Ideally, no one would have to wait for access to a life-saving device, such as a ventilator. But during the coronavirus disease 2019 (COVID-19) crisis, many hospitals have run short of ventilators—as well as respiratory therapists (RTs) who are trained to operate them—while other facilities scramble to prepare for the coming demand surge. Any patient who needs a ventilator might not be able to get one, imperiling their survival. Mathematically, the delay that a patient may experience depends on the number of ventilators at the hospital, the number of patients who need them, and the average time that patients spend on them. COVID-19 increases the arrival rate of ventilator patients at hospitals, and these patients have longer lengths of stay, resulting in the sharp rise in demand for ventilators. Hospitals need to estimate how many ventilators they will need to respond to this crisis, accounting for increased, uncertain patient arrival rates and lengths of stay. At the same time, regional coordinators want to know which hospitals have the greatest need for the next spare ventilator or available capacity for the next ventilator patient.
To address the ventilator allocation problem at both of these levels, we apply an Erlang C queuing model. This model addresses the randomness of ventilator patient arrivals and lengths of stay. We use this model to calculate the number of ventilators and RTs needed to achieve a target wait time—the average delay for a ventilator experienced by a new patient. The target wait time corresponds with a probability that any wait is experienced: If the average wait is small, most patients would in fact experience no wait at all. Given hospital-level inputs, our model estimates the required ventilator and RT resources needed to achieve a target average expected wait time set by the user. At the regional level, planners can use this model to assess needs, allocate patients or resources efficiently across hospitals, or drive hospital protocol decisions. Many states are developing guidelines for ventilator management in pandemics (e.g., Angell, 2020). This model could aid in that management.

While strongly supported by other research and theory, our model has not been validated in a real hospital or regional health planner setting; the fast-moving nature of this crisis precluded this testing at the time of writing. Properly implemented, we believe this model can save lives. Users interacting with it could estimate or vary parameters to examine current conditions, study the implications of disease forecasts, or conduct system stress tests.

The Problem of Allocating Ventilators and Respiratory Therapists in the COVID-19 Crisis

Early studies of outcomes for COVID-19 patients in China describe a great need for ventilators (Yang et al., 2020). Analysis of COVID-19 patient data from Italy demonstrates the value of early intubation in a prone position for improving patient survival (Davenport, 2020). As the pandemic progresses, new studies and protocols are being developed that may result in a decrease of early ventilation (see Meng et al., 2020), but presently, critical care COVID-19 patients require an average of 10–11 days on ventilation (Li et al., 2020). Based on the anticipated number of patients requiring ventilators and the length of time that patients with COVID-19 stay on ventilators, experts anticipated a shortage of ventilators across the United States (Choi and Velasquez, 2020). Such shortages could require medical professionals to make hard decisions about who gets a ventilator and who must wait for one (New York State Task Force on Life and the Law, 2015; White and Lo, 2020).

Regional coordinators need to know how to allocate available ventilators among hospitals. At the federal level, the U.S. Strategic National Stockpile and the U.S. Departments of Defense and Homeland Security store ventilators. States and hospitals may also have ventilators in reserve, and suppliers are gearing up to produce thousands more. A recent New York Times article suggests that federal-level pandemic planning exercises identified the need for improved coordination in the distribution of essential equipment and supplies (Sanger et al., 2020). Our model can help with the development of that strategy.
The planning and execution in response to the pandemic requires knowing how many patients will need ventilators in each state, county, and hospital; the locations of spare ventilators; and a transparent, ethical allocation methodology that can be deployed repeatedly as the crisis evolves.

Ventilators are complex machines requiring operation by trained RTs. Even before the outbreak, RTs were in short supply in the United States. Limited availability of RTs may strain the management of COVID-19 patients requiring mechanical ventilation (Shapiro, 2020). The Inspector General of the U.S. Department of Health and Human Services reports, “Hospitals anticipated that ventilator shortages would pose difficult decisions about ethical allocation and liability, although at the time of our survey no hospital reported limiting ventilator use” (Grimm, 2020). Needs will shift as outbreaks develop across the country, and careful resource allocation and efficient use will be critical to saving lives.

