect this aspect can be used to analyze and understand the impact of capacity constraints on system performance.

Two major types of queueing models are described in this section, which we will refer to as (i) analytical and (ii) simulation models. Analytical models are typically represented by formulae, which means that they are usually easy to apply (e.g. using a spreadsheet), and often provide valuable insights. Simulation models often require specialist software and are designed to allow the generation and evaluation of 'what if …' scenarios and produce quasi-empirical results rather than direct insights. However they are much more adaptable than analytical models and hence, with sufficient work, can be used to produce more accurate predictions of system behavior. The main components and basic behavior of queueing models are introduced in Sect. 2.3.2. Analytical queueing models together with capacity planning insights that they provide are described in Sect. 2.3.3. Section 2.3.4 outlines the potential value of simulation-based queueing models in this context.

2.3.2 Queueing Models in General

Before introducing analytical and simulation queueing models further we start by considering their common characteristics. The key characteristics of a queueing system are a 'service' that takes a period of time to deliver, the 'servers' who deliver the service and 'customers' who demand the service. When customers arrive to find all the servers busy they wait in a queue until a server becomes available. In a health care capacity planning context there are many examples of queueing systems, including outpatient clinics, emergency rooms, waiting lists for elective surgery, telephone advice lines, emergency ambulances, etc.

While in these situations models of unfettered demand give a good starting point for capacity planning, it is often necessary to consider the impact of queueing more precisely to characterize the relationship between capacity and system performance, particularly when providing substantial reserve capacity is not an option. The key components of the analysis of queuing systems are the arrival process (of customers), the service process, the number of servers (the capacity) and the queueing regime. We consider each briefly.

Arrival Process

Almost all analytical queueing models and many simulation models assume that arrivals occur 'at random', or equivalently as a 'Poisson process'. This is often a good representation of reality, especially if arrivals result from a large population within which individuals have a small probability of demanding the service under consideration. Exceptions to this sort of arrival process occur when arrivals are controlled in some manner, e.g. they are given appointments or grouped before arrival, which is where scheduling can have a big impact.

Even in such instances, the level of referrals for a service month-to-month can often be treated as the result of a Poisson process.
A Poisson process is defined by its underlying arrival rate. If the underlying rate is constant it is often denoted by λ; when it varies over time with known peaks and troughs it is referred to as time-dependent and is denoted by $\lambda(t)$. Note that even when the arrival rate is constant, the number of arrivals in a fixed period of time varies stochastically. For example, if arrivals occur as a Poisson process with mean rate of 5 per hour, the number of arrivals in one hour will have a Poisson distribution with mean 5.

Servers

In the capacity planning context, the 'servers' will often relate to the resources that are being planned. So, for example, the beds might be the servers when planning the size of an Intensive Care Unit (ICU), or the staff on duty might be the servers in an Accident and Emergency (A&E) department. However, the question should always be asked about whether the key resource has been identified. In the ICU setting the ability to treat patients is often limited by the number of staff available rather than the number of beds, and in A&E the bottleneck might sometimes be the number of cubicles rather than the number of staff. Often it makes sense to model capacity requirements for a number of resources.

Service Process

The service process refers to the length of time that the server needs to deliver the service. In an ICU where beds are the key resource, the patient's length of stay is the service process of interest from the capacity planning perspective. In the A&E setting, where patients are often managed using different resources (e.g. staff) as they progress, there are a number of service processes, e.g. initial assessment, X-ray, blood tests, final assessment and treatment.

Queueing Regime

Once patients have arrived in a queueing system the 'queueing regime' describes the way in which they are selected from the queue. For many capacity planning decisions the queueing regime is immaterial, and it can often be assumed that the regime is simply first-in-first-out (FIFO) which puts no restriction on the choice of modeling method.

