A Queuing Model for Ventilator Capacity Management during the COVID-19 Pandemic

Samantha Zimmerman · Alexander R. Rutherford · Alexa van der Waall · Monica Norena · Peter Dodek

Received: date / Accepted: date

Abstract We present a queue model to inform ventilator capacity management under different COVID-19 pandemic scenarios. Our model was used to support ventilator capacity planning during the first wave of the COVID-19 epidemic in British Columbia (BC), Canada. The core of our framework is an extended Erlang loss model, which incorporates COVID-19 case projections, along with the proportion of cases requiring a ventilator, the delay from symptom onset to ventilation, non-COVID-19 ventilator demand, and ventilation time. We implemented our model using discrete event simulation to forecast ventilator utilization. The results predict when capacity would be reached and the rate at which patients would be unable to access a ventilator. We further determined the number of ventilators required to meet a performance indicator target for ventilator access. We applied our model to BC by calibrating to the BC Intensive Care Unit Database and by using local epidemic projections. Epidemic scenarios with and without reduced transmission, due to social distancing and other behavioral changes, were used to link public health interventions to operational impacts on ventilator utilization. The results predict that reduced transmission could potentially avert up to 50 deaths per day by ensuring that ventilator capacity would likely not be reached. Without reduced transmission, an additional 181 ventilators would be required to meet our performance indicator target that 95% of patients can access a ventilator immediately. Our model provides a tool for policy makers to quantify the interplay between public health interventions, necessary critical care resources, and performance indicators for patient access.

Keywords COVID-19 · Critical care · Ventilator capacity planning · Erlang loss model · Discrete event simulation · Simulation optimization

S. Zimmerman
Department of Mathematics, Simon Fraser University, 8888 University Dr., Burnaby, BC, Canada V5A 1S6
E-mail: slzimmer@sfu.ca

A. R. Rutherford
Department of Mathematics, Simon Fraser University, 8888 University Dr., Burnaby, BC, Canada V5A 1S6
E-mail: arruther@sfu.ca

A. van der Waall
Department of Mathematics, Simon Fraser University, 8888 University Dr., Burnaby, BC, Canada V5A 1S6
E-mail: avanderw@sfu.ca

M. Norena
Center for Health Evaluation and Outcome Sciences, 588 - 1081 Burrard Street, St. Paul's Hospital, Vancouver, BC V6Z 1Y6
E-mail: mnorena@hivnet.ubc.ca

P. Dodek
Center for Health Evaluation and Outcome Sciences, 588 - 1081 Burrard Street, St. Paul's Hospital, Vancouver, BC, V6Z 1Y6
and
Division of Critical Care, Department of Medicine, Faculty of Medicine, The University of British Columbia, 855 W 12th Ave, Vancouver, BC, Canada V5Z 1M9
E-mail: peter.dodek@ubc.ca

1 Introduction

The World Health Organization declared COVID-19 a global pandemic on March 11th, 2020 [5]. Severe COVID-19 cases may involve critical conditions, including respiratory failure, that require mechanical ventilation for survival [22]. However, there is a limited number

of ventilators available which also need to be used by patients having other medical and surgical conditions unrelated to COVID-19. The potential surge in intensive care unit (ICU) and ventilator demand due to the pandemic heightens the importance of strategic medical resource management.

The importance of mathematical modeling to forecast patient numbers and needed critical care resource capacity during the COVID-19 pandemic has been widely recognized [31, 2, 6]. Mathematical models of disease transmission have been used to predict COVID-19 case counts [11, 27]. Nevertheless, additional modeling is needed to forecast medical resource utilization and inform operational decisions [6]. One approach to estimate ICU demand is to simply scale predicted case counts by the projected proportion of cases admitted to the ICU [27, 7]. However, critical care resource utilization also depends on non-COVID-19 demand, resource use time, available capacity, and the delay from COVID-19 symptom onset to critical care. Furthermore, most of these inputs are stochastic in nature. Queue modeling and discrete event simulation provide an accurate way to predict resource utilization by incorporating these factors [6].

Queue models play an important role in medical resource and ICU management [3]. The life-threatening conditions faced by many ICU patients motivate the common use of Erlang loss models or infinite server models, in both of which agents do not wait to receive service. In loss models, arriving patients are lost to the model if resources are unavailable, whereas in infinite server models resources are always available. Prior to the COVID-19 pandemic, McManus et al. [16] and Julio et al. [13] compared loss model results with historical ICU data and found that they accurately predicted transfer rates. In response to the COVID-19 pandemic, both loss and infinite server queue models have been used to inform critical care resource management, typically addressing different questions. Infinite server models have been used to predict the curve of COVID-19 ICU utilization, independent of available capacity. These models address required capacity based either on expected utilization [20, 17] or use simulation to capture stochasticity in resource utilization [9, 24, 25]. On the other hand, COVID-19 ICU loss models work within a finite resource capacity and are able to explore the relationship between capacity, utilization, and loss probability [30, 1]. Loss probability, or the probability that a patient is unable to access a medical resource when needed, is an important key performance indicator (KPI) that loss models are able to capture accurately. However, to our knowledge loss models have not been used to determine the COVID-19 critical care capacity required to meet a loss probability KPI. In non-epidemic scenarios, extended or generalized loss models have been used to help determine the required number of ICU beds [15, 23, 32] and hospital ward beds [4]. However, the rapid changes in arrival rates of COVID-19 patients present unique challenges to capacity planning using loss models.

The main contribution of this study is the application of a generalized Erlang loss model to address the interaction between ventilator capacity and utilization, as well as identify the capacity required to meet both COVID-19 and non-COVID-19 demand based on a loss probability KPI. We implemented our model using discrete event simulation (DES) to forecast time-dependent ventilator utilization. Simulation results predict when capacity would be reached and the rate at which patients would be unable to access a ventilator. Our approach frames required ventilator capacity as the minimum number of ventilators needed to maintain access targets given by a loss probability threshold. Additionally, our model incorporates a stochastic delay from COVID-19 symptom onset to mechanical ventilation, which is important for translating rapidly changing epidemic case projections into ventilator demand. To our knowledge, no other COVID-19 critical care queue models explicitly capture this stochastic delay.

This paper presents results used to inform ventilator management at a provincial level in British Columbia (BC), Canada. We calibrated the model using local ICU data in order to predict ventilator utilization in BC based on different epidemic projections. We incorporated projections from the BC Centre for Disease Control (BC CDC) of COVID-19 case counts under different levels of transmission due to social distancing and other public health measures. The results demonstrate the ability of our model to link public health interventions to ventilator utilization.

Our general queue model is described in Section 2. Model analysis, including simulation and capacity optimization, is described in Section 3. In Section 4, we describe BC specific model calibration and epidemic projections, which were used to produce the ventilator utilization forecasts presented in Section 5.