In recommendations based on their COVID-19 experiences, Chinese researchers called for better models of resource needs, noting that “it would be very helpful to have mathematical models developed which predict the expected number of patients, and the necessary resources (equipment and personnel) required to treat these patients” (Xie et al., 2020). We note that queuing theory provides an appropriate framework for this problem. Just as a call center experiencing unusually high call volume or unexpectedly longer calls leaves customers waiting longer, a hospital facing an increase in patient arrival rate or lengths of stay will have a higher risk of patients needing to wait for a ventilator to become available. Another familiar example of queuing is the grocery store check-out line: Like customers restricted to a certain number of check-out lines who need to be processed for some amount of time by a cashier, patients arrive at a hospital needing a ventilator from a limited number available and having to spend some amount of time on it.

Our model treats both the number of ventilator patient arrivals per day and their length of stay on the ventilator as random events to capture the unpredictability of the process. Planning ventilator requirements without accounting for this uncertainty may have tragic results. Consider that the arrival rate will exceed its average about half the time, resulting in exploding queue lengths if not accounted for ahead of time. Applying queuing theory can help hospital planners and regional planners understand how many ventilators and RTs they will need in a way that anticipates and probabilistically manages this scenario.

Other researchers and practitioners have applied queuing theory to hospital management. Khan and Callahan (1993) uses queuing theory to plan capacity in a hospital

Although we are confident in the validity of the queuing model, we cannot recommend a target wait time or probability of wait for any hospital: Different hospitals have different procedures, and different patients have different needs. Nonetheless, hospitals might examine the number of ventilators required to ensure that some percentage of patients do not wait at all (e.g., 95 percent or 99 percent) in order to characterize their shortage. A regional coordinator would have to choose a target wait time or probability of waiting in discussion with the regional hospitals and in light of resource constraints (see the section titled “Model Inputs” for a detailed discussion of target wait time).

**Other Models for COVID-19 Capacity Planning**

The current COVID-19 pandemic has placed enormous strain on hospitals and health care systems. The critical question is, will we have enough resources to care for all who become ill? To help answer that question, many ongoing initiatives are currently looking at resource allocation and health care capacity in this pandemic with the intent of helping plan for future pandemics. This section provides a brief review of some of the most relevant work in this area to put our model and its contributions in context.

**Cornell COVID Caseload Calculator (Hupert et al., undated):** A model of demand for health care capacity, this tool uses deterministic simulations and mathematical calculations to estimate the arrival rate of COVID-19 patients and then simulates the load that those patients place on regular and critical care beds.

**RAND Corporation—Critical Care Surge Response Tool (Abir et al., 2020):** This report describes an Excel-based tool to estimate critical care capacity at various levels—hospitals, health care systems, states, and regions. It is a deterministic model that allows the user to input various characteristics—number of doctors, RTs, ventilators, beds, and patient-to-staff ratios. The output is the number of patients who can be cared for per shift.

**RAND Corporation—Mechanism to Reduce Medical Supply Shortfalls During Pandemics (Hastings Roer and Globus-Harris, 2020):** This is not a model but a proposed construct for increasing the value of centralized stockpiles of medical supplies, such as the U.S. Strategic National Stockpile. By implementing a backstopping mechanism that would incentivize the movement of resources from low-impacted areas to areas of greater need, planners can potentially reduce resource shortfalls and increase resource efficiency.

**Stanford University—Hospital Bed Projections (Ferstad et al., 2020):** The model is an exponential growth model of COVID-19 patients requiring hospital beds, ICU
beds, and ventilators. This demand is compared with hospital capacity parameters to estimate when capacity will be exhausted.

University of Pennsylvania—COVID-19 Hospital Impact Model (CHIME, Weissman et al., 2020): This is a hospital-level model that projects the number of new COVID-19 hospital admissions each day, along with the current number of COVID-19 patients. Three cohorts are distinguished—hospitalized, ICU, and ventilator. The user inputs population, hospital characteristics, and infection rate. CHIME is a Monte Carlo simulation of Susceptible-Infected-Recovered patients (sometimes referred to as an SIR model); an analyst can use it to forecast future hospital requirements. CHIME was used in the Philadelphia area to estimate when a three-hospital system would exceed capacity.

University of Washington COVID-19 Forecasting Tool (IHME COVID-19 Health Service Utilization Forecasting Team, 2020): This tool estimates the number of non-ICU and ICU beds at the state level assuming current social distancing. The projections over the next four months assume that social distancing continues until infections are minimized and containment implemented. This model does not provide information at the hospital level.

