ABSTRACT

Estimating the capacity of a region to serve pandemic patients in need of hospital services is crucial to regional preparedness for pandemic surge conditions. This paper explores the use of techniques of stochastic discrete event simulation for estimating the maximum number of pandemic patients with intensive care and/or in-patient, isolation requirements that can be served by a consortium of hospitals in a region before requesting external resources. Estimates from the model provide an upper bound on the number of patients that can be treated if all hospital resources are re-allocated for pandemic care. The modeling approach is demonstrated on a system of five hospitals each replicating basic elements (e.g. number of beds) of the five hospitals in the Johns Hopkins Hospital System in the Baltimore-Washington, D.C. Metropolitan area under settings relevant to the COVID-19 pandemic.

1 INTRODUCTION AND BACKGROUND

The COVID-19 pandemic has alerted communities across the world to the need for an ability to estimate regional capacity to serve pandemic patients requiring hospital services during a surge. An understanding of the upper bound on the number of patients who can be treated during pandemic surge conditions and the most limiting resources preventing greater service capacity are crucial to regional preparedness. This paper explores the use of techniques of stochastic, discrete event simulation (DES) modeling for estimating the maximum number of pandemic patients with intensive care or in-patient and/or isolation requirements that can be served by a consortium of hospitals in a region before requesting external resources. The consortium may have a formal memorandum of understanding in place, simply exist due to proximal locations within a single region, or arise as a consequence of ownership. Estimates from DES models provide upper bounds on the number of patients who can be treated if all hospital resources within the consortium are re-allocated.
for pandemic care. The modeling approach enables the identification of the most limiting resource at each hospital within the consortium in terms of beds, doctors and nurses given existing levels, and quantification of advantages from coordination and resource pooling.

The need to turn nearly all hospital resources toward pandemic patient care, and creating COVID-19 designated hospitals (e.g. Edward et al. 2021), became evident in a variety of locations across the globe during times of COVID-19 hospital patient surge, including in the U.S. (Philips et al. 2020; Assistant Secretary for Preparedness and Response (ASPR (2022)), Canada (Fadaak et al. 2021; Patey et al. 2020), China (Cao et al. 2020), Germany (Möckel et al. 2020), India (Pandey et al. 2020; Goswami and Dutt (2021)), Italy (Tosoni et al. 2020), and Brazil (Falcetta et al. 2021), among other locations. Creating COVID-19 designated hospital services to support these surges is resource-intensive and expensive (e.g. King, (2021)). Determining best actions to alleviate some of the associated costs can be supported through better understanding of most limiting resources and maximum capacities that can be achieved through such service transformation.

In prior work, Fattahi et al. (2022) developed a multi-stage stochastic program to develop the details of strategies for sharing externally provided resources, such as ventilators, and resource sharing through relocation across a healthcare system, or redistributing demand between hospitals in a region during a pandemic surge. Optimal distribution of externally provided resources is also considered by Arora et al. (2010) in relation to allocating supplies from centralized stockpiles among hospitals. Prior to the COVID-19 pandemic, Shahverdi et al. (2020) investigated the potential gains from collaborations between hospitals in a region in managing surge demand from mass casualty incidents or pandemic (imaging then only a bad flu season) or during disaster events damaging hospitals or supporting lifelines. They considered patient transfers and centralized patient allocation strategies, i.e., a form of demand redistribution, and sharing of doctors and nurses across hospitals. This pre-COVID-19 study did not include dedicated care paths for pandemic patients. In later work, Shahverdi et al. (2022) incorporated such care paths, but their work focused on capacity concerns for an individual hospital. While some works have considered the role of collaboration, redistribution of demand, or sharing centralized resources for improving regional hospital capacity, it appears that no prior work has developed methods for or has sought to estimate the maximum capacity of a region for pandemic patient hospital care in a surge as has arisen in the COVID-19 pandemic and is presented here.