2 Queue model

We present a two stage queuing system with two submodels. The core submodel is a generalized Erlang loss model for ventilator use by COVID-19 and non-COVID-19 ICU patients. For COVID-19 patients who require mechanical ventilation, we use a combined queuing system in which an initial delay queue represents the time

from symptom onset to ventilation. The complete queuing system for COVID-19 and non-COVID-19 patients is depicted in Figure 1.

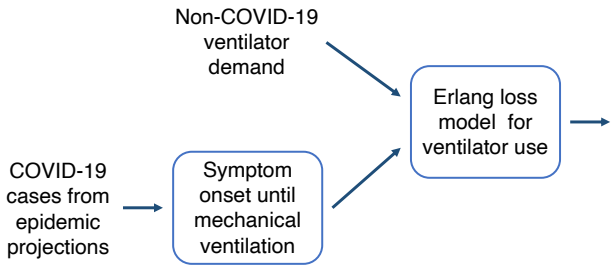

Fig. 1 Diagram of the two-stage queue model for delay from symptom onset and ventilator use.

The ventilator use submodel has the form of a generalized $./G/c/c$ Erlang loss model with a limited supply of c ventilators, where ventilation time has different distributions for COVID-19 and non-COVID-19 patients. If all c ventilators are in use, then arriving patients needing mechanical ventilation are lost to the system, which is motivated by the life-threatening nature of respiratory failure. Non-COVID-19 patients arrive and require a ventilator based on a non-homogeneous Poisson process, which could capture time-dependence in ICU admission rates or COVID-19 responses such as a reduction in elective surgery. For COVID-19 patients needing mechanical ventilation, there is a stochastic time between the onset of COVID-19 symptoms and the presentation of severe symptoms requiring a ventilator.

Epidemiological case projections are translated into ventilator demand using an initial $M_t/G/\infty$ submodel for the time from COVID-19 symptom onset to mechanical ventilation. This infinite server system simply implements a generally distributed stochastic delay and does not correspond to utilization of any physical resource. The arrival process for this submodel is non-homogeneous Poisson and can be based on localized COVID-19 epidemic projections, scaled by the projected proportion of critical patients. The output process is the effective COVID-19 ventilator demand.

3 Model analysis

We implemented our combined queuing model using discrete event simulation to project ventilator utilization based on epidemic projections. Model simulation can demonstrate whether the ventilator supply would be sufficient, estimate when capacity would be reached, and what the rate of patients unable to access a ventilator would be. Additionally, comparing multiple epidemic scenarios with our model can show the effect of different public health policies on ventilator utilization. Subsection 3.1 describes how we implemented a DES of the model and analyzed the results.

We used our model simulation to identify the minimum number of ventilators needed to meet a loss probability KPI target. Our approach to this capacity optimization problem uses a modified Response Surface Methodology (RSM) search procedure initialized at a pointwise-stationary approximation (PSA) for the required number of ventilators. Our methodology is described in Subsection 3.2.

3.1 Discrete event simulation

We built a discrete event simulation of our queue model using the AnyLogic modeling software. The simulation has two source nodes, one each for the COVID-19 and non-COVID-19 patient streams. Each source node generates patients according to a non-homogeneous Poisson process obtained by thinning a homogeneous Poisson process with a time-dependent probability of acceptance. COVID-19 patients are generated based on epidemic case projections scaled by the projected proportion of cases requiring a ventilator. An additional delay block implements the stochastic time between COVID-19 onset and the need for mechanical ventilation. The ventilator use submodel is implemented in a single service block, which sets the ventilation time distribution based on whether a patient has COVID-19. Section 4 describes how we chose model parameters for the epidemic in BC.

The simulation incorporates a "warm-up" period before the start of the community spread of COVID-19. At midnight of each simulated day, we measured the number of ventilators in use and whether the system is currently at capacity. We calculated the mean and variance of these measures across 1000 simulation runs. This yielded a regular time series of point estimates, with confidence intervals, for both the time-dependent expected ventilator use and the time-dependent probability that the system is at full capacity. The combined arrivals for the ventilator use submodel follow a non-homogeneous Poisson process (see Subsection 3.2.1). Therefore, the PASTA theorem for non-homogeneous Poisson processes [29] implies that the time-dependent probability that the system is at capacity equals the time-dependent probability that an arriving patient is unable to access a ventilator. Through this analysis, we

used DES to estimate a time-dependent loss probability KPI and forecast ventilator access for critical patients.

3.2 Identifying required ventilator capacity

We designed a capacity optimization approach to determine the minimum number of ventilators required to maintain a threshold on the time-dependent loss probability for projected COVID-19 and non-COVID-19 demand. Medical resource loss model studies in non-epidemic scenarios have determined required capacity based on the application of steady-state formulas [4] or linear simulation searches [23, 32]. However, in our context, the rapidly changing rate of COVID-19 cases makes applying these approaches less accurate and inefficient. In order to address the unique challenges of capacity optimization in a pandemic, we developed an approach that combines and modifies several existing techniques. The core of our approach is an iterative search procedure to find the minimum number of ventilators required to keep the estimated loss probability from the simulation under 5%. Our search procedure uses a stochastic root-finding method based on response surface methodology (RSM), which we describe in Subsection 3.2.2 and Appendix A. Furthermore, to increase the efficiency of the search, we used an initial proximate required capacity from a pointwise stationary approximation (PSA) described in Subsection 3.2.1.

3.2.1 Proximate required capacity

Our capacity optimization search is initialized at a proximate number of ventilators required to maintain loss probability below 5% in the PSA solution of the model. Pointwise stationary approximations (PSA) obtain proximate expressions for non-stationary system properties, by substituting time-dependent arrival rates into steady-state formulae [10]. We used a PSA to approximate the maximum time-dependent loss probability in the ventilator use submodel as a function of ventilator capacity, by substituting the maximum offered load into a stationary loss formula. A proximate number of required ventilators was found using a linear search until the PSA maximum loss probability is under 5%.

The generalized $M_t/G/c/c$ ventilator use submodel is not a Markovian system, because service time is general and not necessarily exponential. Nonetheless, Takács [26] showed that Erlang's B formula for the steady-state loss probability of an $M/M/c/c$ model also holds for the $M/G/c/c$ model. Thus, the PSA proximate loss probability of the ventilator submodel, with c ventilators and time-dependent offered load $a(t)$, is given by

$$B\big(c, a(t)\big) = \frac{a(t)^c/c!}{\sum_{j=0}^{c} a(t)^j/j!}\,. \tag{1}$$

For ease of calculation, we used the equivalent recursive definition [18]

$$B\big(c+1, a(t)\big)^{-1} = 1 + \frac{c+1}{a(t)} B\big(c, a(t)\big)^{-1}\,. \tag{2}$$

Erlang's B formula is an increasing function of $a(t)$ and a decreasing function of c.