Our model differs from other COVID-19 models in its use of queueing theory. Queueing theory accounts for the random nature of the arrival rate and length of time on a ventilator. The advantage of our queueing model is its simplicity. It can be run in a spreadsheet, allowing for ease of use, transparency, and wide dissemination. It can provide insights at both the hospital and regional (or hospital system) levels. Although some hospitals may have sophisticated enterprise staff planning systems that they might use to plan their ventilator requirements, our model complements these systems by allowing COVID-19-relevant input parameters to be directly incorporated into analyses.

Our model could be used to manage other strained hospital resources, including dialysis machines. One study from China found that 29 percent of COVID-19 patients also had acute kidney injury (Yang et al., 2020). The New York Times recently reported that “20 percent to 40 percent of I.C.U. patients with the coronavirus suffered kidney failure and needed emergency dialysis” (Abelson et al., 2020). The article also stated that supplies are becoming critically low in hospitals in large cities (Boston, Chicago, New Orleans, and Detroit). The model presented in this Perspective could be adapted to assess where dialysis machines are needed most and how to allocate them efficiently, how to best manage patient flows to different dialysis centers, and more.

Model Description

Primarily, an Erlang C model (M/M/N queue) requires three inputs: the number of servers (here, ventilators), the arrival rate of users (here, ventilator patients), and the expected length of service (here, length of stay on a ventilator)—see the appendix for a short exposition. We extend the standard model in two ways. First, we account for the fact that COVID-19 and non-COVID-19 patients arrive at different rates and have different expected lengths of stay on ventilators. We do this by creating a composite patient that embodies the total expected arrival rate and length of service given the mix of patient types. Second, we
take the number of ventilators per RT on shift as a parameter and allow it to operate as a limiting factor: If more ventilators are available or needed than the current RT staff support, the model reports the RT shortage.

We describe the inputs and outputs of our model in detail in the following subsections. We describe how a user might vary these inputs to provide a sense of the model’s richness. To illustrate how the model may be parameterized, we give the default values that we use in our examples where appropriate.

Model Inputs

**The average length of stay (days) on a ventilator for a non–COVID-19 patient:** This value is used to calculate the composite arrival rate. In our examples, we use a default value of 4.5 days (Agency for Healthcare Research and Quality, 2017).

**The average length of stay (days) on a ventilator for a COVID-19 patient:** This value should align with the current best scientific understanding of COVID-19, although planners may affect this value by changing triage and length-of-stay protocols for COVID-19 patients. Based on the present literature, we use a default value of 11 days for this parameter (Li et al., 2020). This average length of stay should include both recoveries and mortalities.

**Arrivals per day, new non–COVID-19 patients needing ventilators:** This value represents the number of non–COVID-19 ventilator patients arriving in the hospital per day. This value can be estimated based on current data or the hospital’s seasonal occupancy history. We select illustrative values for our calculations based on annual ventilator patient counts derived from an analysis of the 2016 Health Care Utilization Project State Inpatient Databases. (For simplicity of discussion, an arrival is a patient who needs a ventilator, whether the patient is newly arrived at the hospital or is a current patient who requires a ventilator.)

**Arrivals per day, new COVID-19 patients needing ventilators:** This value represents the expected number of COVID-19 patients arriving in the hospital per day who are known or expected to require a ventilator. Users can input current values but should note that, in the current context, the arrival rate is expected to grow, particularly if social distancing is relaxed. Users can explore the impact of higher COVID-19 patient counts on ventilator and RT needs by entering projected patient counts here. Given a forecast of disease progression, planners could plot out how needs will change over time. We note that sources suggest that between 5 percent and 10 percent of COVID-19 cases require a ventilator during their care (Wu and McGoogan, 2020). We use illustrative values in our examples in the following section.

**The number of ventilators in the hospital:** This value represents the total number of ventilators available to the hospital, regardless of whether or not they are utilized. Planners could input current values or future values that reflect incoming or outbound resources or examine the implications of unexpected ventilator failures. We choose illustrative values in our examples in the following section.

**The number of RTs available per shift:** This value represents the total number of staff who can serve as RTs at any given time in the hospital.