In addition to these common components of queueing systems, the behavior that they exhibit also has some common features. For example consider a simple,

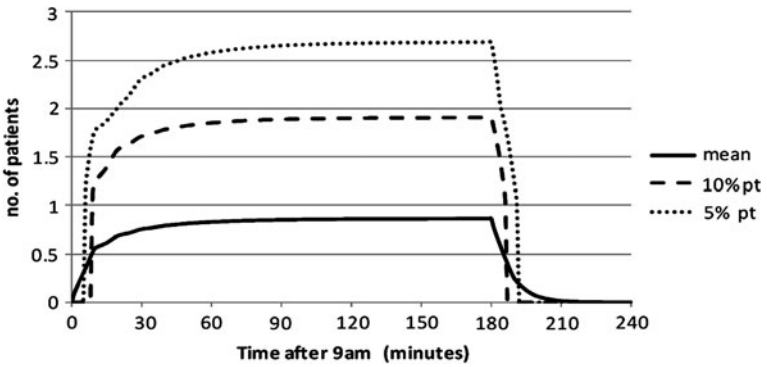

Fig. 2.3 Important features of real queueing systems

single-server outpatient clinic which nominally runs from 9 a.m. to 12 noon and in which 18 patients are given appointments, one every 10 min from 9 a.m. to 11:50 a.m. However, some arrive early, some arrive late and some do not arrive, with the overall effect that they arrive as a Poisson process with an underlying rate of 5 per hour between 9 a.m. and 12 noon. While appointments are every 10 min, consultations are typically shorter, with an average of 6 min. However, the consultation times vary, with a standard deviation of 4 min, so that quite often a consultation takes longer than the allotted 10 min. Using a suitable queueing model (see Brahami and Worthington 1991) for details, Fig. 2.3 graphs the average number of patients in the clinic over the planned duration of the clinic and beyond.

- First we note how the performance is *time-dependent*:
  - Early in the clinic all three performance measures indicate lower levels of congestion than later in the clinic, e.g. after 15 min the mean number of patients in the clinic is 0.61, whereas after 30 min it is 0.75 and after 120 min the mean has grown to 0.85;
  - Also, once patients have stopped arriving, i.e. after 180 min, the levels of congestion then decay away over the next few minutes.
- Between 90 and 180 min the congestion levels stay fairly constant. During this period of time the system has achieved its *steady state*, and indeed it would stay in steady state thereafter for as long as the underlying arrival and service rates do not change.
- However, whether it is in a time-dependent phase or a steady-state phase, the *behavior is stochastic*, i.e. the level of congestion needs to be described by a probability distribution to give understanding to the range of queue lengths, etc., that could be observed at any point of time.

In terms of capacity planning, this form of model could be used to explore the capacity required in the clinic waiting room.

2.3.3 Some Analytical Queueing Models for Capacity Planning

Single-Server Queues

A very useful and insightful queueing model is the Pollaczek-Khintchine formulae which are easy to use formulae for the mean number in the queue $E(q)$ (i.e. excluding anyone in service) and mean number in the system $E(Q)$ (i.e., including anyone in service or in queue) for single-server queues at steady-state in which arrivals occur as a Poisson process at constant rate λ and service times are described solely in terms of their mean $(1/\mu)$ and coefficient of variation (CoV = standard deviation/mean).

$$E(q) = \frac{(\lambda/\mu)^2(1 + \text{CoV}^2)}{2(1 - \lambda/\mu)} \tag{2.5}$$

$$E(Q) = \frac{\lambda}{\mu} + \frac{(\lambda/\mu)^2(1 + \text{CoV}^2)}{2(1 - \lambda/\mu)} \tag{2.6}$$

For example, as the exponential distribution has CoV = 1, if service time is exponential with mean = $1/\mu$, then the equation for $E(Q)$ simplifies to the well known equation:

$$E(Q) = \frac{\lambda/\mu}{(1 - \lambda/\mu)} \tag{2.7}$$

Applying any of these formulae is very straightforward. For example, applying Eq. 2.6 to our outpatient clinic example, where $\lambda = 5$ patients per hour (ph), $\mu = 10$ph and CoV = 4/6 = 2/3, gives:

$$E(Q) = 0.5 + \frac{(0.5)^2(1 + 4/9)}{2(1 - 0.5)} = \frac{31}{36} = 0.861 \tag{2.8}$$

More importantly the formulae also provide important insights into the drivers of congestion, i.e. the drivers of poor performance. Furthermore while these particular formulae are only true for single-server systems, other formulae and further empirical work have shown that the same insights extend to many more queueing situations.

Insights

(i) The arrival rate (λ) and the service rate (μ) only occur as the ratio λ/μ. Thus only the ratio of arrival rate to service capacity is important. This ratio, which is generalized for systems with S servers to $\lambda/S\mu$, is referred to as traffic intensity, and is often denoted by ρ.