In our model, ventilator demand is a non-homogeneous Poisson process which combines non-COVID-19 demand with the effective COVID-19 demand given by the symptom delay submodel output. Ventilation time follows a mixture of two patient-type distributions, with a time-dependent mixing rate based on the proportion of COVID-19 and non-COVID-19 patients. The PSA proximate time-dependent offered load can be expressed as

$$a(t) = \frac{\lambda_2(t) + \lambda_3(t)}{\frac{\lambda_2(t)\mu_2}{\lambda_2(t)+\lambda_3(t)} + \frac{\lambda_3(t)\mu_3}{\lambda_2(t)+\lambda_3(t)}}\,, \tag{3}$$

where $1/\mu_2$ and and $1/\mu_3$ are the mean ventilation times for COVID-19 and non-COVID-19 patients, respectively. Similarly, the time-dependent ventilator demand rates are λ_2 and λ_3 for COVID-19 and non-COVID-19 patients.

The effective COVID-19 demand λ_2 is the output rate of the symptom delay submodel. Since individuals in infinite server models do not affect each other [8], the output process of the symptom delay submodel is a non-homogeneous Poisson process, with rate given by the convolution

$$\lambda_2(t) = E(\lambda_1(t-S)) = \int_0^\infty \lambda_1(t-s) f(s) ds\,, \tag{4}$$

where λ_1 is the arrival rate for the symptom delay submodel, with random delay variable S that has probability density function f. We used equation 4 to calculate the effective, time-dependent COVID-19 demand for ventilators. In our application, we modeled symptom delay time with a uniform distribution between 7 and 12 days. Therefore, the COVID-19 need for ventilators on a given day is the average case count from 7 to 12 days prior (scaled by the proportion of cases needing ventilation). We substituted our calculated values of $\lambda_2(t)$ into equation 3, along with estimates for non-COVID-19 demand and mean ventilation times, in order to identify the maximum time-dependent offered load. Erlang's B formula (equation 2) then provides the approximate maximum loss probability as a function of ventilator capacity.

3.2.2 Capacity optimisation search

Our capacity optimization problem is to minimize the number of ventilators subject to a constraint of 5% on model loss probability. The maximum time-dependent loss probability of our ventilator model is a non-linear and strictly decreasing function of the number of ventilators, which we refer to as the 'response function'. However, the response function can only be estimated through simulation. Our capacity optimization problem can be expressed as a discrete version of a stochastic root finding problem on the response function. Standard stochastic root finding schemes focus on adapting deterministic single point methods, for example adapting Newton's method by approximating the derivative at each value [21]. In our case, since we can only evaluate our response function for integer capacity values, derivative approximations are inaccurate and adaptions of mutli-point root finding procedures are more desirable. Since the response function is non-linear, we guided our search procedure with iterative second-order approximations and utilized a modified response surface methodology (RSM) framework.

The RSM framework uses iterative first and second order approximations of a response function sampled via simulation or random experiment to guide an optimization search [19, 14]. Our approach is based on the RSM framework in Nicolai and Dekker [19]; however, we made several modifications based on our problem context. We only used second order approximations, in order to accurately capture the asymptotic behavior of the true response function for large numbers of ventilators. We used the intersection points of each approximated response function with the targeted loss probability threshold to update the estimated required number of ventilators and the region of interest. The response function is asymptotic to zero for a large number of ventilators. Therefore, our procedure includes an additional check that increases the size of the region of interest if the simulation-based estimate of the response function is zero.

Starting at an initial value given by the proximate PSA solution, we ran our modified RSM procedure with 200 simulation runs per evaluation, until the width of the region of interest reached below 10 ventilators. We then evaluated the maximum loss probability for each capacity value in the final region with 2000 simulation runs to obtain a plot from which the required ventilator capacity could be easily and accurately identified. Furthermore, such a plot provides an indication of the sensitivity of the loss rate to the number of ventilators. This can be useful for gauging the risk associated with fluctuations in the number of operable ventilators. Appendix A discusses the algorithmic details of our modified RSM and linear search approach.

4 Data analysis and case projections

We analyzed data on critical care utilization in BC, in order to apply our model to the BC context. The primary data set used was an extract from the BC ICU Database. We supplemented this with summary data provided by the BC Ministry of Health from the Discharge Abstract Database, published reports, expert opinion, and data from the Provincial Health Services Authority on ventilator capacity. The BC Centre for Disease Control provided case projections for the COVID-19 epidemic in BC as input for the model.

The British Columbia ICU Database was established in 1998 at the Centre for Health Evaluation & Outcome Sciences to provide detailed information on the delivery of critical care in British Columbia [28]. Our data extract consists of records from calendar years 2016–2018. For these years, the database contains ICU data from 20 hospitals in BC, including nearly all major hospitals. However, an additional 21 hospitals in BC with ICUs are not included in our extract. For these hospitals, we used ICU admission data from the Discharge Abstract Database, which is a national database of hospital admissions in Canada. We estimated ventilator utilization for these hospitals by assuming that they have the same fraction of ICU admissions requiring mechanical ventilation as the ICU Database hospitals. Our ICU Database extract includes an entry for each instance of mechanical ventilation. It has fields for the start and stop time of mechanical ventilation, acute respiratory distress syndrome (ARDS) diagnosis, and viral pneumonia diagnosis. We included these diagnoses in our extract, because they are clinically similar to respiratory failure due to COVID-19. Our extract does not include records from the year 2019, because the ICU Database did not have complete 2019 data at the time of this work.

4.1 Non-COVID-19 demand and ventilator capacity

We estimated the rate at which patients required a ventilator prior to the COVID-19 pandemic from the ventilation start frequency in our extract of the BC ICU Database. Patients may have multiple ventilation periods during a single ICU stay; however, we treated these as independent starts. We focused on 2017–2018 data as it was more relevant to our 2020 predictions. Figure 2 shows a monthly time series of mechanical ventilation starts from 2017–2018, which does not demonstrate a

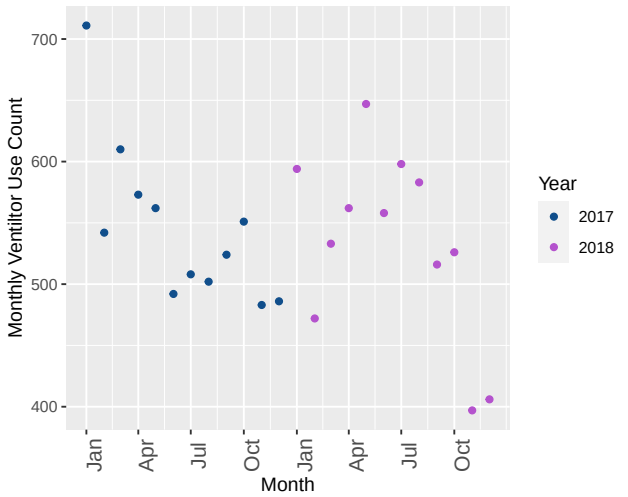

Fig. 2 Monthly time series of mechanical ventilation instances from the ICU Database.

clear trend or seasonality. The average number of ventilator starts per year in 2017–2018 of our extract was 6468, which we multiplied by 1.198 to account for the additional 21 hospitals not in the BC ICU Database. Therefore, we used a constant ventilator demand rate of $\lambda_3 = 21.24$ patients per day for non-COVID-19 patients, prior to the reduction in elective surgeries.