In addition to inputs on hospital resources and patient counts at a given point in time, the calculations depend on the number of ventilators per RT and a target average
wait time. We selected default values for these parameters based on current medical literature, state regulatory requirements, and our own analyses. Users can adjust these parameters in line with any better information they have available or to examine the impacts of protocol changes.

**Number of ventilators per RT:** In our examples, we use a value of 4.88 ventilators per RT (based on California Society for Respiratory Care, 2016, p. 18). In a surge context, this number may be set as high as 12 ventilators per RT (Faculty of Intensive Care Medicine and Intensive Care Society, 2019). Users can vary this parameter to learn how stretching RT responsibilities could improve RTs’ ability to respond to high patient demand.

**Target average wait time, in hours, for a new patient to get on a ventilator:** This value represents the expected delay that a patient needing a ventilator experiences as they wait for one to become available. By default, we set this time to 0.1 hours to reflect a system that can typically respond quickly to patients’ needs. In a crisis, a planner could increase this value, delaying patient access to ventilation but also lowering the required number of ventilators. We note that some states issue guidance relevant for setting this value (e.g., New York State Task Force on Life and the Law, 2015). Planners could aim to equalize wait times across area hospitals to equalize access to critical resources, assuming similar triage procedures.

The average wait time can be difficult to interpret because it can reflect a highly skewed distribution in which most people experience no wait and a few people experience long waits. The user may prefer to think about, and specify, the probability that a patient must wait for a ventilator. Users may be less concerned about average wait and more concerned about the number of people at risk of experiencing a nonsurvivable wait time.

Mathematically, the probability of wait, the average wait time, and the expected length of stay are related through the number of ventilators. With more ventilators, wait time decreases by lowering the chance that patients must wait. The average wait time equals the probability that a patient must wait at all multiplied by the average length of stay on a ventilator. For example, if the average wait time is 1.68 hours and the average time a patient spends on a ventilator is 7 days, the probability of experiencing any wait is 1.68/(7*24) = 1 percent. This means that 99 percent of the patients do not wait at all, but 1 percent of the patients may have long waits. The 1 percent of patients who experience any wait time could be at risk of experiencing a nonsurvivable wait. See the appendix for further details.

**Model Outputs**

**Implied census of non–COVID-19 patients on ventilator:** This is the implied average number of non–COVID-19 ventilator patients in the hospital: a product of the arrival rate of non–COVID-19 patients and the average number of days on a ventilator for non–COVID-19 patients. We include this output in our model so that users can check this value against their patient census to cross-check the rate input if they expect it to remain stable. The implied census and the true census might not match exactly due to variability of patient stays but should seem reasonable given past experience.

**Implied census of COVID-19 patients on ventilator:** This is the implied average long run number of COVID-19
ventilator patients in the hospital: a product of the arrival rate of COVID-19 patients and the average number of days on a ventilator for COVID-19 patients. In the context of increasing arrival rates for COVID-19 patients, this value will underestimate the COVID-19 patient census. Users can look at this number to get a sense of where patient counts are going. With growing arrival rates, this expected average will tend to exceed the current number of patients in the hospital.

**Arrivals per day, all new patients on ventilators:** This value is the sum of non–COVID-19 and COVID-19 patient arrivals, provided for crosscheck purposes.

**Average length of stay (days) on a ventilator for all ventilator patients:** This value is a weighted average of expected days on ventilators for non–COVID-19 and COVID-19 patients.

**Maximum arrivals per day that can be supported with this number of ventilators:** Based on the current average time that patients spend on ventilators at the hospital, this is the maximum number of patients that the hospital can serve while keeping the probability of wait below the specified target.

**Arrivals per day over capacity, patients:** This is the number of new patients per day above the capacity of the available ventilators given the input data.

**Actual expected wait (delay) for a new patient to get a ventilator:** Based on our model and given the entered values, this is the expected time that a patient needing a ventilator will wait for one to become available.

**Minimum number of ventilators needed to reach target average wait time:** Based on our model and given the entered values, this is the smallest number of ventilators needed to satisfy the patient demand for ventilators and achieve the target average wait time.