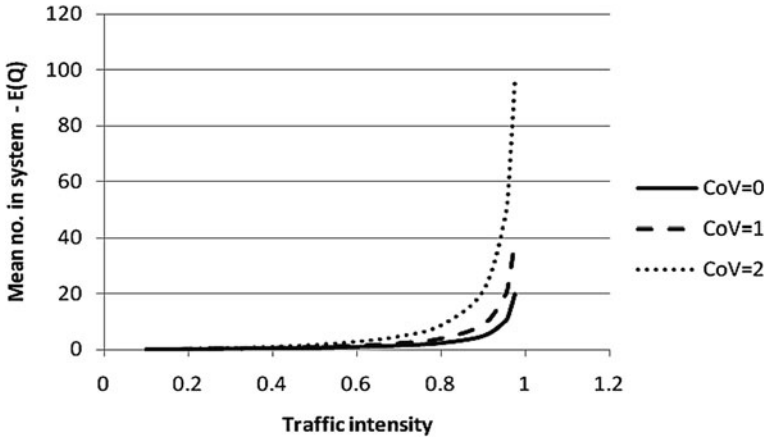

Fig. 2.4 The impact of traffic intensity on congestion

(ii) Simple experimentation with the formulae shows that whatever the value of CoV, the level of congestion grows with ρ, and that it grows increasingly quickly as ρ approaches 1.0; see for example Fig. 2.4 for the cases of CoV = 0, 1 and 2. In terms of capacity planning, this clearly indicates that attempting to achieve high traffic intensity (which equates to high server utilization) is unwise.

(iii) Formula (2.5) also clearly implies (and Fig. 2.4 shows some examples) that for any fixed value of ρ, the level of congestion depends on the CoV of service time, with greater variability (i.e., CoV) leading to higher congestion levels. Further theoretical and empirical work has shown that this insight also applies for multi-server systems.

(iv) A slightly different, but very important, insight that emerges from Eq. 2.5 is that ρ and CoV are the only drivers of this particular measure of congestion. This means that the same result would be obtained for any service time distributions which matched each other in mean and standard deviation. Although this finding is not the case for multiple servers or other performance measures for single-server systems, empirical work shows that service time mean and CoV are key drivers of performance in many other queueing systems.

Multi-server Queues

The main easy-to-use formulae for multi-server queues hold for the steady-state behavior of S server queues with a Poisson arrival process with underlying rate λ, and exponential service times with mean $1/\mu$. In this case the steady-state distribution of number in the system $\{P_n: n = 0, 1, 2,\ldots\}$ is given by:

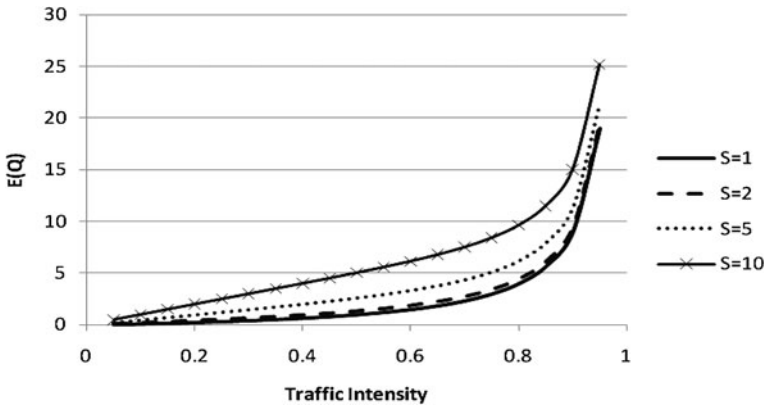

Fig. 2.5 Impact of S and traffic intensity on $E(Q)$

$$P_0 = \left[\sum_{1=0}^{S-1} \frac{(\lambda/\mu)^i}{i!} + \frac{(\lambda/\mu)^S}{S!} \frac{S\mu}{S\mu - \lambda} \right]^{-1};$$

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 \text{ for } n \leq S \\ \dfrac{(\lambda/\mu)^n}{S!\ \ S^{n-S}} P_0 \text{ for } > S \end{cases} \tag{2.9}$$