In the next section, Section 2, details of the proposed stochastic DES methodology for estimating the maximum capacity for serving pandemic patients in a consortium of hospitals are given. In Section 3, the methodology is demonstrated through computational experiments on a simulation of a system of five hospitals replicating basic elements (e.g. number of beds) of the five hospitals in the Johns Hopkins Hospital System (JHHS) in the Baltimore-Washington - D.C. Metropolitan Area, also considered in (Shahverdi et al. 2020), under settings relevant to the Alpha and Omicron strains of COVID-19. The application demonstrates that a simple pooled-resource methodology can provide a tight bound on capacity estimates, and applying the same technique on individual hospitals can inform decision-makers on that hospitals most binding resource. It further shows the advantages of regional coordination and resource pooling.

2 METHODOLOGY

A stochastic DES of the COVID-19 only design for each hospital in the region is first created. In this study, the DES was constructed in ExtendSIM, a microscopic DES platform with underlying queueing network conceptualization. Each hospital is modeled with three key care paths wherein most patients enter the hospital through a critical decision unit (CDU). A predefined portion of patients are sent directly to the ICU as a function of the characteristics of the COVID-19 (or other pandemic) variant. The length of stay (LOS) of patients entering the ICU is determined by a random variate with exponential distribution and a mean
consistent with the variant of concern. Mortality rates are given by a probability aligned with the variant and the patient’s condition at arrival. The remainder of patients entering the hospital will be served in the CDU. Some portion will be sent home while others will be admitted. Admitted patients are given in-patient beds within the internal general ward (IGW) with or without isolation (treated generally here as isolation) before discharge. Again, the length of stay is a random variate with exponential distribution and a mean consistent with the variant and patient condition upon arrival. Mortality rates are also applied as a function of the variant in the isolation rooms. Patients move from the ICU to the IGW before being discharged. Patients in the IGW (or isolation rooms) are discharged once completing their stay and receiving all needed services. Figure 1 shows the hospital layout in terms of these care paths.

This COVID-19 only hospital design and model construct follow from (Ghayoomi et al. (in review)). In Ghayoomi et al. (in review), an optimization methodology is proposed for determining an optimal re-allocation of beds and heterogeneous sets of nurses and doctors to the units of the COVID-19 only hospital. Only a subset of each set of personnel is qualified to serve the ICU. The optimal re-allocation provides the maximum steady-state COVID-19 only daily patient arrival rate that the hospital can serve under predefined levels of acceptable steady-state daily numbers of patients who may be turned away from either the ICU or isolation rooms due to lack of capacity. Such patients who must be turned away are called reneges, and reneges occur when patients have waited an extended period. This approach for determining the allocation and maximum steady-state arrival that can be reasonably served (i.e. with few reneges) relies on concepts of mixed-integer programming, queueing theory and Jackson network decomposition. The optimal allocation of resources within each hospital of the region under study and maximum steady-state daily patient arrival rates are used as input to this proposed regional capacity estimation methodology as depicted in Figure 2.

The generic 200-bed urban U.S. tertiary hospital DES model with Hot Spot design (Shahverdi et al. 2022) and optimal allocation of beds, doctors and nurses (Ghayoomi et al. (in review)) formed a baseline for replicating hospitals of varying size. The original hospital model with key parameters was developed from input provided through extensive interviews with key hospital personnel from the JHHS. Where parameter settings were less well known to the experts, settings suggested in the literature and from national averages were used. The process of model construction, refinement and verification and their outcomes are
described in more detail in (Tariverdi et al. 2019). The designation of 200-beds refers to the pre-pandemic hospital setting and its 200 IGW beds. To this end, each hospital model is created by scaling the 200-bed hospital model in terms of numbers of resources in each unit according to the number of IGW beds in the target hospital. An overview of the methodology used herein is given in Figure 2.

Simulation replications are made under two settings of hospital collaboration over the set of hospitals. The first presumes that hospitals within the coalition or study region operate independently and the second assumes all resources are pooled. In the first case, two versions are considered as a function of patient...
allocation. Using a total of the maximum steady-state daily patient arrival rates over all considered hospitals, the first version presumes that the patients arrive to the hospitals in equal numbers, with no inter-hospital transfers permitted. In the second version, total patient arrivals to the considered hospitals is presumed to be split across the hospitals in proportion to the hospital size as measured by the number of IGW beds, consistent with original maximum individual arrival estimates from the optimizations. This can be seen as an ideal allocation of patients to hospitals or in terms of any arrival pattern involving inter-hospital transfers, but with no negative impacts from transfer efforts. In the second case, all resources are pooled. In line with creating an upper bound, this formulation presumes patient transfers and the movement of resources occur instantaneously. This setting effectively creates a single giant hospital of all resources to which all patients arrive. This case is found to provide a tight upper bound on regional (or coalition) capacity. These two general settings are depicted in Figure 3.