The BC Ministry of Health responded to the COVID-19 pandemic by canceling non-urgent elective surgeries in March, 2020. Detailed data on the impact of this reduction was unavailable at the time of this study. Based on expert opinion, we estimated that this led to a 15% reduction in the number of non-COVID-19 patients requiring mechanical ventilation. We implemented this change as a step reduction in λ_3 by 15% from March 16th, 2020 onwards.

The Provincial Health Services Authority of BC conducted an inventory of ventilators in the province in March, 2020. There were 498 adult ventilators available in 34 hospitals. We were advised that at any given time, approximately 10% of ventilators would be unavailable due to repair or maintenance. Therefore, we set the number of ventilators in the model to 448. Based on consultation with a respiratory therapist, we assumed that the time required to clean a ventilator and prepare it for a new patient is approximately two hours, which we incorporated into the ventilation service time.

4.2 Ventilation time and symptom delay

We characterized the duration of time that patients receive ventilation using start and stop times from the 2017–2018 records of our BC ICU Database extract. Approximately 10.4% of these records were either missing a start/stop time or had zero ventilation time, which we did not incorporate in ventilation time analysis. We divided the remaining records into two groups: one for patients diagnosed with either ARDS or viral pneumonia, and one for patients with neither of these diagnoses. At the time of this analysis, there was very little data for time on ventilation for COVID-19 patients. ARDS and viral pneumonia are clinically similar to patients with respiratory failure due to COVID-19 and therefore we assumed that the ventilation times for patents with these diagnoses are representative of ventilation times for COVID-19 patients.

Table 1 Mixed gamma distribution parameters for time spent on mechanical ventilation.

	Viral Pneumonia or ARDS	Neither Viral Pneumonia nor ARDS
Proportion	6.2%	93.8%
Shape	0.94	0.85
Scale	7.9 days	4.8 days
Mean	7.5 days	4.1 days

For each group of records, we fit ventilation time to a gamma distribution using the maximum-likelihood method implemented in the 'MASS' package in R. The parameters for the distribution fits are given in Table 1 and the distributions are plotted in Figure 3. The ventilation time distributions appear to be reasonably approximated by gamma distributions, and they are substantially different for the two categories of patients. The mean ventilation time for patients with viral pneumonia or ARDS is substantially greater than the mean ventilation time for patients with neither diagnosis.

BC data for the distribution of time between COVID-19 symptom onset and ventilation were not available at the time of this analysis. A study by Phua et al. [22] estimated the median time from symptom onset to ICU admission as 7–12 days. Therefore, we assumed a uniform distribution between 7 and 12 days for our symptom delay model.

4.3 COVID-19 case projections

The BC Centre for Disease Control provided COVID-19 case projections using a stochastic disease model[1] based on Hellewell et al. [11]. Their epidemiological model was calibrated using historical data on cases in BC and can forecast cases under different scenarios by

[1] The BC CDC model is implemented as an R package, which is available from `https://github.com/bcgov/epi.branch.sim`.

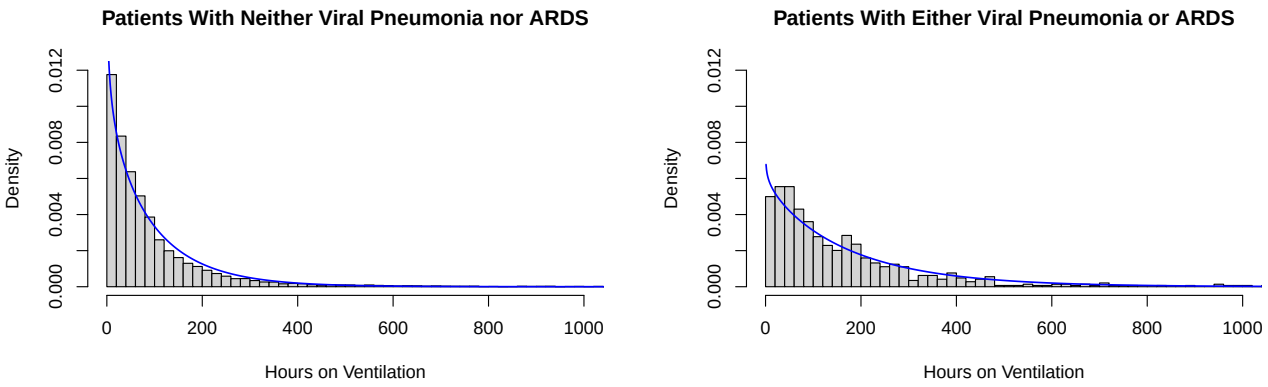

Fig. 3 Density histogram of the time on mechanical ventilation. The gamma distribution fits are plotted as a curve.

varying a transmission rate parameter. Scenarios of reduced transmission are specifically used to illustrate the impact of public health measures such as social distancing and other changes in population behavior. Significant public health measures in BC during the first wave of the epidemic began on March 16th and included restricting gatherings to no more than 50 people and closing of schools.[2] Their model was re-calibrated weekly in March and April 2020, with forecasts extending one month from calibration date. The results in our paper are based on projections released on March 19th 2020, which forecast COVID-19 case counts from March 16th to April 13th.

Based on expert opinion and initial data on the epidemic in BC, we assumed that 6.7% of COVID-19 cases would require ICU care. We further assumed that 70% of COVID-19 ICU patients would require mechanical ventilation, based on the ICNARC Report on COVID-19 (10 April 2020) [12] and expert opinion on hospitalizations in BC. We multiplied the BC CDC projections by the above two proportions, in order to estimate the time-dependent rate of COVID-19 symptom onset for cases which will eventually require a ventilator. We used these scaled projected daily case counts as daily constant arrival rates in a non-homogeneous Poisson arrival process for the symptom delay submodel.

5 Results

Subsection 5.1 presents the validation results for our DES implimentation and Subsection 5.2 presents the simulation results for epidemic projections with and without reduced transmission. Subsection 5.3 presents the results of our capacity optimization search procedure.

5.1 Validation

We validated our DES implementation by comparing mean ventilator utilization in the simulation prior to COVID-19 with a Little's law expected value calculation for the queue model. Furthermore, we compared both of these results with pre-COVID-19 ICU data to demonstrate the soundness of our modeling assumptions for non-COVID-19 ventilator demand. Our validation was based on the 20 hospitals in the BC ICU Database and not on all hospitals BC. These hospitals are estimated to have 356 functional adult ventilators.