Furthermore the mean number in the system and the mean number in the queue are given by:

$$E(Q) = {}^{\lambda}/_{\mu} + \frac{(\lambda/\mu)^S \lambda\mu}{(S-1)!\ \ (S\mu - \lambda)^2} P_0 \tag{2.10}$$

$$E(q) = \frac{(\lambda/\mu)^S \lambda\mu}{(S-1)!\ \ (S\mu - \lambda)^2} P_0 \tag{2.11}$$

Insights Continued

(v) Subtraction of formula (2.11) from formula (2.10) shows that the mean number of customers in service, and hence also the mean number of servers who are busy, is simply λ/μ (or $S\rho$); and this result applies to the steady-state behavior of all queueing systems for which λ and μ are well defined.

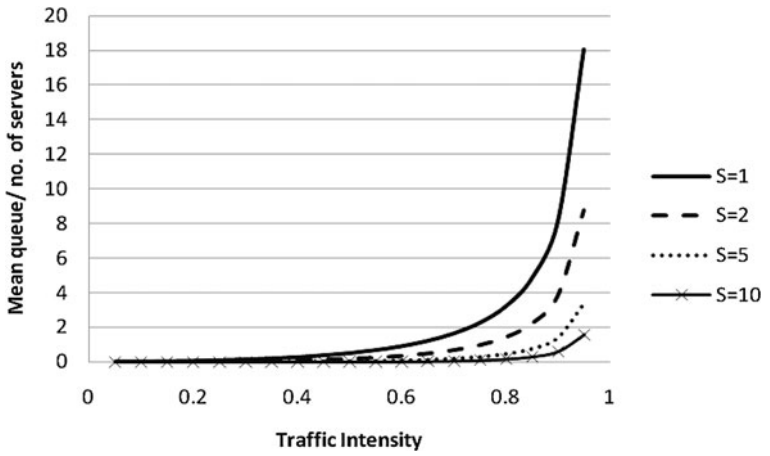

Fig. 2.6 Economies of scale in queueing systems

(vi) Simple experimentation with formula (2.10) supports the insight noted earlier that for any value of S, traffic intensity $\lambda/S\mu$ is a major driver of congestion. See for example Fig. 2.5, which plots E(Q) versus traffic intensity for various values of S.

(vii) Redrawing Fig. 2.5 with its y-axis now showing mean number in the queue per server results in Fig. 2.6, which highlights the sorts of economies of scale that can be achieved by pooling resources of (for example) a number of single-server systems into one multi-server system. We see in Fig. 2.6, for example, when traffic intensity is 0.8 in a single-server system the mean queue length per server is 3.20, whereas in 2, 5 and 10 server systems it is respectively 1.42, 0.44 and 0.16.

Waiting Times and Queueing Times

While number in the system and number in the queue are both important measures of congestion, customers will be much more interested in 'time in the system' (i.e. waiting time) and 'time in the queue' (i.e. queueing time). Except for the special case of exponential service times (see for example Gross and Harris 1985), the distributions of these two performances measures are very difficult to obtain analytically. However, thanks to Little (1961), their mean values ($E(W)$ = mean waiting time, $E(w)$ = mean queueing time) are very easy to obtain. In particular:

$$E(W) = E(Q)/\lambda \tag{2.12}$$

and

$$E(w) = E(W) - 1/\mu \tag{2.13}$$

This very strong relationship to $E(Q)$ means that *all the previous insights* related to $E(Q)$ also carry over to $E(W)$ and $E(w)$.

Networks of Queues

Many real queueing systems involve more than one service operation, and hence can be represented by either a sequence of queues or a network of queues (see Chap. 9). The key analytical model in this area is described in Jackson (1957, 1963), extending works of RRP Jackson (1954). It essentially says that under certain conditions a network of queues can be modeled as if they were independent queues, with arrival rates to each queue calculated from the aggregate arrival rate to that queue from both outside and inside the network. In particular, if a network of queues has:

- K different services (numbered $i = 1, 2,\ldots,K$);
- Customers can also arrive from outside the network to service i, as a Poisson process, at rate λ_i;
- Service i has an exponential distribution, with mean $^1/_{\mu_i}$;
- Service i has S_i servers;
- Having received service i, customers proceed to service j with probability q_{ij}, or leave the network altogether with probability r_i (thus $q_{i1} + q_{i2} + \cdots q_{iK} + r_i = 1$) and these probabilities are independent of the state of the network and the history of the customers.

