A Simulation Optimization Approach for Long-Term Care Capacity Planning

Yue Zhang, Martin L. Puterman
Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, British Columbia, Canada, V6T 1Z2, yue.zhang@sauder.ubc.ca, martin.puterman@sauder.ubc.ca

Matthew Nelson
Centre for Research in Healthcare Engineering, University of Toronto, 5 King’s College Road, Toronto, Ontario, Canada, M5S 3G8, mnelson@mie.utoronto.ca

Derek Atkins
Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, British Columbia, Canada, V6T 1Z2, derek.atkins@sauder.ubc.ca

This paper describes a methodology for setting long-term care capacity levels over a multi-year planning horizon, to achieve target wait time service levels. Our approach integrates demographic and survival analysis, discrete event simulation, and optimization. Based on this methodology, a decision support system was developed for use in practice. We illustrate this approach through two case studies: one for a regional health authority in British Columbia, Canada, and the other for a long-term care facility. We also compare our approach to the fixed ratio approach used in practice and the SIPP (stationary, independent, period by period) approach developed in the call center literature. In addition, a state-dependent model and an adaptive system for capacity planning are considered to achieve better resource utilization. We conclude the paper with policy recommendations.

Key words: long-term care, capacity planning, survival analysis, simulation, optimization, service level

1. Introduction

This paper concerns the planning of long-term care (LTC) capacity for a frail elderly population. We find it is surprising that this issue has received so little attention in the operations literature in spite of the fact that the number of people needing LTC will increase rapidly as a result of advances in medical care. Worldwide, there are 600 million people aged 60 and over, and this number will double by 2025 (World Health Organization 2009). According to Statistics Canada (2009), the 2006 Census shows that seniors aged 65 or over accounted for 13.7% of Canada’s population in 2006; this proportion will start to increase even more rapidly by 2011, when the first wave of baby boomers
born in 1946 reach 65. In British Columbia (BC), Canada, the proportion of the population aged 65 and over will pass 20% in many regions over the next few years; the proportion of the population aged 85 and older is expected to double by 2031 (BC Stats 2008).

Although this population aging is a success of public health policies and socioeconomic development, it is also posing tremendous challenges in providing timely health care for this population. According to the Canadian Medical Association (2009), an estimated 5% of Canadians aged 65 and over live in LTC facilities. These facilities provide health care and support services as well as activities of daily living. As the population ages, the need for LTC will increase.

Moreover, lack of access to LTC is often cited as the major cause of a high level of alternative level of care (ALC) patients, who no longer need acute services but are waiting in acute care to be discharged to a setting more appropriate to their needs. According to Canadian Institute for Health Information (2009), ALC patients accounted for 14% of hospital days in acute care hospitals; 43% of ALC patients were discharged to a LTC facility. This suggests that providing sufficient LTC capacity would have a significant impact on acute care as well as the entire health system. Therefore, there is an urgent need to plan for the LTC capacity that will be required by the elderly population.

The research reported here was the result of two studies we carried out in British Columbia: one for a regional authority and the other for an individual LTC facility where one of the authors (Martin L. Puterman) serves as a member of the board of trustees. The issues in each case were slightly different but in general the question raised was “How many LTC beds are needed over the next 10-20 years to ensure that care is provided in a timely fashion?” Current practice has been to use a fixed ratio of beds per population over age 75 as the basis for planning. This is problematic for several reasons and has resulted in long wait times for admission to care or excess capacity (Cohen et al. 2009). Therefore, development of rigorous mathematical tools for LTC capacity planning is critical.

This research seeks to develop and apply operations research techniques to improve long-term
capacity planning for LTC programs and facilities. Specifically, we propose a simulation optimization approach to determine the minimal capacity level needed each year to satisfy a service level criterion based on clients' wait time. Key elements of the approach include using demographic and survival analysis to predict arrival and length of stay (LOS) distributions for input to the simulation model. Based on the combined approach, a decision support system was developed for managers of LTC programs or facilities to use in practice. This decision support system provides service levels for any capacity levels specified by the user, determines the optimal capacity levels to meet the service level criterion, and enables sensitivity or “What if?” analysis to explore the sensitivity of results to assumptions.

The main methodological contribution of the research lies in the integration of simulation, optimization, and demographic and survival analysis. Through two case studies, we illustrate our approach, compare it to the ratio and SIPP approaches (see below) and derive policy insights. We believe that the decision support system will result in improved access to care, reduced wait times, and optimum capital investment. Also, optimizing the level of LTC capacity should reduce the number of ALC patients in acute care and thus improve entire health system access.

The paper is organized as follows. The next section reviews related literature on capacity planning in health care and other areas. Section 3 describes the problem and the system. The methodology is introduced in Section 4, including the discrete event simulation model, demographic analysis, survival analysis, and two optimization techniques. Section 5 presents two applications of this methodology and compares it to two other approaches. A state-dependent model and an adaptive policy for capacity planning are investigated in Section 6 to achieve better resource utilization. In the final section, concluding remarks and policy recommendations are summarized, and future research directions are discussed.

2. Related Literature

There is little published research on capacity planning for LTC services. As far as we know, the only study addressing this issue is Hare et al. (2009), which presents a deterministic system dynamics
model of the HCC system to predict future client counts. Since their model focusses on strategic provincial level decisions is at a high level of aggregation, no client is individually identifiable and no service level target is considered.

In contrast, several studies focus on capacity planning or utilization analysis for other specific medical services (Ridge et al. 1998, Green 2003, and Harper and Shahani 2002), as well as capacity allocation for various services or departments (Kao and Tung 1981 and Vassilacopoulos 1985). These papers usually do not consider the dynamics of the systems over time. Refer to Smith-Daniels et al. (1988) and Green (2004) for a thorough review of capacity planning in health care.

On the other hand, the problem investigated here is similar to the operator staffing problem in call centers and other multi-server queueing systems with time-varying arrival rates, where the minimum staffing level (number of servers) in each period needs to be determined to ensure a satisfactory service level, usually based on customers’ waiting time. Moreover, other service sectors, where such a staffing problem is encountered, include toll plazas, airport check-in counters, retail check-out counters, banking, telecommunications, and police patrol (Green et al. 2001).

Prior to determining the optimum staffing level, an important issue is to evaluate the performance of the system for a specified staffing level. In the literature of queueing theory, analytical approaches have been used to study non-stationary systems if they are Markovian. For example, by numerically solving the Chapman-Kolmogorov forward equations (i.e., a standard set of differential equations), the steady state probability of the number of customers in the system can be calculated, and then various performance measures can be obtained based on that. See Green et al. (1991), Green and Kolesar (1991), and Ingolfsson et al. (2002). In contrast, many papers also focus on using simple stationary queueing models as approximations to evaluate and manage non-stationary systems, especially for non-Markovian or more general systems. These include the pointwise stationary approximation that uses the instantaneous arrival rate, the simple stationary approximation that uses the long-run average arrival rate, and the infinite-server approximation that estimates the distribution of the number of busy servers with respect to time. See Jennings et al. (1996) and Ingolfsson et al. (2007) for review of these approximation methods.
When the required staffing levels are decision variables, they are typically determined by using available analytical results based on simple stationary queueing models, as in Green et al. (2001). Specifically, the planning horizon is divided into multiple homogeneous periods. Then, a series of stationary queueing models, usually $M/M/s$ queues, are constructed, one model for each period. Each of these models is independently solved for the minimum number of servers needed to meet the service level target in that period. They referred to this method of setting staffing requirements as the SIPP (stationary, independent, period by period) approach. The SIPP approach is closely related to the approximations mentioned above, and it usually results in the form of the “square-root rule” (Atlason et al. 2008). However, the SIPP approach does not always work well. Many papers have compared the achieved performance measures derived from the solutions by the SIPP approach with the ones derived from the exact analytical approaches or simulation, including Kwan et al. (1988), Whitt (1991), and Green et al. (2001). For instance, Green et al. (2001) and Atlason et al. (2008) were mainly concerned with the linkage between staffing decisions in consecutive periods. The latter also mentioned that the SIPP approach does not work well when service times are long relative to the period length.