**Ventilators short, the additional ventilators needed:** This number is the difference between the hospital’s total number of ventilators and the minimum number needed given the input data and target average wait time. A negative value in this field indicates that the hospital has a surplus of ventilators given the input information. Regional-level decisionmakers can therefore identify where ventilators are available to meet needs elsewhere. Decisionmakers should also maintain short-term forecasts to address changing arrival rates.

**RTs needed to support this number of patients on ventilators:** This is the number of RTs who would need to be available on any given shift to care for patients on the minimum number of ventilators needed to reach the target average wait time.

**RTs short, the additional RTs needed:** This is the difference between the number of RTs needed per shift and the number of RTs available. As with ventilators short, a negative value in this field indicates that the hospital has a surplus of RTs given the input information.

**Examples: One Hospital, a Regional Planner, and Increasing Arrivals over Time**

This section provides three examples that demonstrate our model. First, we give an example for one hospital, showing how the increase in arrival rate affects the number of required ventilators. Then, we show a regional coordinator example, examining the allocation of ventilators across
three hospitals. Finally, we discuss how the analyst would address a dynamic arrival rate, increasing or decreasing over time.

In the following examples, we will assume
- the average length of stay on a ventilator for a non–COVID-19 patient is 4.5 days (Agency for Healthcare Research and Quality, 2017)
- the average length of stay on a ventilator for a COVID-19 patient is 11.0 days (Li et al., 2020).

The increases in both the average length of stay and the arrival rate explain the dramatic need for ventilators.

Example for a Hospital

Table 1 shows example inputs for the queuing calculations at an imaginary hospital. Table 2 shows outputs of the queuing calculations for this example.

With these inputs, the hospital has an average of 7 patients arriving per day, each with the length of stay of 9.1 days on a ventilator. We see that the hospital has more patients than it typically does and that those patients also require longer time on the ventilator. With an average of 7 patients per day, 9.1 days on a ventilator, the hospital can expect to reach an average census of 64 ventilator patients, possibly within a week, depending on conditions. Table 2 shows the example outputs. The target average wait time for a ventilator includes cleaning and repair time.

Using an Erlang function with load = 9.1*7 and 67 ventilators, we find the probability of waiting to be 62 percent, meaning that new patients will have to wait on average 0.62*7.0*24 = 135.4 hours, which is almost certainly unacceptable. Patients must wait so long because of the variability of the number of arrivals needing ventilators and the increased length of time that the ventilator is required.

To ensure that patients wait only about 2 hours on average, the hospital would need 85 ventilators. This calculated value of 85 ventilators is not a forecast. The calculation assumes a relatively steady state. If the hospital had only 64 ventilators, any slight increase above the average arrivals or days on a ventilator would immediately force any newly arriving patient to wait. We can interpret the

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Example Inputs for Queuing Calculations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inputs</td>
<td>Value</td>
</tr>
<tr>
<td>Average length of stay (days) on a ventilator, non–COVID-19 patient</td>
<td>4.5</td>
</tr>
<tr>
<td>Average length of stay (days) on a ventilator, COVID-19 patient</td>
<td>11.0</td>
</tr>
<tr>
<td>Arrivals per day, new non–COVID-19 patients on ventilators</td>
<td>2.0</td>
</tr>
<tr>
<td>Arrivals per day, new COVID-19 patients on ventilators</td>
<td>5.0</td>
</tr>
<tr>
<td>Target average wait time for a ventilator, hours</td>
<td>2.0</td>
</tr>
<tr>
<td>Current number of ventilators</td>
<td>67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Example Outputs of Queuing Calculations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outputs</td>
<td>Value</td>
</tr>
<tr>
<td>Average length of stay (days) on a ventilator for any ventilator patient</td>
<td>9.1</td>
</tr>
<tr>
<td>Arrivals per day, all new patients on ventilators</td>
<td>7.0</td>
</tr>
<tr>
<td>Expected steady-state census of patients</td>
<td>64</td>
</tr>
<tr>
<td>Probability of wait for new patients</td>
<td>62%</td>
</tr>
<tr>
<td>Desired number of ventilators for a two-hour target wait</td>
<td>85</td>
</tr>
</tbody>
</table>
calculation as a type of safety stock, to be sure the hospital has enough.

From this, the hospital analyst can calculate the required number of RTs using the hospital’s preferred number of ventilators per RT. For example, if the hospital requires no more than 4.9 ventilators per RT, then the hospital will need $85/4.9 = 17.4$ RTs. The hospital analyst can make similar calculations for dialysis machines or any other kind of equipment.