In this case the overall arrival rate to service i, say α_i, is given formally by:

$$\alpha_i = \lambda_i + \sum_{k=1}^{K} \alpha_k q_{ki} \quad \text{for } i = 1, 2, \ldots, K$$

and steady-state will then exist if $^{\alpha_i}/_{S_i\mu_i} < 1$ for each service i. In this case, steady-state behavior of system i is obtained using the multi-server queueing formulae (2.9), (2.10) and (2.11) presented earlier (with arrival rate $= \alpha_i$, mean service time $= {}^1/_{\mu_i}$, and S_i servers), and it can be treated as if it is independent of the rest of the network. Clearly this analytical model again provides easy-to-use formulae which can be implemented in a spreadsheet, although the nature of the assumptions means that any results will be approximate at best. However these can be obtained quite quickly, can be valuable in producing at least an initial assessment of a problem, and will on occasions be accurate enough for practical purposes.

Furthermore the close links between these network results and the previous single-node results at least suggest that many of the previous insights will carry over to networks of queues.

Time-Dependent Queueing Models

In Sect. 2.2, we outlined a time-dependent model for unfettered demand but there are very few easy-to-use analytical models for the time-dependent behavior of fettered queueing systems despite their obvious importance in practice. Here we introduce one such approach and use it to provide insights into the time-dependent behavior of queueing systems. We also flag up important situations where the approach does not work and discuss how this can impact on the insights.

There are some queueing systems with time-dependent arrival rates $\lambda(t)$ in which the rate of change of arrival rate relative to the throughput of the system is sufficiently slow that the system more or less achieves the steady state associated with any arrival rate instantaneously. In these cases system behavior can be approximated by a pointwise stationary approximation (PSA). This implies that the behavior at any time t_0 is simply approximated by the steady-state behavior of the equivalent constant arrival rate system, with fixed arrival rate $\lambda(t_0)$.

The implication of this approximation is that *whenever the PSA is appropriate*, *all the insights previously identified as associated with steady-state behavior continue to hold.*

However there are many important time-dependent queues in health care for which this approach will not work well, and hence for which the insights will not necessarily hold. For example, for the outpatient clinics modeled previously in Fig. 2.3, the PSA would predict that the mean number of patients in the clinic immediately jump to their steady-state values at 9 a.m., and immediately drop back to zero at 12 noon. Figure 2.3 shows how the queue size can be expected to lag substantially behind the response predicted by the PSA. In cases where it is important to reflect this lagged behavior (for instance in determining staffing requirements over the course of a day in an emergency room), simulation models are often used.

2.3.4 Simulation-Based Queueing Models

As noted earlier, simulation models often require specialist software and are used in 'what if…' mode to produce quasi-empirical results rather than direct insights. However they are much more adaptable than analytical models and hence with sufficient work are capable of producing more accurate results.

One of the earliest queue modeling studies in health care was by Bailey (1952), who used manual simulation experiments to investigate the queueing process occurring in hospital outpatient departments. He came to the conclusion that "a substantial amount of the patients' waiting time may be eliminated without appreciably affecting the consultant."

Using simple simulation analyses of the clinic bottleneck (i.e. the consultation with the doctor) he revealed a number of characteristics of the clinics, including:

- Disproportionate patient waiting time compared to actual consultation time;
- An over-riding consideration to the requirement that the consultant is kept fully occupied;
- A large amount of room (which is often in short supply) just for those waiting.

These insights about the running of outpatient clinics have in fact proven to hold good in many subsequent studies, with simple recommendations such as ensuring that doctors arrive in time for the start of a clinic, and giving patients appointment times that ensure a reasonable balance between the time that patients will wait and the chance that the doctor is idle. See for example Worthington et al. (2005) for one of a number of reviews of such work.