In addition to the SIPP approach, analytical methods have also been investigated for staffing problems. For example, a function of the staffing level with respect to time can be derived based on the infinite-server approximation in Jennings et al. (1996), when the probability of delay is the service performance measure. More recently, Parlar and Sharafali (2008) proposed an exact analytical approach based on a stochastic dynamic programming model, to determine the optimal number of check-in counters needed for each flight over time to minimize a certain expected cost function. De Vericourt and Jennings (2008) considered a nurse staffing problem by modeling medical units as closed queueing systems. Their results suggest that nurse-to-patient ratio policies cannot achieve consistently high service level. Yankovic and Green (2008) investigated a more complicated nurse staffing problem with taking new arrivals, departures, and transfers of patients into account. Based on their queueing model with a two-dimensional state space, they can evaluate the system...
performance analytically and then choose optimum staffing levels. Again, their analysis shows that nurse-to-patient ratio policies cannot achieve satisfied performance.

Simulation is another methodology used in the literature, especially to study complex non-stationary queueing systems. However, instead of using simulation to optimize staffing, most of these papers for call centers use simulation to evaluate the system performance with the staffing levels identified by approximate analytical approaches, so as to verify whether the suggested staffing levels indeed produce the desired performance. A few exceptions in recent years include Feldman et al. (2008) and Atlason et al. (2008), who used simulation to study their specific staffing problems. The former proposed a flexible simulation-based iterative-staffing algorithm for models with non-homogeneous Poisson arrival process and customer abandonment. They divided the time horizon into many small intervals. Running multiple independent simulation replications, they can estimate the distribution of the total number of customers in the system with respect to time, based on which the staffing function with respect to time can be derived. By generating multiple simulation replications, Atlason et al. (2008) transformed their staffing problem into a deterministic one, which computes the staffing level in each period to ensure that the average service level is satisfied. Both papers use the probability of delay as the service measure, whereas their approaches may be hard to be applied to solve the problems based on other measures.

See Gans et al. (2003) and Green et al. (2007) for further references about operations, performance evaluation, and staffing rules in call centers.

3. Model Description

This paper focuses on LTC capacity planning for an individual facility or a geographic region in aggregate over a multi-year planning horizon. We model the system as a multiple server queue with time varying arrival and service rates. Since the past study on LTC (Hare et al. 2009) and our analyses have shown that people with different ages and genders may have different arrival and LOS distributions, clients are classified into $I$ classes based on age and gender, each with its own arrival and LOS distribution. We assume no constraints on waitlist size, that there are no departures from the waitlist, and that queue discipline is first-come first-served (FCFS).
We consider a $T$ year planning horizon and denote the year index as $t$, $t = 1, \ldots, T$. We further assume that the number of beds can only be changed at the start of each year. Initially, there are a known number of existing clients in each class, who are either in care or in the waitlist. During each year, we assume that the number of arrivals in each class follows a Poisson distribution with a constant rate. We also assume a class-dependent and possibly time varying LOS distribution. The number of beds available in year $t$ is denoted by $s_t$; let $s_0$ denote the initial number of beds. Thus, the system can be modeled as a series of multi-class $M/G/s$ queueing systems as shown in Figure 1.

In this research, we assume that capacity can be changed at the start of each year and by any amount. In practice, this is not feasible. Hence, our results provide decision makers with capacity targets that can serve as inputs to optimal capacity planning decisions which take costs and constraints into account. The proposed methods seek to find values for $s_t$, $t = 1, \ldots, T$, that achieve a pre-specified service level. Service levels may be based on the probability of waiting, the number of people in the system or in the queue, and average wait time. We focus on service levels
for clients' wait time expressed as:

\[
\Pr(W_t(s_t) \leq \gamma) \geq \tau \quad t = 1, \ldots, T, \tag{1}
\]

where \(W_t(s_t)\) denotes the wait time in year \(t\) given \(s_t\), \(\gamma\) denotes a wait time threshold (in days), and \(\tau\) denotes a probability threshold. This means that the probability that an arbitrary client in year \(t\) will be placed in care within \(\gamma\) days is greater than or equal to \(\tau\). In other words, \(\tau \times 100\) percent of arriving clients receive service within \(\gamma\) days each year.

In contrast, the following stronger criterion based on the simultaneous probability over the planning horizon may be preferred:

\[
\Pr(W_1(s_1) \leq \gamma, \ldots, W_T(s_T) \leq \gamma) \geq \tau'. \tag{2}
\]

Nevertheless, using the Bonferroni approach (Miller 1981), expression (2) will hold if:

\[
\Pr(W_t(s_t) \leq \gamma) \geq 1 - \frac{1 - \tau'}{T} \quad t = 1, \ldots, T, \tag{3}
\]

which has exactly the same form as expression (1). Therefore, this approach provides a simple way to deal with more complex service level criteria. Consequently, we focus on expression (1) in this paper.

Let \(s_t^*\) denote the minimum number of beds so that the service level criterion is met in year \(t\). As discussed in Atlason et al. (2008), the relationship between the number of beds and the resulting steady state service level, \(\Pr(W_t(s_t) \leq \gamma)\) for fixed \(\gamma\) and variable \(s_t\), typically follows an “S-shaped” curve, i.e., a convex arc flowing into a concave arc. If an exact closed-form expression for this steady state probability is available, the number of beds required to meet the service level criterion in each year can be directly determined from

\[
s_t^* = \arg\min\{k \in \mathbb{N} : \Pr(W_t(k) \leq \gamma) \geq \tau\} \quad t = 1, \ldots, T. \tag{4}
\]
We note that this is the basis for the SIPP approach (Green et al. 2001). In our context, it would divide the planning horizon into $T$ years and use the closed-form expression which applies to a stationary queue in the steady state to set the capacity in each year.

One significant challenge of using the SIPP approach in our setting is that closed form expressions needed in expression (4) are not available. This is because the system contains several classes of clients with different arrival and LOS distributions. To use the SIPP approach, we could aggregate them in a single class and further assume that the LOS follows an exponential distribution, so that the system over time is modeled by a series of $M/M/s$ queues.

However, it is doubtful whether the SIPP approach provides good approximations in our setting, because several of its implicit assumptions are violated for the reasons below.

1. **Independent Periods**: The system is not empty at the start of each year. LOSs are long relative to period length so that clients may remain in the system for several periods.

2. **Homogeneity of Clients**: Aggregating multiple classes of clients into a single class ignores widely varying client classifications and resource requirements.

3. **Exponentiality**: LOS distributions appear to be better modeled by a fat-tailed Weibull rather than an exponential distribution. Hence, the memoryless property will not hold.

Because of these reasons, we developed a simulation-based optimization approach to determine $s^*_t$ and compared the resulting capacities to those based on the SIPP and ratio approaches.