Figure 2: Methodological overview.

Figure 3: Hospital settings - idealized pooling versus operating independently.
3 DEMONSTRATION ON CASE STUDY

3.1 Case Study Setting

The proposed methodology is demonstrated on a coalition of the five hospitals of the JHHS in the Washington, D.C.-Baltimore Metropolitan Areas. This system includes Suburban Hospital, Howard County General Hospital, Sibley Memorial Hospital, Bayview Medical Center and Johns Hopkins Hospital with 228, 243, 288, 462 and 1162 beds, respectively. Their locations are shown in Figure 4.

Table 1: Number of resources at each hospital.

<table>
<thead>
<tr>
<th>Resources</th>
<th># IGW Beds</th>
<th>ICU</th>
<th>Isolation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base case</td>
<td>Optimized</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Base</td>
<td>Suburban Hospital</td>
<td>Howard County General Hospital</td>
</tr>
<tr>
<td>Bed</td>
<td>90</td>
<td>35</td>
<td>39</td>
</tr>
<tr>
<td>Doctor</td>
<td>21</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Nurse</td>
<td>45</td>
<td>30</td>
<td>34</td>
</tr>
<tr>
<td>Bed</td>
<td>235</td>
<td>290</td>
<td>330</td>
</tr>
<tr>
<td>Doctor</td>
<td>28</td>
<td>46</td>
<td>52</td>
</tr>
<tr>
<td>Nurse</td>
<td>48</td>
<td>63</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 2: Model settings by COVID-19 variant.

<table>
<thead>
<tr>
<th></th>
<th>Earlier Variants</th>
<th>Omicron (Peralta-Santos et al. 2022)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ICU</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOS</td>
<td>Exponential Mean 9 days</td>
<td>Exponential Mean 4 days</td>
</tr>
<tr>
<td>Mortality rates</td>
<td>30%</td>
<td>10%</td>
</tr>
<tr>
<td>% patients transferred directly to ICU upon arrival</td>
<td>10%</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Isolation Room</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOS</td>
<td>Exponential Mean 5 days</td>
<td>Exponential Mean 5 days</td>
</tr>
<tr>
<td>Percentage of patients admitted</td>
<td>60%</td>
<td>65%</td>
</tr>
</tbody>
</table>

Using the base 200-bed urban U.S. tertiary hospital DES model with Hot Spot design and optimal
allocation of beds, doctors and nurses as a baseline, each of the five hospitals of the JHHS case study were created. Details of the resource settings by unit in each hospital and model parameters as a function of COVID-19 variant are given in Tables 1 and 2, respectively.

### 3.2 Experimental Design

50 runs of the simulation were completed across three sets of run categories from the two collaboration settings as depicted in Figure 5. The first two assume the hospitals operate independently, but arrivals to the hospital are either even or in proportion to hospital size. The last employs the idealized resource pooling concept. In all runs, 1,098 patients are presumed to arrive to the hospital coalition daily and split evenly across the day, arriving in four intervals. The specific allocation of patients to hospitals for each setting of the runs is shown in the figure.

Each run was conducted over 62 days, including a 20-day warm-up period for reaching steady state. Average results are reported over the 42 remaining days and across the 50 replications. A single batch of runs required approximately 5 hours on a personal computer. In all runs, the 1,098 daily patients are presumed to arrive to consortium hospitals as dictated. This number was derived from the total of the maximum optimized arrival rates by hospital given its size in relation to the baseline hospital.

Figure 5: Hospital settings.

### 3.3 Experimental Results

Results are given in Figures 6 through 9 for the settings associated with the Omicron variant. Results of the earlier Alpha variant were found to be quite similar and were, thus, not reported. The results of Figure 6 indicate the importance of coordination. Comparing the independent hospital results where patients arrive evenly across hospitals regardless of hospital size to the case where they arrive in proportion to the hospital size shows a 45% increase in patient throughput and nearly 70% decrease in patients who must renege. These factors provide an indication of the gains that can be achieved through transfers between hospitals under ideal transfer conditions (i.e., with unlimited resources and rapid completion).