Since that time there have been many studies using simulation-based queueing models to address a wide range of health care capacity planning issues. These have been reviewed on a number of occasions and from various perspectives. Jun et al. (1999) review about 30 years of research, identifying work that focuses on allocation of resources (including bed, room and staff sizing and planning). Fletcher and Worthington (2009) concentrate on simulation modeling of emergency flows of patients, categorizing models under the following areas of hospital activity: A&E, bed management, surgery, intensive care, diagnostics and whole systems models. Günal and Pidd (2010) have a wider remit and use the categories: A&E, inpatients, outpatients, other hospital units and whole systems. Both papers note the very limited amount of work on whole systems models. Brailsford et al. (2009) have an even more ambitious remit and apply a very structured process to review a stratified sample of simulation and other modeling approaches in health care.

The potential of simulation-based models to provide 'what if' analyses of many health care capacity planning issues is clear. However this approach needs to be viewed with care. The record of application of models reviewed in these various surveys is not high, and their flexibility can easily tempt the modeler to pursue an unnecessarily complex model. As suggested by Proudlove et al. (2007) in health care modeling, a combination of simplicity and supportive presentation is more important than aiming at a complex and detailed representation.

2.4 Populating Capacity Planning Models

The models outlined in the previous sections vary in terms of the amount of data or parameter estimates required for their use. These range from having to know mean arrival rates and resource use/service time per patient to knowing or estimating seasonal patterns of arrival, full distributions of service times and resource use and transfer rates between different environments for different sub-groups of the patient population.

For this reason, in addition to the strengths and weaknesses of alternative modeling approaches for capacity planning set out above, the level of information or data available to populate a model is an important consideration in adopting a modeling approach. As demonstrated in Sects. 2.2 and 2.3 it is clear that in

planning capacity for a service, it is important to have estimates for the expected referral or arrival rate, the variability in arrivals, the mean use per patient of the key resources considered and the variability in this use from patient to patient. In instances where a network of environments or services are associated with delivering care for the patient population concerned, the anticipated flows between each service might be required, or at the very least an estimate of how many times (on average) a patient uses each service in the network during an episode of care.

Although not a parameter required in order to run many of the models discussed, an understanding of the desired service standards is essential for using models to inform capacity plans. These might be expressed as maximum cancelation rates for elective patients, or average waiting times or queue sizes for an outpatient clinic.

Typically, those planning a new service will have limited data from which to work and there may be considerable uncertainty in terms of model parameters. In these circumstances, the additional insights of complex models above what can be obtained from simpler models may be limited, or rather the relevance and credibility of these insights may be undermined by doubts as to whether the calibration of the model is sufficiently robust.

Planners looking to estimate future capacity requirements for an existing service are in a better position to estimate parameters but face difficulties of a different nature. The existence of reliable local data concerning the current operation and performance of a system can lead to planners and modelers using these data as direct estimates for model parameters. However, typically an organization will have data on, say, the number of patients admitted to a service rather than the actual demand for admission, since records may not be kept of patients turned away/diverted to another provider or of those who, seeing the queue in a walk-in center, decide not to join the queue. As another example, bed capacity models require estimates concerning clinically necessary length of stay, whereas historical data on length of stay will include any delays to discharging or transferring a patient caused by a shortage of capacity at another point in the system and perhaps instances where discharge was expedited due to pressure to admit an urgent case. Being aware of these potential limitations of historical data is important (particularly if one aim of the capacity planning exercise is to improve rather than simply scale up or scale down a service) since they are often linked to historical capacity provision and the service standards accepted in the past. Essentially, if one is not careful when interpreting data as model input, one risks carrying forward undesirable features of past system performance.

Another issue in calibrating capacity planning models is that changing capacity may influence the demands made of a service. For services where it is known or likely that there is latent, unmet demand or the prospect of genuine supply-induced demand, historical referral or arrival rates should clearly be used with particular care. The influence of increased capacity on referral rates can be modeled explicitly using dynamic models in which referrer behavior is influenced by, for example, waiting times. Another approach is to simply explore scenarios where

referral/arrival rates are increased above and beyond any demographic trends to account for latent demand.

If one of the motivations or likely consequences of the capacity planning exercise concerned is to widen (or restrict) access to a service, it should also be recognized that case-mix (in terms of the severity of patients and resource use per patient) may change.