4. **Methodology**

4.1. **Discrete Event Simulation**

We first describe the discrete event simulation model with a fixed pre-specified capacity sequence $s_t, t = 1, \ldots, T$. The simulation has three main inputs: arrival distributions, LOS distributions, and pre-loaded existing clients. The simulation logic can be summarized as follows.

1. **Initialization**: At the beginning of a planning period, the model pre-loads existing clients representing those in care and in the waitlist. To each, it randomly assigns a remaining LOS based on the appropriate age and gender specific *conditional* LOS distribution, as discussed in Section 4.2.3.
2. **Client Generation**: In each year, arrivals occur at exponentially distributed times with rate appropriate for that year. They are assigned age and gender on the basis of the relative arrival rates as discussed in Section 4.2.1.

3. **LOS Generation**: The model randomly assigns a LOS to each new client based on the age and gender specific LOS distribution determined by methods discussed in Section 4.2.2.

4. **Assignment to Queue and Beds**: Each new client enters the queue and waits until a bed becomes available. Upon receiving a bed, the LOS starts and the client remains in the system for that period. Length of time in the queue is recorded to measure performance.

5. **Annual Summaries and Updating**: At the end of each year, the service level is computed and arrival rates and LOS distribution parameters are updated as necessary.

Given a capacity level in each year, the simulation model provides the resulting service level, i.e., the percentage of clients in each year who are placed in care within $\gamma$ days. Due to variability, we ran multiple simulation replications and calculated the average service level each year. We also considered a percentile of the service level distribution as a performance summary (i.e., the 80th percentile). We later found the average to be more stable than the percentile value and chose it as our measure.

The discrete event simulation by itself cannot determine the appropriate capacity level each year. Hence, we developed and integrated optimization techniques into the discrete event simulation to accomplish this. Then, by running multiple simulation replications, the problem becomes that of determining the minimal capacity level required each year so that on average $\tau \times 100$ percent of clients are placed in care within $\gamma$ days. See Section 4.3 for the description of our approach to optimization.

### 4.2. Simulation Inputs

In this section, we discuss our approach for estimating arrival and LOS distributions as well as pre-loading existing clients into the simulation.
4.2.1. Arrival Analysis  Typically, we can assume that new arrivals follow a Poisson distribution with a constant rate during each year; hence, the time between arrivals follows an exponential distribution. This assumption appears reasonable in that most arrivals to LTC are unscheduled. Alternatively, it can be tested using data and modified if necessary.

Let $\lambda_i(t), i = 1, \ldots, I, t = 1, \ldots, T$, denote the Poisson arrival rate parameter for clients in class $i$ in year $t$. We considered several approaches to estimating it. Primarily, we used the historical per capita arrival rate $\lambda_i$ for class $i$ and multiplied it by a population forecast $N_i(t)$ for class $i$ in year $t$ to obtain,

$$\lambda_i(t) = \lambda_i N_i(t) \quad i = 1, \ldots, I. \quad (5)$$

To estimating $\lambda_i$ requires two data sources: the historical number of clients per year entering care or the waitlist and historical population sizes. The former should be available from long-term care facility data or appropriate regional records. Population data by age, gender, and year is usually available from census or administrative data, such as BC Stats (BC Stats 2008). The simplest estimate of $\lambda_i$ would be obtained for each class $i$ by dividing total arrivals to the LTC facility by the population. This could be refined by using a weighted average of the past several years. Alternatively, expression (5) could be refined to allow $\lambda_i$ to vary with $t$. These refined estimates could be based on rigorous forecasting models or a damping function of the form $e^{-\beta t}$ could multiply $\lambda_i$ to model a decreasing trend that might result from improved population health, enhanced community services or policy changes. The effect of this could be explored through sensitivity analysis.

One challenge in using this approach is that reliable historical waitlist data may not be available. Often, data records indicate when a client entered care but not when he/she entered the waitlist. Some adjustment is necessary depending on what data is available. For instance, we may seek historical waitlist information or have to assume that the waitlist size remains constant. In the worst case, we would have to make an explicit assumption about the change in waitlist size.
In the simulation, two approaches are available for generating arrivals. Either the time between arrivals of class $i$ follows an exponential distribution with rate $\lambda_i$, or instead, clients are generated according to exponential distribution with rate $\lambda(t) = \sum_{i \in I} \lambda_i(t)$ and assigned to class $i$ with probability $\lambda_i(t)/\lambda(t)$. Due to the separation property of Poisson distributions, the arrival process for each class still follows a Poisson distribution.

**4.2.2. LOS Analysis** In most settings, LOS distributions for clients in care can be estimated from historical data for clients who have exited care. However, this ignores the effect of active clients (those who have entered care and are still in care at the end of the most recent fiscal year). If there is a considerable amount of historical data and it is believed that LOS distributions have been constant over time, this would be valid. But if LOS distributions are changing over time or there is not a lot of data, data on clients still in care must be taken into account. Unfortunately, we will not have full LOS information about these active clients, only lower bounds (start dates). Data of this type is said to be *right censored*. We observed that ignoring censored data leads to underestimation of LOS, since the censored clients tend to be those with longer LOSs.

The statistical field of survival analysis (Klein and Moeschberger 2003) addresses the problem of modeling of time-to-event data, when a portion of data is censored. We used parametric models to estimate LOS distributions. These enable us to generate LOS in the simulation model. We recommend using a Weibull distribution, because it is determined by two parameters, can be tuned to have several shapes, contains an exponential distribution as a special case, and can represent data with long tails. Its adequacy can be investigated through Q-Q plots or formal goodness-of-fit tests. The cumulative distribution function (CDF) of a Weibull distribution is given by:

$$F(x) = 1 - e^{-(x/\alpha)^\beta}, \quad (6)$$

where $\alpha$ is the scale parameter and $\beta$ is the shape parameter.

Potentially, $\alpha$ and $\beta$ can be time varying and class specific. They can be estimated separately for each class. However, consistent with the Weibull regression model, we assume constant $\beta$ over
all clients and allow $\alpha$ to vary according to a Weibull regression model, where age, gender, and year of admission are independent or explanatory variables (Klein and Moeschberger 2003). Let $y_1, \ldots, y_k$ denote values for $k$ explanatory variables; $\alpha$ can then be represented by:

$$
\alpha = \exp(\gamma_0 + \gamma_1 y_1 + \ldots + \gamma_k y_k),
$$

where $\gamma_0, \gamma_1, \ldots, \gamma_k$ are regression coefficients to be estimated from data. This expression implies that each observation or client may have its own $\alpha$ but $\beta$ is constant over all observations.

Regression parameter estimates are obtained using maximum likelihood. Specifically, let $x_i$ be the LOS of the $i$th individual in the data set, $NC$ be the set of all non-censored cases, and $RC$ be the set of all right censored data; the likelihood function denoted by $L(\gamma_0, \gamma_1, \ldots, \gamma_k, \beta)$ can then be written as:

$$
L(\gamma_0, \gamma_1, \ldots, \gamma_k, \beta) = \prod_{x_i \in NC} \Pr(X = x_i|\gamma_0, \gamma_1, \ldots, \gamma_k, \beta) \prod_{x_i \in RC} \Pr(X > x_i|\gamma_0, \gamma_1, \ldots, \gamma_k, \beta),
$$

where the first item can be expressed by the probability density function and the second one can be expressed by the survival function. As a result of this analysis, we can obtain a distinct Weibull distribution for each class of clients and can also test whether it varies with time.