To validate our DES implementation, we first computed the expected number of ventilators in use in the non-COVID-19 component of our queue model using Little's law, with the mean ventilation time and the non-COVID-19 ventilator demand rate estimated from the 2017–2018 BC ICU Database extract. The loss model adjustment term to Little's law is negligible in this calculation since the number of ventilators is large compared to the offered load. We also ran our DES implementation for two simulated years with only non-COVID-19 demand, using a constant rate set to the average ventilation starts from the 2017–2018 BC ICU Database extract. We measured the number of patients on a ventilator at the end of each simulation run and calculated the mean and standard deviation of this measurement across 2000 simulation runs. The first two columns of Table 2 compare the Little's law expected utilization and the simulation mean, and show that are not significantly different.

Using our BC ICU Database extract, we inferred the number of patients on a ventilator for each day in

[2] Additional details on public health measures undertaken in BC are available at `https://news.gov.bc.ca/releases/2020HLTH0086-000499`.

Table 2 Pre-COVID-19 mean utilization comparison. Simulation mean ventilator use is estimated from the number of patients on a ventilator at last simulated time point, across 2000 DES runs. The theoretical model mean is calculated using Little's Law for the expected number of patients in the queue. The historic mean is a time average of the inferred number of patients on a ventilator during March and April 2018.

	Simulation	Little's Law	Historic
Mean	77.50	77.36	77.70
Standard Error	0.20		

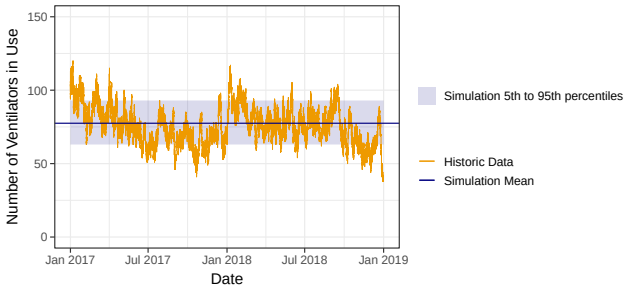

Fig. 4 Comparison of pre-COVID simulation and inferred historic ventilator use. The mean and 5th and 95th percentiles of simulation ventilator use are measured from 2000 runs.

2017 and 2018 from the recorded start and stop times of ventilation. For the 10.4% of records without a recorded stop time, we assumed a stop time proxy equal to the start time plus the annual mean ventilation time. Figure 4 compares the inferred historic ventilator use with the mean, 5th, and 95th percentiles from the simulation. The historic data is within the percentile range 75.3% of the time. For March and April 2018, the time average ventilator use shown in the third column of Table 2 is not significantly different from the simulation mean. Thus, for the purpose of predicting non-COVID-19 ventilator demand in March and April 2020 without a reduction in elective surgery, a homogeneous Poisson arrival process with average demand is a reasonable choice.

5.2 Model simulation

Model simulation results for the March 19th 2020 case projections from the BC CDC are shown in Figure 5, under scenarios with and without reduced transmission. The March 19th projections start on March 16th and project COVID-19 case counts until April 13th. We ran the simulation model until April 20th, because our delay between symptom onset and ventilation has a minimum of 7 days. The effect of the cancellation of non-urgent elective surgeries on March 16th is noticeable in both scenarios as a slight decrease in ventilator utilization, before an increase occurs due to projected rising COVID-19 cases. Without reduced transmission, the projected number of COVID-19 cases requiring a ventilator reaches approximately 112 patients per day by the end of the projection. In this scenario, the estimated probability of reaching ventilator capacity is negligible until approximately April 14th, when it begins to rise dramatically. By the end of the simulation, the mean number of patients unable to access a ventilator reaches approximately 50 per day. However, with reduced transmission (through public health measures including social distancing), the projected rate of COVID-19 ventilator cases reaches a substantially lower rate of approximately 36 patients per day. In this scenario, the estimated probability of reaching ventilator capacity remains negligible and all the simulated patients are able to access a ventilator.

5.3 Capacity optimization

We developed a combined search procedure to find the ventilator capacity required to meet demand in the epidemic scenario without reduced transmission. Our approach combines PSA, RSM, and linear search techniques to identify the minimum number of ventilators required to maintain a loss probability under 5%. In the first stage of our procedure, we identified that at least 818 ventilators are needed to meet the access target in a PSA proxy for the time-dependent model solution. We used this proximate capacity as an initial value for a modified RSM procedure which is guided by iterative approximations of maximum loss probability as a function of the number of ventilators, estimated using 200 DES runs at strategically chosen evaluation values. Using 6 search iterations of RSM, we identified that the required capacity is between 623 and 633 ventilators. We simulated the values in this range using 2000 runs per evaluation, the results of which are displayed in Figure 6. From this plot we identified that at least 629 ventilators are required to keep the estimated loss probability KPI within 5%. A reduction in this capacity of up to 5 ventilators will keep the loss probability within 6%; however, once the capacity has been reduced to 623 ventilators the estimated loss probability rises sharply to 8%.

Figures 7 and 8 compare the PSA and simulation ventilator utilization and loss probability for 629 ventilators. The PSA estimate of ventilator utilization increases more quickly than the simulation results, because it assumes instantaneous response towards stationarity with changes in arrival rate.