With this approach, the hospital analyst can calculate the maximum number of patient arrivals per day that the hospital could have with its current number of ventilators. Using the Erlang function with the current number of ventilators, the analyst (or program) simply increases the number of arrivals until the expected wait time reaches the target wait time. Using the data from Table 1, we calculate 5.3 patients per day. Similarly, the hospital analyst could calculate the maximum number of patient arrivals for its current number of RTs.

**Example for a Regional Planner**

Consider now the view of the regional planner with access to only a few new ventilators during a generally recognized shortage of ventilators. Assume further that the planner has access to the data we have described for each hospital.

Where should the planner put the next free ventilator? We think the planner should put the next ventilator at the hospital with the longest wait time. Where should the planner direct the next patient? Without being able to comment on the needs of different patients, we think the planner should direct the patient to the hospital with the shortest wait time. Knowing estimated wait times is important input for planners who are accustomed to balancing hospital needs based on demographics and perhaps infection hot spot information. As previously mentioned, states are preparing guidelines for ventilator management. Thus, the wait time calculation can be used in conjunction with the guidelines for better management of this scarce resource.

The planner may wish to request loans of ventilators from hospitals with short wait times and then move those ventilators to hospitals with long wait times. This would spread the supply of ventilators.

The planner could also seek to direct new patients to hospitals with short wait times rather than hospitals with long wait times. This would spread the load of patients.

In theory, repeatedly placing the next ventilator at the hospital with the longest wait time and directing new
patients to hospitals with short wait times would eventually result in all hospitals having the same expected wait time. This insight suggests a practical allocation mechanism.

During a crisis affecting many hospitals across a region simultaneously, the regional planner could choose a target wait time over all hospitals. The target wait may be longer than anyone prefers, but it allows the planner to allocate ventilators as effectively as possible, spreading new patients and the supply of ventilators to balance the loads across hospitals. The same approach applies to RTs. Table 3 shows calculations for an imaginary region with three hospitals. From Table 3, we see that increasing the target wait time decreases the number of ventilators required, but unevenly across hospitals. Queuing theory provides the regional planner with a mechanism to understand the available capacity at each hospital, how requirements for ventilators change with the number of patients arriving at each hospital, and a method to balance the loads across hospitals.

Suppose Hospital B had 32 ventilators rather than only ten. Then, it would have a surplus of ventilators, which it could conceivably share with the other hospitals.

### Planning for Increasing Arrivals

The basic queuing formulas assume a steady average arrival rate and a steady length of stay. When the analyst forecasts the arrival rate to increase, the analyst must exercise care in forecasting the need for ventilators.

At the beginning of a crisis, the analyst may observe ten arriving COVID-19 patients and may research the literature to find 11 days as their expected length of stay. These values imply an average census of 110 COVID-19 patients. But at the start, the actual census would be only ten patients, thus overestimating the number of ventilators immediately needed for a few days, and perhaps as long as a week or even 11 days. At that point, with a relatively steady arrival of ten patients per day, the analyst can expect to need many more than 110 ventilators to ensure short wait times for new arrivals. Thus, if the analyst expects an arrival rate of ten for the near future, the queuing formula provides a helpful guide for the near future. See the appendix for ways to use the model to account for either increasing or decreasing arrivals.

### Table 3

<table>
<thead>
<tr>
<th>Calculated Variables</th>
<th>Hospital A</th>
<th>Hospital B</th>
<th>Hospital C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal patient arrival rate, patients per day</td>
<td>0.042</td>
<td>0.99</td>
<td>17.7</td>
</tr>
<tr>
<td>Crisis patient arrival rate, all increase being COVID-19</td>
<td>0.083</td>
<td>1.98</td>
<td>35.4</td>
</tr>
<tr>
<td>Number of ventilators currently at each hospital</td>
<td>2</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Probability of wait with current number of ventilators</td>
<td>16%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Desired number of ventilators for a two-hour target wait</td>
<td>4</td>
<td>26</td>
<td>315</td>
</tr>
<tr>
<td>Probability of wait with two-hour wait time</td>
<td>0.5%</td>
<td>0.9%</td>
<td>1.0%</td>
</tr>
<tr>
<td>Desired number of ventilators for a four-hour target wait</td>
<td>4</td>
<td>25</td>
<td>310</td>
</tr>
<tr>
<td>Probability of wait with four-hour wait time</td>
<td>0.5%</td>
<td>1.6%</td>
<td>2.1%</td>
</tr>
</tbody>
</table>
Conclusion