Note that the LOS in this research indicates the length of time in care, excluding the length of time in the waitlist. This is primarily because data records in practice often indicate when a client entered care but not when he/she entered the waitlist. This implies that LOS distributions may be slightly underestimated by using this method. Some adjustment to historical data may be needed. It could also be explored through sensitivity analysis.

### 4.2.3. Pre-Loading Existing Clients

Often, the system being modeled has been in operation for many years and contains people who were admitted to care in the past. Thus, we believed that simply allowing the system to warm-up and reach a “steady state” is not practical.

To incorporate these clients who are currently in care and in the waitlist, their information is pre-loaded into the simulation before it starts. For each existing client in care, the information
includes his/her gender, age, and amount of time in care so far. In order to assign a remaining LOS to each of them, we used a conditional LOS distribution for each class. Specifically, denoting the amount of time in care so far by \( u \), the CDF of a conditional Weibull distribution can be obtained as follows:

\[
F(x|u) = 1 - e^{-(x/\alpha)^\beta + (u/\alpha)^\beta} \quad x > u.
\] (9)

Thus, based on expression (9), a remaining LOS can be randomly generated for each client who is in care.

On the other hand, each client in the waitlist is treated the same as a new client arriving after time zero, and a LOS for each of these clients is randomly generated according to his/her age and gender.

4.3. Optimization

For several decades, simulation has been used as a descriptive tool in the modeling and analysis of complex systems. However, simulation has also been criticized for the lack of optimization capability. In this problem, given a capacity level in each year, the simulation model can output the resulting service level, whereas it cannot determine the minimal capacity level required each year by itself.

With recent advances in computing technology, it now becomes feasible to integrate simulation models and optimization techniques together for decision-making. A variety of simulation optimization approaches have been proposed, and there are also several review papers in the literature that discuss theories and applications of these techniques, including Fu (1994) and Tekin and Sabuncuoglu (2004). The techniques for continuous decision spaces include gradient-based methods, response surface methodology, and stochastic approximation; those for discrete decision spaces are mainly related to meta-heuristics.

Based on the structure of this problem, we developed two specific optimization techniques.
4.3.1. A Sequential Bisection Search Algorithm  In this section, we describe a simulation-based sequential bisection search algorithm to solve this problem. Let $k$ denote the iteration index, $\bar{s}_t$ and $\underline{s}_t$ denote initial upper and lower bounds for the number of beds in year $t$, $\epsilon$ denote a tolerance level, $N$ denote the number of simulation replications to be run, and $n$ denote the replication index.

Algorithm 1

Step 0: Choose appropriate values for $\bar{s}_t$ and $\underline{s}_t$ for each year $t$, $\epsilon$, and $N$; set $t = 1$ and $k = 0$.

Step 1: If $\bar{s}_t \leq \underline{s}_t + 1$, set $s^*_t = \bar{s}_t$ and go to Step 4; otherwise, set $k = k + 1$ and $s^*_t = (\bar{s}_t + \underline{s}_t)/2$.

Step 2: Run the simulation from time zero to the end of year $t$ with $s^*_l$, $l = 1, \ldots, t-1$, and $s^*_t$, for $N$ independent replications; for each replication $n$, record the achieved service level denoted by $\pi_n$ (i.e., the fraction of clients who are placed in care within the time threshold $\gamma$) in this year $t$; calculate the mean service level $\bar{\pi}$.

Step 3: If $\tau - \epsilon \leq \bar{\pi} \leq \tau + \epsilon$, set $s^*_t = s^*_t$ and go to Step 4; otherwise, if $\bar{\pi} < \tau - \epsilon$, set $\underline{s}_t = s^*_t$ and go to Step 1; otherwise, if $\bar{\pi} > \tau + \epsilon$, set $\bar{s}_t = s^*_t$ and go to Step 1.

Step 4: If $t = T$, stop; otherwise, set $t = t + 1$ and $k = 0$ and go to Step 1.

The idea of this algorithm is to find $s^*_t$ year by year. At the beginning of every simulation run, existing clients who are in care and in the waitlist are pre-loaded into the system. The algorithm first finds $s^*_1$ through the bisection search. To find $s^*_2$, the simulation starts over with $s^*_1$ as the number of beds in year 1, to capture carryover clients in the system at the end of the first year. The algorithm continues until $s^*_T$ is found. In particular, for the sake of efficiency, we set the tolerance $\epsilon$ in the algorithm, so that $s^*_t$ is the first number during the process which makes the average service level achieved in year $t$ lie within $[\tau - \epsilon, \tau + \epsilon]$.

In addition, we note that the achieved service level in each year and in each simulation replication is represented by the percentage of clients who receive the service, i.e., leave the queue, within the time threshold in each year, regardless of the time when they enter the system. This implies that there may be a small portion of these clients who enter the system in the previous year, i.e., they wait in the queue until receiving the service in the current year.
4.3.2. A Simultaneous Search Algorithm  Since the bisection search algorithm above determines the optimal capacities sequentially, many intermediate results obtained during the process are wasted. Therefore, we designed a more efficient algorithm by adjusting $s_t$ simultaneously in each iteration.

Let $\theta_1$ and $\theta_2$ denote two step-size parameters and $K$ denote a maximum iteration number.

Algorithm 2

Step 0: Choose appropriate values for $\epsilon$, $\theta_1$, $\theta_2$, $N$, and $K$; for each year $t$, set $s^1_t = s_0$; set $k = 0$.

Step 1: Set $k = k + 1$; run the simulation for the entire time horizon with $s^k_t$, $t = 1, \ldots, T$, for $N$ independent replications; for each replication $n$ and each year $t$, record the achieved service level denoted by $\pi_t n$ (i.e., the fraction of clients who are placed in care within the time threshold); calculate the mean service level $\bar{\pi}_t$ for each year $t$.

Step 2: If $\tau - \epsilon \leq \bar{\pi}_t \leq \tau + \epsilon$ for any year $t$ or $k \geq K$, set $s^*_t = s^k_t$ and stop; otherwise, for each year $t$, set

$$s^{k+1}_t = \begin{cases} s^k_t + \theta_1 (\tau - \bar{\pi}_t) & \text{if } \bar{\pi}_t < \tau - \epsilon \\ s^k_t - \theta_2 (\bar{\pi}_t - \tau) & \text{if } \bar{\pi}_t > \tau + \epsilon \\ s^k_t & \text{otherwise} \end{cases}$$

and go to Step 1.

In each iteration, this algorithm simulates the entire time horizon, and adjusts the capacity level in each year based on the resulting service level in that year. The search mechanism of this algorithm is similar to that of gradient-based methods (Tekin and Sabuncuoglu 2004), whereas the gradient information is represented by the deviation between the current service level and the target. Moreover, we note that this algorithm would also work for more complex service level criteria or for solving capacity allocation problems.

To balance efficiency and reliability, we chose the following parameter values based on computational experiments: $\bar{s}_t = 2s_0$ and $s_t = 1$ in Algorithm 1, set $\theta_1 = 0.2s_0$, $\theta_2 = 0.1s_0$, and $K = 50$ in Algorithm 2, and set $N = 100$ and $\epsilon = 3\%$ in both. Due to variability, the two algorithms may not produce exactly the same solution, but usually they are very close. However, the simultaneous
search algorithm is much more efficient than the sequential bisection search algorithm, typically requiring only 30% of the run time that the other does.