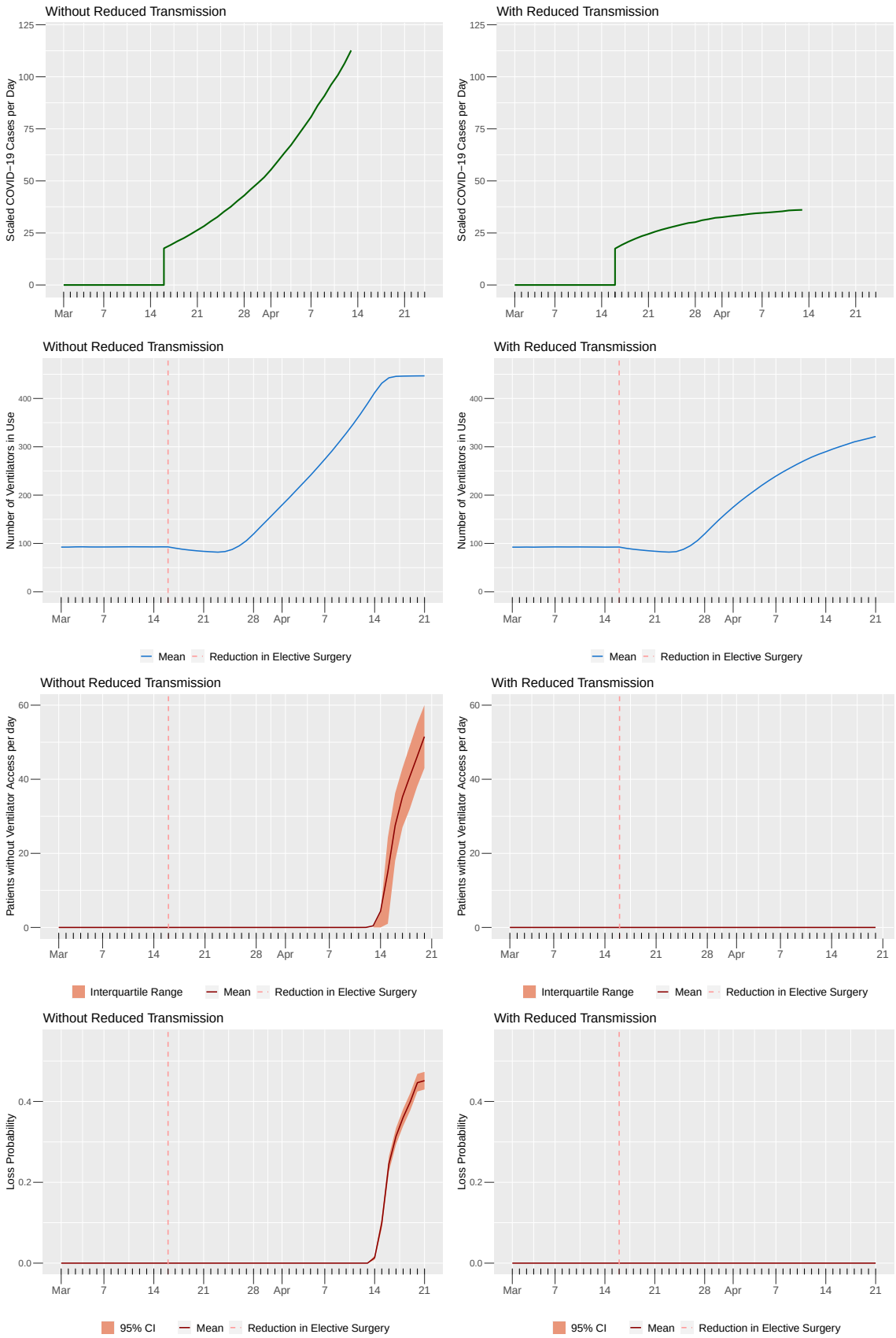

Fig. 5 Epidemic projections and simulation results for scenarios with and without reduced transmission due to public health measures including social distancing, using 448 ventilators. The topmost panels present epidemic projections of daily COVID-19 cases requiring a ventilator, from March 16th to April 13th 2020. Historical cases are not plotted. The second pair of panels show estimated expected ventilator utilization over time from simulations based on the epidemic projections. The third and fourth pairs of panels show estimated expected rates and probabilities of patients unable to access a ventilator. The dotted line represents the start of both the projections and the reduction in elective surgeries. All of the simulation results are obtained using 1000 runs.

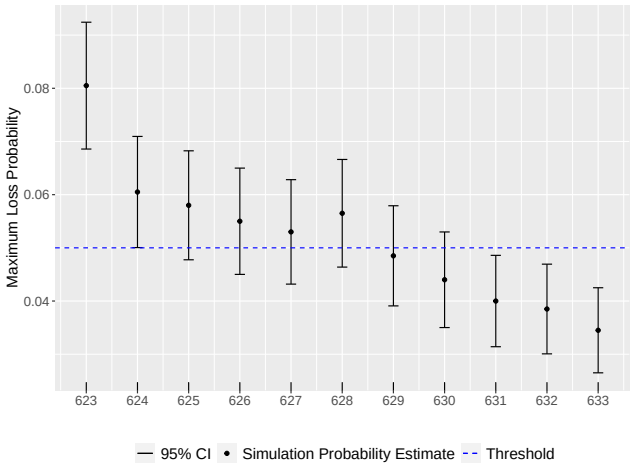

Fig. 6 Linear search across the final region of ventilator capacities identified as near optimal by the modified RSM search procedure. For each number of ventilators, the maximum loss probability is estimated using 2000 simulation runs.

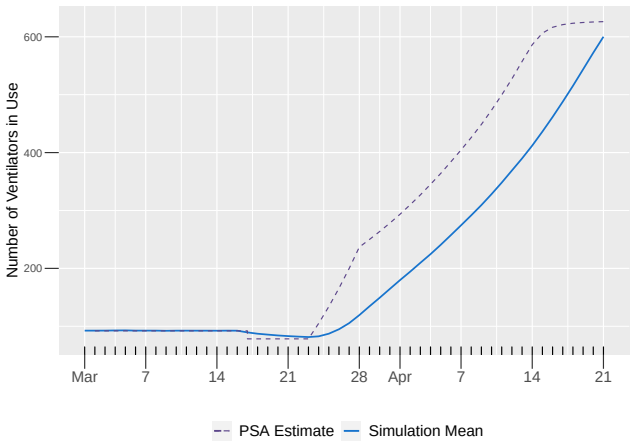

Fig. 7 Comparison of mean ventilator utilisation from 2000 simulation runs with the PSA expected utilisation, for 629 ventilators.

6 Discussion

We developed a loss model based queuing system to inform ventilator capacity management during the COVID-19 pandemic. Our model incorporates projections of COVID-19 cases, the proportion of cases needing ventilators, the delay from symptom onset to mechanical ventilation, non-COVID-19 demand, ventilation time, and ventilator capacity. Using simulation, we predicted ventilator utilization under different epidemic scenarios, which informed whether the available ventilator capacity would be sufficient to meet demand and when capacity would be reached. We used the loss model framework to address capacity optimization in terms of a performance indicator given by the time-dependent probability of a patient being able to access

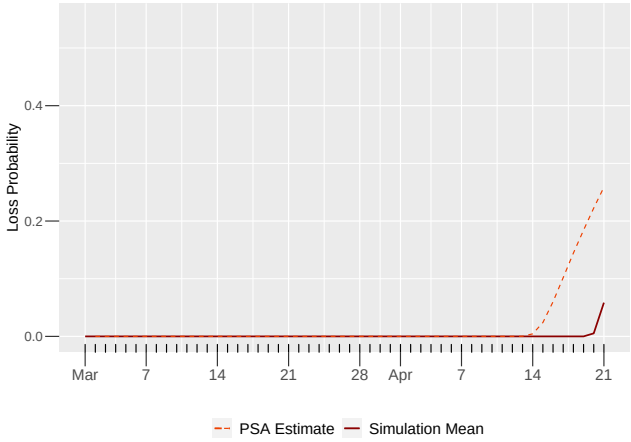

Fig. 8 Comparison of estimated loss probability from 2000 simulation runs with the PSA loss probability, for 629 ventilators.

a ventilator. To navigate the computational challenges posed by capacity optimization in a pandemic, we developed a stochastic search procedure to efficiently find the required number of ventilators when patient demand is changing rapidly.

We worked closely with analysts at the BC Ministry of Health, the Provincial Health Services Authority, and the BC Centre for Disease Control in March and April 2020, to provide weekly reports on projected ventilator utilization during the first peak of the COVID-19 pandemic. We collaborated with these organizations to obtain up-to-date BC specific data and parameter values for our model, including weekly updated COVID-19 case projections provided by the BC CDC. The results presented in this paper are based on the BC CDC's March 19th epidemic projections under different levels of transmission. Simulation results predict that under a scenario of reduced transmission of COVID-19 (through means such as social distancing, but also other public health measures and changes in population behavior), ventilator capacity would likely not be reached, thereby helping to avert as many as 50 deaths per day within the time frame of the March 19th projections. The accuracy of our results are dependent on the underlying epidemic projections and need to be interpreted in the light of challenges faced in predicting the epidemic trajectory.