2.5 Final Thoughts

As indicated in the earlier sections of this chapter, it is possible to build capacity planning models that are highly complex and that incorporate the fine detail of a system as well as the key drivers to system performance. It is perhaps worth reflecting that, although additional complexity and detail in models can add value and in some circumstances will be warranted, there can be diminishing returns in terms of the utility of model output as models get more and more complex, take longer to develop and require more parameter estimates, etc. Indeed, there is an argument that, rather than use or configure explicit capacity planning models each time they undertake a capacity planning exercise, it might sometimes be just as beneficial for organizations to adopt a set of capacity planning rules based on the cumulative experience of capacity modeling of various forms. Such rules might include pooling resources wherever this is consistent with good clinical management, planning on the basis of working at bed utilization of 85%, with lower targets for smaller environments and services where instant access is imperative. Another rule might be to accept lower utilization of relatively cheap resources (for instance portering) where shortages hinder access to or the efficient use of very expensive or scarce resources.

Service configuration or reconfiguration that is informed by the considerations outlined in this chapter should achieve the aim of establishing a reasonable balance between demand and capacity, providing a basis for effective and efficient scheduling to enhance performance.

That said, planning is never enough and the right operational culture is required for the balance of capacity provided across a system to be associated with the performance anticipated from models. Reserve capacity will not have the desired effect on system performance in an organization where the flow of patients is only stimulated by the push of new arrivals at the front door. Proactive discharge planning and a pull through the system must be maintained somehow even when the system seems to be running smoothly. Having appropriate financial arrangements in place helps but ultimately this is a challenge for leadership within health care organizations.

In some health systems, the shift in working culture between the world as it is and the world as Operations Researchers would like it to be is vast and modelers, in their work, should be mindful of the barrier that this represents. The challenge of getting managers and clinicians who have spent their lives working at capacity

to realize that an idle porter, a few empty beds downstream and a nurse spending half a day doing some online training because they are not needed on the ward are good things is one that goes way beyond the technical challenges of building and using models.

References

Bailey NTJ (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. J Roy Stat Soc Ser B 14:185–199

Brahimi M, Worthington D (1991) Queueing models for out-patient appointment systems: a case study. J Oper Res Soc 42:733–746

Brailsford SC, Harper PR, Patel B, Pitt M (2009) An analysis of the academic literature on simulation and modelling in health care. J Simul 3:130–140

Fletcher A, Worthington D (2009) What is a 'generic' hospital model?—A comparison of 'generic' and 'specific' hospital models of emergency flow patients. Health Care Manage Sci 12:374–391

Gallivan S, Utley M (2005) Modelling admissions booking of elective in-patients into a treatment centre. Inst Math Appl J Manage Math 16:305–315

Gallivan S et al (2002) Booked inpatient admissions and hospital capacity: mathematical modelling study. BMJ 324:280–282

Gross D, Harris CM (1985) Fundamentals of queueing theory, 2nd edn. Wiley, New York

Günal MM, Pidd M (2010) Discrete event simulation for performance modelling in health care: a review of the literature. J Simul 4:42–51

Jackson RRP (1954) Queueing systems with phase-type service. Oper Res Quart 5:109–120

Jackson JR (1957) Networks of waiting lines. Oper Res 5:518–521

Jackson JR (1963) Jobshop-like queueing systems. Manage Sci 10:131–142

Jun J, Jacobson S, Swisher J (1999) Application of discrete-event simulation in health care clinics: a survey. J Oper Res Soc 50:109–123

Little JDC (1961) A proof for the queueing formula $L = \lambda W$. Oper Res 9:383–387

Proudlove NC, Black S, Fletcher A (2007) OR and the challenge to improve the NHS: modelling for insight and improvement in in-patient flows. J Oper Res Soc 58:145–158

Utley M, Gallivan S, Treasure T, Valencia O (2003) Analytical methods for calculating the capacity required to operate an effective booked admissions policy for elective inpatient services. Health Care Manage Sci 6:97–104

Utley M, Jit M, Gallivan S (2008) Restructuring routine elective services to reduce overall capacity requirements within a local health economy. Health Care Manage Sci 11:240–247

Worthington DJ, Goulsbra R, Rankin J (2005) Scheduling appointments in outpatient clinics. In: Vissers J, Beech R (eds) Health operations management. Routledge, London, pp 223–248

http://www.springer.com/978-1-4614-1733-0

Handbook of Healthcare System Scheduling
Hall, R. (Ed.)
2012, X, 334 p., Hardcover
ISBN: 978-1-4614-1733-0