In this Perspective, we propose the use of a queuing model for planning hospital requirements for ventilators, RTs, and other resources in a crisis. A hospital analyst could implement the queuing model in a spreadsheet and use it to estimate the patient wait times given the current number of ventilators and RTs. The analyst can also use the model to estimate the required number of ventilators and RTs when arrival rates and lengths of stay change. Although the queuing model theoretically assumes steady arrival rates, the careful analyst could use forecasted changes in arrival rates at the hospital or regional level. We further observe that a regional coordinator could use a target wait as a metric for allocating scarce ventilators and for directing patients to hospitals. This methodology could assist federal planners who allocate scarce resources across states and regions.

Appendix

Potential Ways to Use This Model

If the analyst expects an increasing arrival rate for the foreseeable future, the analyst may wish to develop a more sophisticated analysis of the number of patients on hand and use that value as the load (the product of the average arrival rate and the expected length of stay) in the queuing formula. The analyst forecasts the expected census for each future day. The analyst uses the expected census as the load in the Erlang function; with the targeted wait time, the analyst can use the Erlang function to determine the number of required ventilators. With this analysis, the analyst can also calculate the runout date, when the ventilators required exceed the number of ventilators available. The analyst can also calculate how the expected wait time will change should new ventilators not appear.

If the analyst expects a decreasing arrival rate at some point in the future, the analyst will have to decide how much risk the hospital is willing to take in surrendering ventilators early or in anticipating too few; patients would experience the likely poor outcome as waits longer than targeted—possibly much longer. If the encompassing region experiences the crisis simultaneously so that the arrivals occur simultaneously across hospitals, the regional planner would probably have little use of excess ventilators and, therefore, would not be likely to request loans from the hospital. If hospitals peak at different times, the planner may want to press the hospital administrator about whether the hospital’s peak has passed; in this case, we expect the hospital administrator to be extremely risk-averse.

Model Formulation

In our model (an M/M/N queue where the first M identifies the arrival distribution, the second M identifies the server distribution, and N indicates the number of servers in this case ventilators), ventilator patient arrivals follow a Poisson process, and their lengths of stay on the ventilator follow an exponential distribution. We account for the differential arrival and use rates of non–COVID-19 versus COVID-19 patients. We want to determine the ventilators and respiratory staff necessary to satisfy a specified, expected target wait time $E[\text{Wait}]_{\text{Target}}$. In general:

$$E[\text{Wait}] = P(\text{Wait} > 0) \cdot E[\text{Wait} | \text{Wait} > 0],$$
where the probability of waiting $P(\text{Wait} > 0)$ depends on the number of ventilators, the number of patients, and the length of time the patient typically needs on the ventilator; and the expected waiting time $E[\text{Wait}|\text{Wait} > 0]$ is a weighted average of the typical service time for non–COVID-19 and COVID-19 ventilator patients.

Our assumptions conform to the Erlang C queuing model staffing (see Robbins, Medeiros, and Harrison, 2010). For a given composite arrival rate of patients $\lambda$, an average length of stay on a ventilator $\mu$, and a number of ventilators $N$, define the offered load as $R = \lambda/\mu$ and the traffic intensity as $\rho = R/N$. For $N > R$,

$$P(\text{Wait} > 0) = 1 - \frac{\sum_{m=0}^{N-1} \frac{R^m}{m!}}{\sum_{m=0}^{N-1} \frac{R^m}{m!} + \frac{R^N}{N!} \cdot \frac{1}{1 - \frac{R}{N}}} .$$

We calculate $\lambda$ and $\mu$ based on a composite of COVID-19 and non–COVID-19 patient arrival rates and lengths of stay, and $N$ is the number of ventilators. We calculate the number of ventilators needed by finding the lowest value for $N$ such that the expected wait time $E[\text{Wait}] \leq E[\text{Wait}]_{\text{Target}}$. We refer to this value of $N$ as the minimum number of ventilators needed to reach the target average wait time. The ventilators short equals the minimum number of ventilators needed minus the hospital’s current number of ventilators. Given the input parameters, if the hospital has a surplus of ventilators, ventilators short is negative.