Figure 7 shows the change in utilization rates across hospitals under each setting. In every hospital and in both units, the doctors are the most limiting factor, and the cause of nearly all reneges. In a few cases, open beds are equally in short supply. This figure shows the utilization rate for the most binding resource averaged over the day and over all 42 days and over all 50 runs. The utilization rate will vary across days and over runs, and perhaps more so within a single day.

Figure 8 shows changes in throughput and reneges across the hospitals. The most notable difference is for the main Johns Hopkins Hospital, the largest of the hospitals. When demand is split across hospitals evenly, this hospital has great excess capacity, which can be utilized if inter-hospital transfers are supported. The results in these three figures also show that the isolation rooms have much less excess capacity than...
does the ICU in all hospitals. This is by design, as turning patients away who are in critical care is avoided. Re-allocation strategies within the hospitals, thus, put greater emphasis on avoiding too many reneges from the ICU as compared with the isolation room, and greater utilization rates are noted for the isolation rooms.

Results from the idealized pooling of hospital resources show a marginal decline in throughput with a small reduction (72 versus 75) in reneges from independently run hospitals with idealized transferring. Considering the ICU utilization rates between these runs in Figure 8, utilization rates drop under idealized resource pooling, leaving more open capacity across the board. Upon further investigation, it is found that the average LOS under idealized resource pooling is 4.08 days. In comparison, the average LOS for the independently run hospitals with proportional arrivals ranged between 3.99 and 4.05. The noted slight decline in throughput, thus, is likely due to the increased experienced LOS in the ICU, which is likely due to increased backups from full isolation rooms. The full isolation rooms prevent patients from finishing their care before discharge, thus, slightly reducing total throughput.

Significant variability exists in the run results as illustrated by considering the utilization rates in the ICU in Figure 9. Generally, when only average results by run or across hospitals are considered, encounters where resources are limited may not be recognized. Figures 9 and 10 also indicate lower utilization values of the most limiting resource in the case of idealized pooling. This is likely due to a leveling of resource use that occurs from pooling the resources.
Figure 6: Regional capacity for different strategies.

Figure 7: Average utilization rates by strategy for each hospital.

Figure 8: Capacity of individual hospitals under different strategies.
Added runs were made to investigate the number of additional arrivals that could be absorbed under the idealized resource pooling setting with the same 75 average daily steady-state renege rate as obtained from the independent hospital setting with transfers. An additional 10 daily patient arrivals could be absorbed before exceeding the 75 renege average. These findings related to the idealized resource pooling runs are two-fold. First, they show that the giant, single hospital representation of all resources can provide a tight upper bound on the maximum capacity for surge demand for the coalition. Second, they indicate that there are very significant gains to be had from transferring patients to hospitals with excess capacity, and potential gains from putting all resources in one location as in centralized resource pooling. A less distributed approach to hospital care can enable more efficient use of idle resources in such high utilization settings.

Figure 9: Utilization rate variability in most limiting resource by run in ICU.

Figure 10: Utilization rate variability in most limiting resource by run in isolation room.
4 CONCLUSIONS AND EXTENSIONS

This work provides a stochastic DES approach to estimating regional or coalition-based hospital capacity for pandemic patient care for extreme surge scenarios, where all hospital resources need to be turned toward pandemic patient care. A comparison is made between two modeling approaches, the first that models each hospital of the region independently and the second that uses a single giant hospital model that ideally pools all resources. Results of numerical experiments on a case study modeled from the JHHS hospitals in the Baltimore-Washington, D.C. Metropolitan area demonstrate that the simpler giant hospital modeling approach provides very tight bounds on maximum regional pandemic patient capacity. The results also show the benefits that can be achieved from collaboration through transferring patients and/or pooling resources. While an idealized resource pooling approach may not be practically implementable, dispatching patients centrally in proportion to the number of hospital beds (perhaps in proportion to the number of open beds in a real-time setting) may also increase regional capacity. Centralizing operations may also aid in exploiting idle resources that are critical in such high utilization situations as arise in pandemic surges.