4.4. Implementation

We developed a decision support system for LTC managers to use in practice based on the methods above.

The decision support system consists of two main components: a discrete event simulator and a front-end interface. The discrete event simulation model was developed in Arena 10.0, where the optimization techniques (Algorithm 2 is preferred) were coded in Visual Basic for Applications (VBA). The front-end interface developed in Excel contains all data and information to be used by the Arena simulation model, including population data and per capita arrival rates to generate arrival distributions, parameter values of LOS distributions obtained from survival analysis, information on existing clients in care and in the waitlist, a current capacity level, as well as all other relevant inputs. In addition, the LIFEREG procedure in SAS was used for survival analysis. When the simulation ends, the front-end interface stores the optimal solution and other outputs from the simulation in both graphical and tabular forms. In addition, it allows the user to set the capacity levels and determine the resulting service performance, as well as to modify the parameter values for sensitivity analysis or scenario testing.

In summary, this decision support system enables the user to:

- Estimate service performance with any sequence of pre-specified capacity levels;
- Determine a sequence of optimal capacity levels required to meet the service level criterion;
- Conduct sensitivity analysis of important system input parameters, such as service levels, initial conditions, population growth rates, per capita arrival rates, and LOS distributions.

5. Applications

5.1. Application I: Determining Bed Capacity for the Vancouver Island Health Authority

5.1.1. Background and Data The Vancouver Island Health Authority (VIHA) is one of six health authorities in BC, providing health services, such as public and environmental health,
home care and support, long-term care, hospital care, and mental health, across a widely varied geographic area covering approximately 56,000 square kilometers and consisting of 15 local health areas (LHAs).

LTC is managed by the Home and Community Care (HCC) program. One challenge they faced at the time we initiated this research was to plan LTC bed capacity to meet future needs. More precisely, they required a decision support system that could “forecast” long-term capacity requirement and allow ongoing scenario testing. We arrived at the objective of having a sufficient number of beds each year to ensure that 85% of clients would be placed in care within 30 days every year. In terms of expression (1), this corresponded to $\gamma = 30$ and $\tau = 0.85$.

Our approach was applied at the geographic region level for 2009-2020. Data sources for estimating the arrival and LOS distributions include: the Population Extrapolation for Organization Planning with Less Error (P.E.O.P.L.E.) 32 database from BC Stats and the Continuing Care Information Management System (CCIMS) database collected by VIHA. These data sources are updated on a yearly basis, are readily available to VIHA personnel, and were utilized in a manner that required the minimum of data cleaning activities.

The P.E.O.P.L.E. database provides population forecasts and historical population sizes by geographic area, age, and gender in one year increments. This data was aggregated to the LHA level and broken down by gender and age (less than 55, 56-65, 66-75, 76-85, and greater than 85). The CCIMS database was used to estimate the per capita arrival rates to the system by age group, gender, and LHA, which were calculated based on the weighted moving average of the last four years. However, this database only contained information of admitted clients. The waitlist length in each LHA was only available for the date the database was given; the historical waitlist information was not available. Hence, the number of arrivals in the past was assumed to be the number of clients admitted. This implies that our analysis may have underestimated the arrival rates. The impact of this assumption was investigated through sensitivity analysis.

The CCIMS database was also used to estimate the LOS distributions. It includes information of more than 40,000 clients in all the LHAs since 1990. Again, because the historical individual
wait time data was unavailable, the LOS distributions may be slightly underestimated by only using the CCIMS database. Another challenge that we faced was that LTC admission criteria were implemented in BC in 2003 so that only clients with higher acuity levels were admitted. Hence, we would expect shorter LOSs for post 2003 clients. To estimate LOS distributions, we split the clients into two groups: pre-2003 and post-2003, and estimated those for each group separately. In particular, 36% of clients in the database were admitted after 2003 and one third of them were still in care.

Figure 2, based on the post-2003 data, shows that not considering the information contained in the censored cases seriously underestimates the LOS; median LOS changed from 400 days to 110 days by ignoring the censored cases. Note that the survival curve based on all the data was based on the Kaplan-Meier method (Klein and Moeschberger 2003). In addition, Figure 3 shows the difference in the survival curves for male and female clients, based on the post-2003 data. It is clear that the LOS of females is normally double that of males. Moreover, analyses also showed that LOS differs significantly by age and region. This provided the reasons for separating clients into different groups by age, gender, and geographic region.

In order to estimate LOS distributions as accurately as possible, the pre- and post-2003 data was analyzed at the aggregate level by including age group, gender, and LHA as explanatory variables. Thus, LOS distributions for all the classes were represented by Weibull distributions with a common shape parameter value and distinct scale parameter values.

For the existing clients in care, the simulation model pre-loaded their information and randomly generated a remaining LOS for each of them based on the corresponding conditional Weibull distribution (9). Moreover, although the waitlist length in each LHA was available, there was no information about the age and gender of each client in it. We had to assume that the existing clients in the waitlist can be represented by the people who entered care in the last year. Using this assumption, we split these clients in each LHA by age group and gender and then applied the same LOS distributions as for new arrivals to them.
Figure 2  Kaplan-Meier Survival Curves with and without Censored Data

Figure 3  Kaplan-Meier Survival Curves of Male and Female Clients
5.1.2. Results and Analysis  In this section, we describe results for one particular geographic region of VIHA, where there were 2392 existing clients in care (i.e., the number of available beds) and 240 clients in the waitlist. Our intention in this study was not only to show the result of using our simulation methodology, but also to compare it to the fixed ratio and SIPP approaches. In this regard, the forecasts (the number of beds required during 2009-2020) obtained using these approaches are displayed in Figure 4. We also show the forecasted number of arrivals per year from each class in Figure 5 (since the number of arrivals aged less than 65 is small and stable, we only show the classes over age 65).

The curve of the optimal capacity levels over time based on our approach is “U-shaped”. From Figure 5, we believe that the required capacity rises during 2014-2020 mainly due to the rapid increase of the population aged between 65 and 85. On the other hand, we identified several reasons for the initial drop in capacity.

1. Since the service level at that moment was much lower than the target (VIHA 2009) and also there were a large number of clients in the waitlist, much extra capacity is needed to reduce the wait time and meet the service level criterion in 2009.

2. The second reason for this initial drop is the change in admission criteria instituted in 2003. The people admitted pre-2003 are of lower acuity and thus generally have a longer LOS. Based on our analysis, a large number of these people will be leaving the system over the next a few years, causing the drop in required capacity.

3. From Figure 4, in spite of the increase in the population aged below 75, the size of the population aged over 75 continues decreasing during 2009-2013.

Based on these observations, we recommended not relaxing admission criteria during this period. The consequence of doing that would be that clients with lower acuity would be admitted. Since these clients would have longer LOSs, more capacity would be needed in the future when arrival rates increase.

5.1.3. Comparison to the Ratio Approach  We calculated the capacity levels based on the current planning ratio of 75 beds per 1000 population aged over 75 (Legislative Assembly of British
Columbia 2006) as shown in Figure 4. Compared to the simulation approach, it is clear that this approach significantly underestimates the capacity requirement during 2009-2018. Furthermore, according to the trend of the curve, it may significantly overestimate the capacity requirement after 2020. Nevertheless, we observed overestimation in some other LHAs by using this approach. This suggests that the ratio value used at this moment is not accurate. The problems of this approach are many-fold:

1. It ignores geographic specific differences in arrival and LOS.
2. It does not account for clients in care and in the waitlist at the beginning of each year.
3. It ignores the population aged below 75, who still account for 20% of total clients.
4. It also ignores differences in arrival and LOS between the two age groups (75-85 and greater than 85) and between the two gender groups.