We determine RTs short in a similar manner, identifying the minimum number of RTs needed by using the model parameter number of ventilators per RT and dividing as follows: minimum number of ventilators needed / number of ventilators per RT. Then, the respiratory therapists short equals the RTs needed minus the number of RTs in hospital per shift.

Notes

1 More precisely, the arrival rate of COVID-19 patients and their length of stay on ventilators are stochastic variables in the sense that they can be statistically modeled but not fully determined. We leave aside this aspect of the problem, but, where implemented, the accuracy of our model’s outputs could be improved by more-accurate estimates of these variables.

2 The Erlang C model assumes a Poisson arrival process with mean $\lambda$. The median $M$ of this distribution is such that $\lambda - \ln 2 \leq M \leq \lambda + 1/3$. As $\lambda$ increases, the median and mean approach each other in the limit.

3 At present, testing for COVID-19 may vary, from a rapid one-hour test to one that takes up to nine days to complete, and testing capacity is in short supply. We recommend users of the tool think of this value as known or suspected COVID-19 cases.

4 The model does not account for death in the queue or other changes to the queue.

References


About the Authors

John F. Raffensperger is a senior operations researcher at the RAND Corporation. He has worked on problems of national security and emergency preparedness, including disaster recovery and military readiness and logistics. His research interests include optimization, risk analysis, logistics, and ecological economics. Raffensperger holds a Ph.D. in business.

Marygail K. Brauner is a senior operations research analyst at RAND. Her current research helps inform the rebuilding of Puerto Rico using evidence-based analysis, and past research has focused on military-force planning and logistics. Her research interests include optimization, risk analysis, logistics, scheduling, and resource allocation. Brauner holds a Ph.D. in engineering and operations research.

R. J. Briggs is an economist at RAND. His research focuses on cost estimation for disasters and recovery, risk and insurance, and energy and environmental policy. He has worked on a diverse array of RAND projects, from modeling the returns to education to measuring the efficacy of U.S. Naval communications networks. Briggs holds a Ph.D. in economics.
About This Perspective

This Perspective proposes a queuing model for planning ventilator requirements in a crisis for hospitals, state and regional coordinators, and federal planners. A hospital analyst could implement the queuing model in a spreadsheet and use it to estimate the hospital’s current capacity and the required resources when arrival rates and lengths of stay change, in both the number of ventilators and the number of respiratory therapists. A state or regional coordinator could use wait time as a metric for allocating scarce ventilators to hospitals. The metric would suggest additional ventilators for hospitals with long expected wait times and would suggest directing patients to hospitals with short expected wait times. At the federal level, a recent pandemic planning exercise noted the need for a national distribution and allocation strategy; this model could help with development of that strategy.

Although the queuing model theoretically assumes steady arrival rates, the careful analyst can use the model with forecasted changes in arrival rates, whether at the hospital level or the regional level. The authors developed an early version of software that outputs the calculations described in this Perspective. Hospital and regional health administrators interested in working with the RAND Corporation to fully realize its potential can contact the authors for more information (John F. Raffensperger, jraffens@rand.org; Marygail K. Brauner, marygail@rand.org; R. J. Briggs, rjbriggs@rand.org).

RAND’s mission is to help improve policy and decisionmaking through research and analysis. This research was funded by the U.S. Department of Health and Human Services and carried out within the Payment, Cost, and Coverage Program in RAND Health Care. RAND Health Care, a division of the RAND Corporation, promotes healthier societies by improving health care systems in the United States and other countries. We do this by providing health care decisionmakers, practitioners, and consumers with actionable, rigorous, objective evidence to support their most complex decisions. For more information, see www.rand.org/health-care, or contact

RAND Health Care Communications
1776 Main Street
P.O. Box 2138
Santa Monica, CA 90407-2138
(310) 393-0411, ext. 7775
RAND_Health-Care@rand.org

The authors thank RAND Health Care for funding this work and gratefully acknowledge helpful comments from Carl Berdahl, Chrissy Eibner, Emmett Keeler, Paul Koegel, Lisa Turner, Emily Ward, and Alex Gamble.