Results were derived from only 50 runs due to computational burden. Additional runs would reduce overall variation and size of confidence intervals in estimates. The hospitals used in this study were based on scaling from a baseline, generic hospital in the system. More realistic estimates could be gained from conversion from actual hospital design details. The most limiting resource in these runs was almost always the doctors. Absenteeism due to illness or other causes during surge conditions could very dramatically reduce the accuracy or attainability of maximum capacity estimates. Runs of the models with reduced personnel resource pool scenarios could produce meaningful estimates to plan for these circumstances. With efforts to obtain real-world data on throughputs, wait times, reneges, and other metrics become available for hospitals in periods in which they were designated for COVID-19 patient care, the outcomes of estimates obtained through DES runs can be assessed for their ability to replicate actual capacity conditions. With feedback from such data, the models can be further refined and additional limiting resources or different hospital care path configurations can be integrated.

The models can be modified to assess the potential of telemedicine to expand capacity for pandemic patient care in the region. Critical care telemedicine programs have been successfully implemented and are found to produce lower ICU and hospital mortality and shorter ICU LOS (Lilly et al. 2014). These systems often work by enabling intensive care physicians to interact with patients and bedside patient care teams virtually. This is especially useful for locations where intensive care physicians are in short supply, as might be the case in more rural locations, for example, during routine circumstances. ICU telemedicine could be a viable option for increasing capacity of COVID-19 designated hospitals. Results of the numerical experiments indicated that doctors were nearly always the limiting resource in the case study hospitals. Runs of the idealized pooling model with an expanded doctor resource pool through telemedicine and related changes in unit-based LOS and mortality rates, where relevant, could provide estimates of increases that can be attained in maximum regional hospital capacity through telemedicine.

REFERENCES


Proceedings of the 2022 Winter Simulation Conference


AUTHOR BIOGRAPHIES

BAHAR SHAHVERDI, M.S. received her Bachelor of Science in Industrial Engineering from Sharif University of Technology in Iran in 2016. She is currently a Ph.D. student at George Mason University in Fairfax, completing her Master of Science degree in the Systems Engineering and Operations Research in 2019. Her expertise and interests are in disaster preparedness planning, DES modeling of healthcare systems, and mathematical modeling and algorithms. Her email address is bshahver@gmu.edu.
HADI GHAYOOMI, M.S. received his B.S. in Civil Engineering from Isfahan University of Technology in Iran (2014) and M.S. in Transportation Engineering at Sharif University of Technology (2017). He is a Ph.D. student at George Mason University. His expertise and interests are in resilience in extreme environments, DES modeling of healthcare systems, cybersecurity, queuing theory, and mathematical modeling and algorithms. His email address is hghayoom@gmu.edu.

ELISE MILLER-HOOKS, Ph.D. holds the Bill and Eleanor Hazel Endowed Chair in Infrastructure Engineering at George Mason University. With a Ph.D. (1997) from the University of Texas – Austin, she has expertise in: disaster planning and response; multi-hazard resilience quantification; transportation and infrastructure systems engineering; mathematical modeling and optimization; and more. She is the founding Editor-in-Chief of IFORS/Elsevier’s Sustainability, Analytics and Modeling journal and serves on numerous additional editorial boards. Her email address is miller@gmu.edu. (http://civil.vse.gmu.edu/miller)

MERSEDEH TARIVERDI, Ph.D is senior data scientist in Health, Nutrition, and Population Global Engagement at the World Bank Group working at the intersection of resilient health systems and climate change. She works on evidence-based and data-driven decision making in health and nutrition, and helps World Bank teams and clients employ Artificial Intelligence. Previously, she was principal data scientist in the private sector working on big data and trend analysis. She holds a Ph.D. specializing in health and infrastructure systems resiliency and capacity enhancement strategies. Her email address is mtariverdi@worldbank.org.

THOMAS D. KIRSCH, M.D., MPH, is Professor and Director of the National Center for Disaster Medicine and Public Health and a Professor of Military and Emergency Medicine at the Uniformed Services University of the Health Sciences. He is a board-certified emergency physician. Before coming to NCDMPH, he was a Professor of Emergency Medicine, International Health and Civil Engineering at Johns Hopkins University. Dr. Kirsch is a globally recognized researcher and educator with keen interest in disaster health, public health and emergency medicine. His email address is Thomas.kirsch@usuhs.edu