Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

PHYSICIAN STAFFING IN TELEMEDICINE: A SIMULATION-BASED APPROACH FOR A NETWORK OF CVS MINUTE CLINICS

Shuwen Lu[1], Mark E. Lewis[2], and Jamol Pender[2]

[1]Systems Engineering, Cornell University, Ithaca, NY, USA
[2]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, USA

ABSTRACT

Telemedicine has seen rapid expansion partially because of the COVID-19 pandemic, which helped to reduce regulation regarding telemedicine availability. In particular, we are inspired by the minute-clinic model of CVS-Aetna, and in this paper, we build a minute-clinic simulation model to understand the proportion of time with which a nurse practitioner can obtain additional medical advice from a collaborating physician. We use this simulation to construct staffing policies for collaborating physicians to address staffing issues nationally. We compare our simulated staffing schedules with those from Gaussian and Binomial approximations and assess their quality. Finally, we develop staffing procedures when the physicians are restricted by state regulations and compare them to situations where there are no restrictions to understand the impact of the regulations.

1 INTRODUCTION

Before the coronavirus pandemic, many Americans might have been more hesitant to interact with their doctors via virtual appointments since they were uncertain about the quality of care. However, nearly four years into the pandemic, the necessity for virtual appointments (mainly to reduce COVID-19 exposure risks for front-line workers such as nurses, doctors, and medical personnel) has made the adoption of telemedicine in the United States much more common. As a result, many medical practices are investing heavily in telehealth platforms to reach as many patients as possible.

Despite most people viewing telemedicine as a way of replacing in-person medical care with a virtual option, some companies such as CVS-Aetna view it as an opportunity to expand medical care to patients. The new minute clinics by CVS-Aetna provide a new platform that allows patients to receive high-quality medical care with a combination of in-person and virtual care. With the increase in training of medical service providers (MSPs), nurse practitioners (NPs) and registered nurses (RNs) can handle a significant proportion of patient medical needs without a physician. However, even when the patient is of low acuity, there are still instances when a physician's expertise needs to be consulted. As network communication capabilities have increased, so has the ability to deliver quality healthcare without requiring every patient to meet in person with a medical doctor (referred to here as general physicians (GPs). Together these advances have made it more prevalent for NPs and GPs to collaborate to deliver quality healthcare with at least one of them doing so virtually. This hybrid option of care already exists in the context of Mayo Clinic where physicians are primarily used to validate or verify the proposed treatments of physicians in small town or rural areas, see for example Haddad et al. 2021 and Locke et al. 2021.

As is the case with any service system, there is a question of the quality of service. For example, people can be impatient and have the tendency to abandon the queue waiting for service. In health-related systems, this is more than the loss of income - there is also potential health-threatening danger from the perspective of patients. Therefore, this work aims at answering the question of how many collaborating physicians (CPs) are needed to guarantee that a large fraction of patients who need additional care from

979-8-3315-3420-2/24/$31.00 ©2024

a virtual CP have no wait for this service. In queueing theory, this is a question of *server staffing* with a quality of service constraint.

By providing telemedicine consultation, the CPs can be accessed completely by the clinic locations in the same network, for example, within one state since the physicians are equipped with the same state license. Such complete resource access potentially reduces the wait and improves the utilization of resources despite its complicating the staffing decisions. Hence, this work proposes an efficient simulation-based staffing approach that easily generalizes to the network case. When it extends to nationwide staffing, however, strict state regulations might be a barrier to the expansion of telemedicine. The state with restricted practice limits an out-of-state physician to render services in that state, which hinders the providers from practicing telemedicine. On the other hand, the state with full practice allows eligible physicians to deliver high-quality care across state lines via telemedicine in a state where they are not physically located. The discrepancy in flexibility motivates our investigation of the impact of state regulations on the service quality of telemedicine.

Nowadays, telemedicine offers three main types of healthcare delivery. The store-and-forward option, also called asynchronous telemedicine, is based on the data transmission of patient information and is often used in the medical fields including radiology, pathology, etc. Remote monitoring relies on technical devices that remotely monitor the health concerns of patients. The last type is real-time (or synchronous) interactive services, which are similar to traditional in-person appointments except usually provided remotely. Telemedicine in this paper only refers to the synchronous option unless otherwise noted.