A parallel study (Zhang et al. 2009) compares the service-based and ratio-based approaches in detail and shows that a more accurate ratio value or a better ratio-based policy could be achieved after rigorous analysis, but a universal ratio-based policy that can be applied in general may not be achievable. As it lacks a rigorous foundation, using this approach in general is not recommended.

5.1.4. Comparison to the SIPP Approach  To calculate the capacity levels based on the SIPP approach, we aggregated the clients into a single class. We used the overall arrival rate and estimated the LOS distribution for this single class based on an exponential distribution. Using the closed-form expression of the $M/M/s$ queueing system, we calculated the number of beds required to satisfy the service level criterion in the steady state in each year. Figure 4 shows that the required capacities based on the SIPP approach are significantly lower than those based on our approach. Several reasons for this include:

1. Similarly, the SIPP approach does not consider existing clients in care and in the waitlist at the beginning of each year.
2. It ignores differences in arrival and LOS among the five age groups and between the two gender groups.
3. Most importantly, Figure 6 shows that the exponential distribution provided a much poorer fit to the Kaplan-Meier curve than the Weibull distribution did.

In addition to the original SIPP approach, we investigated two modified “SIPP’’ approaches. For the SIPP’ approach, we modeled the system as multiple independent $M/M/s$ queueing systems, one for each class. The LOS distribution for each class was again estimated based on an exponential distribution. We then calculated the number of beds required to satisfy the service level criterion in each year for each class independently. Summing up these numbers, we obtained the total capacity required in each year. For the SIPP” approach, we modified the SIPP approach by replacing the mean of the estimated exponential distribution by that of the estimated Weibull distribution.

The required capacities by using these two approaches are also displayed in Figure 4. The SIPP’ approach predicts the required capacities closer to the one obtained using our simulation approach. One possible reason is that it accounts for differences in arrival and LOS among the five age groups and between the two gender groups. Nevertheless, the loss of the risk pooling causes more capacity required as well. In contrast, the solution based on the SIPP” approach is the closest to the one based on our approach. Note that it would perform even better if there were no clients in the waitlist. However, similar analyses in other regions, which we do not report herein, and the next section suggest that it does not work well all the time.

Therefore, although the SIPP approach typically works well in the context of call centers, it may not be an appropriate approach for this problem. In addition to the reasons above, one of the key differences between this two contexts deserves more attention. LOS in LTC (in years) is much longer than service time in call centers (in minutes), while service level for LTC is measured yearly and that for call centers is usually measured hourly. Therefore, it is almost impossible for the LTC system to reach the steady state in each year. This suggests that, in general, the SIPP approach may not be suitable to use when the service time is very long compared to the period length. This is consistent with observations in Atlason et al. (2008).

In summary, we believe that the simulation approach uses the most information and consistently performs most reliably, and thus recommend using it in practice.
5.1.5. Sensitivity Analysis  A number of scenarios were chosen to investigate the sensitivity of the model to assumptions. These scenarios were chosen to represent the most likely changes to occur in the coming years based on knowledge of the systems, including the base case, increasing
Figure 6  Estimated Exponential Distribution versus Estimated Weibull Distribution Compared to the Kaplan-Meier Curve

Figure 7  Capacity Levels for Different Scenarios
and decreasing the LOS by 5%, increasing and decreasing the per capita arrival rates by 5%, and a diversion of clients into possible future dementia housing. All scenarios were run using the methodology above to find the required capacity in each year to meet the service level criterion. Figure 7 compares the base case to increasing and decreasing the LOS by 5% as well as increasing and decreasing the per capita arrival rates by 5%. It seems that there is no significant difference in the required capacity between changing the LOS and changing the per capita arrival rates.

5.2. Application II: Supporting Capacity Decision Making for the Louis Brier Home and Hospital

5.2.1. Background and Data The methodology developed in this paper also enabled us to assist the board and executive of the Louis Brier Home and Hospital (LBH) decide when and where to add capacity. LBH provides long-term care to elderly Jewish and other residents in Vancouver, BC. In 2005, the Board of Trustees of LBH created a task force to investigate the future LTC needs of Jewish seniors in Vancouver and provide recommendations of how best to meet them. An issue was whether the 218 bed facility should expand on site to 350 beds, move to a new site, or develop a second site. For reasons identified through focus groups, the first option was considered the most desirable. Thus, before ruling out other options, it was necessary to determine how far into the future the current site could satisfy community needs.

Our analysis focused on a 11-year planning extending to 2020. Data needed for estimating population sizes was obtained from Shahar and Gerber (2004), which provided estimates of the Jewish population of Vancouver for 2011 and 2021 based on actuals from the 2001 census and demographic analysis. LBH historical records were used to estimate per capita arrival rates and LOS distributions. Historically, the waitlist had been very short so its effect was ignored in our analysis, and it was observed that approximately 85% of residents of LBH were Jewish.

The client data contained 921 records 36% of them admitted after 2003 and 23% of them still in care. For analysis, we again split the client data into two groups: pre-2003 and post-2003, for the
reasons described above, and estimated the LOS distributions for each groups separately. Due to
the limited amount data for some age groups, we only considered gender as an explanatory variable
in the Weibull regression model.

5.2.2. Results and Analysis  Figure 8 displays the number of beds required for 2009-2020
based on the different approaches. Clearly, the solution based on the fixed ratio approach (75
beds per 1000 population aged over 75) significantly underestimates the capacity required. This
demonstrates that the provincial target is not applicable at this facility. On the other hand, the
solution based on the original SIPP approach is the closest to the one based on our recommended
simulation approach. Consistent with the previous case, the solution based on the SIPP’ approach
is larger than that based on the SIPP approach. As indicated earlier, the SIPP” approach does not
work well in this case, where it significantly overestimates the capacity required. In addition, the
figure shows that, regardless of the approach used, the capacity required increases almost linearly
by 4% every year.

This result is different from the previous case. Potential reasons include:
Figure 9  Estimated Exponential Distribution versus Estimated Weibull Distribution Compared to the Kaplan-Meier Curve

1. There were just two clients in the waitlist for the LBH case compared to 240 clients for the VIHA case. Therefore, there is no need of much extra capacity for LBH in the first year.

2. For LBH, we only considered the two gender groups without taking age into account due to the data limit for the survival analysis. This simplification makes the system of LBH much closer to the $M/M/s$ queueing system.

3. Arrival rates were based on interpolating population forecasts of the future years based on the numbers in 2001, 2011 and 2021, i.e., the number of arrivals linearly increases by 4% every year. In contrast, the population data from P.E.O.P.L.E. used in the previous case shows great differences among the different age groups.