We addressed capacity optimization by using an iterative search procedure to identify the number of ventilators required to meet demand within access constraints. Under the March 19th BC CDC projections without reduced transmission, 181 additional ventilators would have been required to ensure that the probability of immediate patient access to a ventilator is at

least 95%. However, with public health interventions and social distancing, the current ventilator supply was sufficient.

Epidemic projections are frequently updated during the COVID-19 pandemic, and efficient model analysis is important for timely decision making support. Our capacity optimization uses loss model properties to find a proximate initial value for our simulation-based iterative search procedure. This combined approach increases the computational efficiency of our model analysis without compromising accuracy in the context of rapidly increasing COVID-19 rates. Future work could incorporate more sophisticated stochastic root finding or simulation optimization techniques for even greater computational efficiency.

The results in this paper demonstrate the ability of our model to address critical care resource management during the COVID-19 epidemic. Simulation can predict details of ventilator utilization within the ICU, including and if (and when) capacity will be reached and what the rate will be of patients unable to access a ventilator. These forecasts can be used by health system analysts and policy makers to inform capacity management decisions, including whether to expand the current ventilator supply. Capacity optimization can address how many ventilators are required to meet projected demand, based on access targets. Additionally, our model links public health interventions, including social distancing, to operational impacts on ventilator utilization. Our work provides a tool for policy makers to quantify the interplay between public health interventions, necessary critical care resources, and performance indicators for patient access.

Overall, our queuing model provides a valuable approach for translating epidemic curves into resource utilization and capacity optimization. It can be easily implemented using data for other regions and updated as more information about COVID-19 becomes available. It can also be used for projecting demand for other COVID-19 resources, such as ICU beds. Future work could extend our model to consider joint resource utilization, for example modeling the joint use of both ventilators and ICU beds.

Acknowledgements We are grateful to Sandra Feltham and the Hospital & Diagnostics Analytics Team at the BC Ministry of Health for supporting this project and providing summary statistics from the Discharge Abstract Database. We thank the epidemiological modeling team led by Michael Otterstater at the BC CDC for providing weekly COVID-19 case projections from their model; Ognjenka Djurdjev at the Provincial Health Service Authority for data on the number of ventilators available; and Lena Farina at St. Pauls Hospital, Vancouver, for expert opinion on ventilator maintenance and cleaning times.

This project was funded in part by a Partnership & Innovation Grant from the BC Ministry of Health.

References

1. Alban A, Chick SE, Dongelmans DA, Vlaar APJ, Sent D, Study Group (2020) ICU capacity management during the COVID-19 pandemic using a process simulation. Intensive Care Medicine pp 1–3, DOI 10.1007/s00134-020-06066-7
2. Aziz S, Arabi YM, Alhazzani W, Evans L, Citerio G, Fischkoff K, Salluh J, Meyfroidt G, Alshamsi F, Oczkowski S, Azoulay E, Price A, Burry L, Dzierba A, Benintende A, Morgan J, Grasselli G, Rhodes A, Møller MH, Chu L, Schwedhelm S, Lowe JJ, Bin D, Christian MD (2020) Managing ICU surge during the COVID-19 crisis: rapid guidelines. Intensive Care Medicine 46:1303–1325, DOI 10.1007/s00134-020-06092-5
3. Bai J, Fügener A, Schoenfelder J, Brunner JO (2018) Operations research in intensive care unit management: a literature review. Health Care Manag Sci 21(1):1–24, DOI 10.1007/s10729-016-9375-1
4. de Bruin AM, Bekker R, van Zanten L, Koole GM (2010) Dimensioning hospital wards using the Erlang loss model. Ann Oper Res 178:23–43, DOI 10.1007/s10479-009-0647-8
5. Cucinotta D, Vanelli M (2020) WHO Declares COVID-19 a pandemic. Acta Bio-medica : Atenei Parmensis 91:157–160, DOI 10.23750/abm.v91i1.9397
6. Currie CSM, Fowler JW, Kotiadis K, Monks T, Onggo BS, Robertson DA, Tako AA (2020) How simulation modelling can help reduce the impact of COVID-19. J Simul 14(2):83–97, DOI 10.1080/17477778.2020.1751570
7. Davies NG, Kucharski AJ, Eggo RM, Gimma A, Edmunds WJ (2020) Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. Lancet Public Health 5(7):e375–e385, DOI 10.1016/S2468-2667(20)30133-X
8. Eick SG, Massey WA, Whitt W (1993) The physics of the $M_t/G/\infty$ queue. Oper Res 41(4):731–742, DOI 10.1287/opre.41.4.731
9. Garcia-Vicuña D, Esparaza L, Mallor F (2020) Hospital preparedness in epidemics by using simulation. the case of COVID-19. medRxiv DOI 10.1101/2020.08.12.20173328
10. Green L, Kolesar P (1991) The pointwise stationary approximation for queues with nonstationary ar-