Despite its recent importance, there is relatively little literature on telemedicine from the perspective of operations research and applied probability. For instance Zhou et al. 2021 analyze whether it is profitable from a social welfare perspective to introduce telemedicine options for patients within a healthcare provider. Their work is closely related to the literature on the economic feasibility of telemedicine and is comparable to work by Dowie et al. 2008 (both synchronous and asynchronous options), Labiris et al. 2005, Sun et al. 2020, and Theodore et al. 2015. These studies empirically investigate the cost effectiveness and socioeconomic benefits of telemedicine. However, there are limited theoretical studies on telemedicine services. These studies include Tarakci et al. 2009, who use approximations in queueing theory to explore the optimal investment level of telemedicine and staffing policies when considering various cost components, including staffing, technology investment, incorrect treatment, and waiting. Rajan et al. 2013 analyze the impact of telemedicine on the market share of a specialty hospital deploying this technology and on its competing hospitals in the region. Later, Rajan et al. 2019 turn to explore the impact of telemedicine on the speed-quality trade-off when chronically ill patients have heterogeneous treatment utilities in both revenue-maximizing and welfare-maximizing settings. In addition, their technical contribution includes analyzing queueing systems with rational customers and establishing the equilibrium behavior of patients when introducing telemedicine. Based on stochastic modeling, recently Zychlinski et al. 2023 study a hybrid hospital that includes in-person and virtual meetings with physicians and present staffing methods for this situation.

1.1 Contributions of Our Work

By using stochastic simulation, we address the following server staffing related problems of minute-clinics:

- What proportion of time will a minute clinic need the help of a collaborating physician?
- How many collaborating physicians are needed to satisfy a quality guarantee in a network of minute clinics?
- How do state regulations impact the staffing of a network of minute clinics?

1.2 Organization of the Paper

The remainder of the paper is organized as follows. Section 2 introduces a single server minute clinic model, which is adapted from CVS-Aetna minute clinics. We also derive the proportion of time that a collaborating physician is needed to provide additional care. In Section 3, we present a new model for a network of minute clinics. Here we also provide Gaussian-based approximations and Binomial approximations for the staffing problem. We use simulation to verify that these approximations are quite accurate at estimating the number of collaborating physicians needed to satisfy the quality of service that is required. Section 4 gives an example of a minute clinic network with restrictions to illustrate the influence of different types of state regulations on staffing results, by using Gaussian and binomial-based approximations and simulation. Section 5 summarizes the staffing results with more performance measures in an Excel-based PivotTable with real data inputs from CVS-Aetna groups, demonstrating the advantages of full independence states. Finally, a conclusion is given in Section 6 along with some new directions for future research.

2 TELEMEDICINE WITHIN A SINGLE MINUTE CLINIC

It is common in clinics that upon arriving patients meet a licensed practical nurse (LPN). The LPNs do basic intake by taking the patient's height, weight, temperature, blood pressure, etc. If the patient's visit is for something simple, like administering a vaccine, they receive this service and leave the clinic. If they need further medical attention, they are passed to a queue waiting for an initial screening/diagnosis from an NP. To avoid a large number of model parameters, our simulation only models the phase after the clinical intake as depicted by Figure 1.

2.1 Stochastic Model

In this subsection, we describe a telemedicine operation with dedicated doctors in a single minute clinic. Patients arrive to see NPs for the initial consultation at the telemedicine station (beyond the LPN station) at rate λ per unit time and join the queue of waiting patients if one exists. All the services are conducted in a first come first serve fashion. Once a patient reaches the front of the queue, they interact with an NP who does an initial diagnosis of the health concerns of the patient with service rate μ_0. Starting from this point, this NP guides and remains with the patient until the patient finishes all their service and leaves the clinic. Upon completion of the initial consultation, the NP understands (through examination) whether or not the patient needs collaborative service with a (possibly online) CP (independent of all other patients). This additional consultation with the CP is required in cases where the NP is unsure about possible treatments and needs additional advice. If the patient is classified as low acuity and does not require the CP's expertise, the NP-patient pair can still go to the collaborative station for a second opinion to verify their initial diagnosis.

If the NP-patient pair ends up going to the collaborative station, where we assume the probability is p, they (both) join the queue waiting for a CP to become available; otherwise (with probability $1-p$), the NP sees the patient themselves. The service rates of non-collaboration and collaboration are denoted as μ_1 and μ_2, respectively. Figure 1 provides a visual description of this stochastic model.