4. Figure 9 depicts the estimated exponential distribution, the estimated Weibull distribution, and the Kaplan-Meier curve based on all the post-2003 data. It is clear that the estimated exponential distribution is much closer to the estimated Weibull distribution and the Kaplan-Meier curve than in the VIHA case.
Table 1  Achieved Service Levels of the Optimal Solution

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.62</td>
<td>0.45</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.09</td>
<td>0.63</td>
<td>0.46</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.94</td>
<td>0.88</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.10</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.82</td>
<td>0.62</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.45</td>
<td>0.90</td>
<td>0.48</td>
<td>0.85</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.97</td>
<td>0.74</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0.96</td>
<td>0.64</td>
<td>0.93</td>
<td>0</td>
<td>0</td>
<td>0.74</td>
<td>0.18</td>
<td>0.48</td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0.89</td>
<td>0.83</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>0.79</td>
<td>0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>0.94</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
<td>0.21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>0.86</td>
<td>0.84</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.84</td>
<td>0.84</td>
<td>0.86</td>
<td>0.87</td>
<td>0.85</td>
<td>0.84</td>
<td>0.86</td>
</tr>
</tbody>
</table>

A blank cell represents “one”. A gray row represents that the service level criterion cannot be met in at least three consecutive years.

After considering this analysis, the Board recommended redevelopment of the current site. Detailed implementation plans are under development.

5.3. Insights from Both Applications

Based on the solution obtained using our approach, the mean service level achieved in each year meets the target level. However, looking more closely at the simulation output, a few interesting findings deserve more discussion. Since these findings are common for both applications, we only focus on the former.

Table 1 shows the resulting service level in each year in the first 20 simulation replications as well as the average service level in each year over all the replications. A blank cell in the table represents “one”, i.e., all the clients in this year in this replication are placed in care within the wait time threshold. For clarity, we remove “one” in the table. A gray row in the table represents that the service level criterion cannot be met in at least three consecutive years.

While the average service levels in the period normally meet the criterion, we observed that there is great variation in the achieved service level between the replications. In particular, there are
many blank cells (representing “one”) except in the first year. In all the other years, around 80% of the replications achieve a 100% service level. This demonstrates a high possibility of overcapacity. Second, from the gray rows, it is clear to see a high correlation between the achieved service levels in consecutive years. Whenever a low service level is achieved in a year, there is a high chance of achieving a low service level in the next year as well. There are 28 such rows among all the 100 replications in total. In many replications, the service levels are consistently low over the planning horizon. This phenomenon demonstrates that once clients accumulate in the queue, much extra capacity is needed in order to meet the service level criterion rapidly. We believe that the main reason for these two observations is because the service time is very long compared to the period length.

One simple way to reduce the likelihood of having replications with consistently low service levels is to impose a stronger service level criterion based on the simultaneous probability, such as:

$$\Pr(W_t(s_t) \leq \gamma, W_{t+1}(s_{t+1}) \leq \gamma) \geq \tau', \quad t = 1, \ldots, T - 1,$$  \hspace{1cm} (10)$$

or a even stronger one as expression (2). As mentioned earlier, expression (2) or (10) can be approximated by expression (1) with a different value of $\tau$. Therefore, in this case, we only need to solve the problem with a higher value of $\tau$, i.e., increasing capacity in each year. For instance, Figure 10 compares the required capacities of the base case to those by using the Bonferroni approach with $\tau' = 0.85$ and the service level criterion defined as (2) and (10), respectively. In particular, for the latter case, the number of beds required in each year increases almost 50, and in around 90 replications over 100, the achieved service level in any year is 100%. This suggests that although this method reduces the likelihood of having replications with consistently low service levels, it simultaneously raises the possibility of overcapacity.

6. Adaptive System

6.1. State-Dependent Model

To find a more effective method to overcome the problems alluded to in Section 5.3, we now describe a modified approach to address the LTC capacity planning problem. Specifically, we seek a series
of policies for managers instead of a series of capacity levels. A policy, normally derived in dynamic programming problems, refers to a set of selected decisions for each state of the system. A state refers to some information about the system that can be observed by the decision maker and used for decision-making. A commonly-used state is the number of carryover clients at the end of each year. However, due to the multiple classes and the necessity of separating carryover clients in care and in the waitlist, there would be a massive state vector for this problem.

To simplify this, we chose the achieved service level in the previous year as the state of the system. For tractability, we aggregated service levels into $J$ states. Let $s^j_t$ denote the number of beds in year $t$ and state $j$ ($j = 1, \ldots, J$). The problem is then to determine the value of $s^j_t$ in year $t$ and state $j$ (a series of policies) so as to satisfy the service level criterion, i.e.,

$$\Pr(W_t(s^j_t) \leq \gamma|j) \geq \tau \quad \forall j = 1, \ldots, J \quad t = 1, \ldots, T.$$  \hfill (11)

In other words, a look-up table can be provided, from which managers can determine the required capacity with respect to the year and the observed state. On the other hand, this look-up table
provides an upper and lower bound for the required capacity level in each year, which are also very useful for managers to establish a long-term capacity plan.

6.2. Optimization

Let \( s_j^t^* \) denote the number of beds so that the service level criterion is met in year \( t \) and state \( j \). The algorithm below (Algorithm 3) that we developed to find \( s_j^t^* \) is a variation of Algorithm 1. Specifically, in each year, \( s_j^t^* \) for state \( j \) is iteratively derived by the simulation-based bisection search. Since the initial state is known, this approach starts with the second year, and \( s_1^1^* \) is derived by Algorithm 1. For each year \( t \) \((t \geq 2)\) and state \( j \), let \( s_j^t \) and \( s_{jk}^t \) denote the initial upper and lower bounds for the number of beds, respectively; for each state \( j \), let \( z_j \) denote a binary indicator. Other parameters remain the same as before.

**Algorithm 3**

**Step 0:** Run Algorithm 1 for year 1, and obtain \( s_1^1^* \); for each year \( t \) \((t \geq 2)\) and state \( j \), choose appropriate values for \( s_j^t \), \( s_{jk}^t \), \( \epsilon \), and \( N \); set \( t = 2 \), \( k = 0 \), and \( z_j = 1 \) for each state \( j \).

**Step 1:** For each state \( j \), if \( z_j = 0 \): if \( s_j^t \leq s_{jk}^t + 1 \), set \( s_j^t^* = s_j^t \) and \( z_j = 1 \); otherwise, set \( k = k + 1 \) and \( s_{jk}^t = (s_j^t + s_{jk}^t)/2 \).

**Step 2:** Run the simulation from time zero to the end of year \( t \) with \( s_j^1^*, s_j^t^* \) \((l = 2, \ldots, t - 1)\), \( s_j^t^* \) (for \( j \), \( z_j = 1 \)), and \( s_{jk}^t \) (for \( j \), \( z_j = 0 \)), for \( N \) independent replications; for each replication \( n \), according to the state (i.e., the achieved service level in year \( t - 1 \) of this replication), record the achieved service level denoted by \( \pi_n \) (i.e., the fraction of clients who are placed in care within the time threshold) in this year \( t \) into the corresponding set \( j \); for each set \( j \), calculate the mean service level \( \bar{\pi}_j \).

**Step 3:** For each state \( j \), if \( z_j = 0 \): if \( \tau - \epsilon \leq \bar{\pi}_j \leq \tau + \epsilon \), set \( s_j^t^* = s_{jk}^t \) and \( z_j = 1 \); otherwise, if \( \bar{\pi}_j < \tau - \epsilon \), set \( s_j^t = s_{jk}^t \); otherwise, if \( \bar{\pi}_j > \tau + \epsilon \), set \( s_j^t = s_{jk}^t \).