rivals. Manag Sci 37(1):84–97, DOI 10.1287/mnsc.37.1.84
11. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, Munday JD, Kucharski AJ, Edmunds WJ, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Funk S, Eggo RM (2020) Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Global Health 8(4):E448–E496, DOI 10.1016/S2214-109X(20)30074-7
12. ICNARC (2020) ICNARC report on COVID-19 in critical care 10 April 2020. Tech. rep., Intensive Care National Audit and Research Centre (ICNARC)
13. Julio B, Guergué JM, Esparza L, Azcárate C, Mallor F, Ochoa S (2012) A mathematical model for simulating daily bed occupancy in an intensive care unit. Critical Care Medicine 40(4):1098–1104, DOI 10.1097/CCM.0b013e3182374828
14. Kleijnen JPC (2015) Response surface methodology. In: Fu MC (ed) Handbook of Simulation Optimization, Springer, New York, chap 4, pp 81–104
15. Litvak N, van Rijsbergen M, Boucherie RJ, van Houdenhoven M (2008) Managing the overflow of intensive care patients. Eur J Oper Res 185(3):998–1010, DOI 10.1016/j.ejor.2006.08.021
16. McManus ML, Long MC, Cooper A, Litvak E (2004) Queuing theory accurately models the need for critical care resources. Anesthesiology 100(5):1271–1276, DOI 10.1097/00000542-200405000-00032
17. Meares HDD, Jones MP (2020) When a system breaks: queueing theory model of intensive care bed needs during the COVID19 pandemic. Medical Journal of Australia 212(10):470–471, DOI 10.5694/mja2.50605
18. Medhi J (2002) Stochastic Models in Queueing Theory, 2nd edn. Mathematics in science and engineering, Elsevier Science & Technology, San Diego
19. Nicolai R, Dekker R (2009) Automated response surface methodology for simulation optimization models with unknown variance. Quality Technology & Quantitative Management 6(3):325–352, DOI 10.1080/16843703.2009.11673203
20. Palomo S, Pender J, Massey W, Hampshire RC (2020) Flattening the curve: Insights from queueing theory. ArXiv abs/2004.09645v1
21. Pasupathy R, Kim S (2011) The stochastic root-finding problem: Overview, solutions, and open questions. ACM Trans Model Comput Simul 21(3), DOI 10.1145/1921598.1921603
22. Phua J, Weng L, Ling L, Egi M, Lim CM, Divatia JV, Shrestha BR, Arabi YM, Ng J, Gomersall CD, Nishimura M, Koh Y, Du B (2020) Intensive care management of coronavirus disease 2019 (COVID-19): Challenges and recommendations. Lancet Respiratory Medicine 8(5):P506–517, DOI 10.1016/S2213-2600(20)30161-2
23. Ridge JC, Jones S, Nielsen MS, Shahani AK (1998) Capacity planning for intensive care units. Eur J Oper Res 105(2):346–355, DOI 10.1016/S0377-2217(97)00240-3
24. Römele C, Neidel T, Heins J, Heider S, Otten V, Ebigbo A, Weber T, Müller M, Spring O, Braun G, Wittmann M, Schoenfelder J, Heller AR, Messmann H, Brunner JO (2020) Bed capacity management in times of the COVID-19 pandemic : A simulation-based prognosis of normal and intensive care beds using the descriptive data of the University Hospital Augsburg. Anaesthesist 69(10), DOI 10.1007/s00101-020-00830-6
25. Stang A, Stang M, Jöckel KH (2020) Estimated use of intensive care beds due to COVID-19 in Germany over time. Dtsch Arztebl Int 117:329–335, DOI 10.3238/arztebl.2020.0329
26. Takács L (1969) On Erlang's formula. Ann Math Stat 40(1):71–78, DOI 10.1214/aoms/1177697805
27. Weissman GE, Crane-Droesch A, Chivers C, Luong T, Hanish A, Levy MZ, Lubken J, Becker M, Draugelis ME, Anesi GL, Brennan PJ, Christie JD, Hanson III CW, Mikkelsen ME, Halpern SD (2020) Locally informed simulation to predict hospital capacity needs during the COVID-19 pandemic. Ann Intern Med 173(1):21–28, DOI 10.7326/M20-1260
28. Wenner JB, Norena M, Khan N, Palepu A, Ayas N, Wong H, Dodek P (2009) Reliability of intensive care unit admitting and comorbid diagnoses, race, elements of Acute Physiology and Chronic Health Evaluation II score, and predicted probability of mortality in an electronic intensive care unit database. J Crit Care 24(3):401–407, DOI 10.1016/j.jcrc.2009.03.008
29. Wolff RW (1982) Poisson arrivals see time averages. Oper Res 30(2):223–231, DOI 10.1287/opre.30.2.223
30. Wood RM, McWilliams CJ, Thomas MJ, Bourdeaux CP, Vasilakis C (2020) Covid-19 scenario modelling for the mitigation of capacity-dependent deaths in intensive care. Health Care Manag Sci 23:315–324, DOI 10.1007/s10729-020-09511-7
31. Xie J, Tong Z, Guan X, Du B, Qiu H, Slutsky AS (2020) Critical care crisis and some recommendations during the COVID19 epidemic in China. Intensive Care Med 46:837–840, DOI 10.1007/s00134-020-05979-7

32. Zhu Z, B HH, Teow KL (2012) Estimating ICU bed capacity using discrete event simulation. International Journal of Health Care Quality Assurance 25(2):134–144, DOI 10.1108/09526861211198290

Appendix A Response Surface Methodology

We used a modified response surface methodology (RSM) procedure to identify the minimum ventilator capacity required to maintain a threshold on the loss probability KPI. We used only second-order approximations of the response function (maximum model loss probability as a function of ventilator capacity). Our approach is based on the RSM framework in Nicolai and Dekker [19]; however, we made several modifications based on our problem context. This appendix describes the algorithmic details of our implementation; higher level discussion is in Subsection 3.2.2.

To describe our RSM approach, we denote the iterative estimates of the required ventilator capacity by c_i and the iterative radii of the successive regions of interest by r_i, for $i = 0, 1, 2, ...$ We begin with initial estimated capacity c_0 and radius $r_0 = \text{round}(0.1\, c_0)$. The target loss probability is denoted by α. The details of our implementation are as follows:

Iterate the following over $i = 0, 1, 2, ...,$ until $r_i \leq 5$:

(a) Estimate the maximum loss probability at each ventilator capacity value in a one dimensional central composite experimental design [19]. This design involves five estimates made at the center value c_i, and one estimate made at each of the values $c_i - r_i, \text{round}(c_i - 0.5\, r_i), \text{round}(c_i + 0.5\, r_i)$, and $c_i + r_i$. For each capacity value, the maximum loss probability is estimated using 200 simulation runs.

(b) Using these estimated points, perform a least-squares regression to fit a second-order model

$$y = \beta_0 + \beta_1 c + \beta_2 c^2,$$

where y is the maximum loss probability for c ventilators.

(c) If the estimated coefficients satisfy $\hat{\beta}_0 = \hat{\beta}_1 = \hat{\beta}_2 = 0$, then set $r_{i+1} = 2\, r_i$ and move to the next iteration.

(d) If the overall regression fit is statistically significant (F-statistic has p-value at most 0.05), then set

$$d = \hat{\beta}_1^2 - 4\, \hat{\beta}_2 (\hat{\beta}_0 - \alpha)$$

$$z_1 = \frac{-\hat{\beta}_1 - \sqrt{d}}{2\, \hat{\beta}_2}$$

$$z_2 = \frac{-\hat{\beta}_1 + \sqrt{d}}{2\, \hat{\beta}_2}$$

$$c_{i+1} = \begin{cases} z_1, & \text{if } d \geq 0 \text{ and } z_1 \geq 0 \text{ and } \hat{\beta}_2 \geq 0 \\ z_2, & \text{if } d \geq 0 \text{ and } \big(z_1 < 0 \text{ or } \hat{\beta}_2 < 0\big) \\ \frac{\hat{\beta}_1}{2\, \hat{\beta}_2}, & \text{if } d < 0 \end{cases}$$

$$c_{i+1} = \text{round}(c_{i+1})$$
$$r_{i+1} = \text{round}(0.5\, r_i),$$

and move to the next iteration.

(e) If the overall regression fit is not statistically significant (F-statistic has p-value greater than 0.05), then double the number of runs per evaluation until all are statistically significant.

We implemented our search procedure by iteratively using AnyLogic to perform simulation runs and R to analyze the results, perform regression fits, and determine the updated center values. Once the region of interest had width under 10, we ran a linear simulation search over the region with ten times the simulation runs, and plotted the results to identify the required ventilator capacity.