**Step 4:** If \( z_j = 1 \) for all \( j \), go to Step 5; otherwise, go to Step 1.

**Step 5:** If \( t = T \), stop; otherwise, set \( t = t + 1 \), \( k = 0 \), and \( z_j = 0 \) for each state \( j \), and go to Step 1.
6.3. Results and Analysis

Results of using the adaptive approach for the first application are discussed here. We set \( J = 3 \) and defined State 1 as the service level achieved in the previous year is greater than 0.9, State 2 between 0.5 and 0.9, and State 3 less than 0.5. The values of the parameters remain the same. This aggregation was chosen so that using the optimal policy in each year results in roughly 60 replications out of 100 in State 1 and roughly equal sizes for States 2 and 3. Since we seek an average service level of 85\%, the majority of replications have to be in State 1.

The optimal policy for each year in the form of a look-up table appear in Table 2. The table starts with \( s_1^* = 2508 \) obtained by Algorithm 1. For any other year, \( s_2^t \) is between \( s_1^* \) and \( s_3^* \). Based on these policies, a similar tabular output was derived as shown in Table 3. Note that we used the same random number streams for the 100 replications as before.

From Table 3, clearly, the average service levels in the period still normally meet the criterion. In contrast, there are significantly fewer blank cells (representing “one”) in this table than Table 1. This suggests that the possibility of overcapacity is lower. Also, it is clear that the correlation between the achieved service levels in consecutive years is much lower. In most of the replications, whenever a low service level is achieved, the service level in the next year meets the criterion. There are only 9 gray rows where the service level criterion cannot be met in three consecutive years. This result can even be improved by increasing the number of states considered and simultaneously the number of simulation replications.

Note that \( s_1^* \) and \( s_3^* \) can be viewed as the best and worst cases of the required capacity in year \( t \), respectively. In other words, an upper bound and a lower bound for the required capacity in each year can be obtained. Moreover, we found that, using both the adaptive and the non-adaptive approaches, the average total capacities required during the planning horizon are very close. This
Table 3  Achieved Service Levels of the Adaptive Policy

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.62</td>
<td>0.53</td>
<td>0.52</td>
<td>0.02</td>
<td>0.98</td>
<td>0.27</td>
<td>0.53</td>
<td>0.58</td>
<td>0.36</td>
<td>0.97</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.94</td>
<td>0.47</td>
<td>0.92</td>
<td>0.02</td>
<td>0.93</td>
<td>0.03</td>
<td>0.93</td>
<td>0.03</td>
<td>0.93</td>
<td>0.03</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.82</td>
<td>0.75</td>
<td>0.35</td>
<td>0.02</td>
<td>0.92</td>
<td>0.88</td>
<td>0.92</td>
<td>0.35</td>
<td>0.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.97</td>
<td>0.92</td>
<td>0.02</td>
<td>0.99</td>
<td>0.90</td>
<td>0.01</td>
<td>0.91</td>
<td>0.09</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.91</td>
<td>0.96</td>
<td>0.50</td>
<td>0.66</td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.96</td>
<td>0.59</td>
<td>0.38</td>
<td>0.94</td>
<td>0.92</td>
<td>0.98</td>
<td>0.94</td>
<td>0.99</td>
<td>0.74</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.89</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
<td>0.49</td>
<td>0.75</td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.89</td>
<td>0.06</td>
<td>0.92</td>
<td>0.02</td>
<td>0.98</td>
<td>0.92</td>
<td>0.42</td>
<td>0.95</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.95</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.66</td>
</tr>
<tr>
<td>15</td>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.86</td>
</tr>
<tr>
<td>16</td>
<td>0.94</td>
<td>0.78</td>
<td></td>
<td></td>
<td>0.47</td>
<td>0.94</td>
<td>0.77</td>
<td>0.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
</tr>
<tr>
<td>18</td>
<td>0.90</td>
<td>0.55</td>
<td>0.46</td>
<td>0.96</td>
<td>0.64</td>
<td>0.99</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.90</td>
<td>0.47</td>
<td>0.99</td>
<td>0.39</td>
<td>0.59</td>
<td>0.95</td>
<td>0.11</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.89</td>
</tr>
</tbody>
</table>

Mean: 0.86 0.87 0.86 0.85 0.86 0.88 0.86 0.85 0.87 0.84 0.87 0.86

A blank cell represents "one".
A gray row represents that the service level criterion cannot be met in at least three consecutive years.

implies that resource utilization is greatly improved by using an adaptive policy.

The results and analysis presented here demonstrate that this adaptive policy depending on the state information is useful and worth further investigation. In other words, in addition to base capacity, flexible surge capacity is very critical. However, it may be difficult to provide flexible surge capacity in practice, especially for an individual LTC facility. Nevertheless, this is probably easier for a regional level LTC program, such as within a LHA of VIHA. Also, third party delivery of LTC services is an established practice. To create or expand surge capacity, a public LTC program could purchase temporary beds from the private sector. Note that there have been discussions on issuing temporary licenses by government for LTC beds in the private sector in Canada (Prince Edward Island Department of Health 2009).

7. Conclusion

This paper describes a methodology for setting LTC capacity levels over a multi-year planning horizon to achieve target wait time service levels. We proposed and applied an approach that integrates demographic and survival analysis, discrete event simulation, and optimization. Based on this methodology, a decision support system was developed for use in practice. We illustrated this
approach through two case studies. We also compared our approach to the current ratio approach and the SIPP approach developed in the call center literature. Finally, a state-dependent model and an adaptive system for capacity planning were considered to achieve better resource utilization.

From a methodological perspective, the innovation of this research is the combination of several operations research and statistical methods. Since our methods are driven by service levels, they are preferable to commonly-used ratio-based methods. Also, because LOS distributions tend not to be exponential and LOS is long relative to the period length, it is preferable to SIPP approaches. From a practical perspective, this is the first attempt to develop a rigorous tool that can be used by managers of LTC programs or facilities to evaluate system performance and to make long-term capacity planning. We are hopeful that using the tool will result in both improved access to LTC and reduced ALC patients in acute care. We hope that this research will guide capacity planning and allocation decisions in this industry at all BC health authorities and more broadly.

We believe that the following observations and recommendations follow from our research.

• Survival analysis reveals that LOS varies considerably by age, gender, and geographic region. This must be accounted for in estimating future capacity needs.

• The current approach based on a fixed ratio of beds per population should not be used because it ignores differences in population characteristics by region and historical data.

• The SIPP approach should not be used because it ignores a large number of clients in care at the beginning of each year, the differences in arrival and LOS by age and gender, and our observation that LOS tend to follow a Weibull distribution.

• Service levels from year to year will be highly correlated so that the results in practice may differ significantly from results based on averages over simulations. An adaptive policy based on historical service levels may improve performance. To execute this requires availability of surge or temporary capacity.

• System managers must avoid relaxing the admission criteria even when capacity utilization is low. Admitting lower acuity clients could result in increased lengths of stay and the need for more capacity in the future.
Our intent is to build on this approach to develop methods for planning for a series of interrelated services including assisted living, dementia care, and home care and as well to plan for different levels of acuity within a single facility.

Acknowledgments

We gratefully thank Steve Atkinson, Director of Operations Research, Vancouver Island Health Authority, and Peter Kafka, Chief Executive Officer of the Louis Brier Home and Hospital for providing data and many helpful discussions. This research has been partially supported by NSERC grant RGPIN 5527.

References


worker requirements in a service operation with time-dependent customer demand. *Queueing Systems*, 3, 265-276.