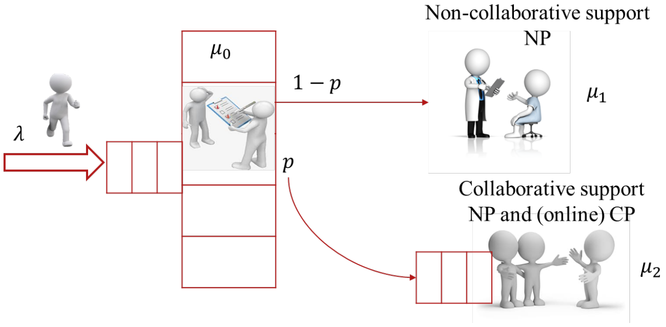


Figure 1: Diagram of (telemedicine) collaborative support model.

2.2 Closed-form Approximation for a G/G/m Queue

Now consider a clinic location equipped with m NPs and m CPs so that the queueing dynamics become a general G/G/m queue. We have the following proposition.

Proposition 1 Suppose $\rho := \frac{\lambda}{m}\left(\frac{1}{\mu_0} + \frac{1-p}{\mu_1} + \frac{p}{\mu_2}\right) < 1$. The proportion of time (or equivalently, the probability in the steady state) that any NP spends with CPs equals $\frac{\lambda p}{m\mu_2}$. Hence the expected number of NPs at the collaborative station in the steady state is $\frac{\lambda p}{\mu_2}$.

Proof. For any $j \in \{1,...,m\}$, define

$$X_j(t) = \begin{cases} 1, & \text{if at time } t \text{ the } j^{th} \text{ NP is busy,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

and

$$\tilde{X}_j(t) = \begin{cases} 1, & \text{if at time } t \text{ the } j^{th} \text{ NP is busy at collaborative station,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Define also $Y(t)$ as the number of NPs at the collaborative station at time t. Clearly $Y(t) = \sum_{j=1}^{m} \tilde{X}_j(t)$. Notice that the service times follow a distribution with mean $\frac{1}{\mu_0} + \frac{1}{\mu_1}$ with probability $1-p$ and with mean $\frac{1}{\mu_0} + \frac{1}{\mu_2}$ with probability p. Hence the utilization is $\rho = \frac{\lambda}{m}\left(\frac{1}{\mu_0} + \frac{1-p}{\mu_1} + \frac{p}{\mu_2}\right)$ which is assumed less than one. Assume that the minute-clinic is operating in stationarity (so no time dependence). Note that $\mathbb{P}(X_j = 1) = \rho$. Consider the proportion of those services that are with CPs yields,

$$\mathbb{P}(\tilde{X}_j = 1 | \overbrace{\text{The } j^{th} \text{ NP is busy}}^{X_j=1}) = \frac{\frac{p}{\mu_2}}{\frac{1}{\mu_0} + \frac{1-p}{\mu_1} + \frac{p}{\mu_2}}.$$

Thus,

$$\begin{aligned} \mathbb{P}(\tilde{X}_j = 1) &= \mathbb{P}(\tilde{X}_j = 1 | \text{The } j^{th} \text{ NP is busy})\mathbb{P}(\text{The } j^{th} \text{ NP is busy}) \\ &= \mathbb{P}(\tilde{X}_j = 1 | \text{The } j^{th} \text{ NP is busy})\rho \\ &= \frac{\lambda p}{m\mu_2} =: \tilde{\rho}. \end{aligned}$$

Therefore,

$$\mathbb{E}[Y] = \mathbb{E}[\sum_{j=1}^{m} \tilde{X}_j] = \sum_{j=1}^{m} \mathbb{E}[\tilde{X}_j] = \sum_{j=1}^{m} \mathbb{P}(\tilde{X}_j = 1) = \frac{\lambda p}{\mu_2} = m\tilde{\rho}.$$

□

As an illustrative example, consider an M/G/m queue where the inter-arrival times and all the service times are exponential and $m = 10$. Let $\lambda = 20$, $p = 0.5$. Table 1 compares the proportion of time that an NP spends at the collaborative station given by the formula in Proposition 1 and by simulation. Table 2 summarizes the expected number of NPs at the collaborative station in the steady state in a single clinic location. The simulation runs for 4,000 unit time and 20,000 replications.

Table 1: Proportion of time an NP spends at the collaborative station in a clinic.

	Parameters	Theoretical Formula	Simulation
Case 1	$\mu_0 = \mu_1 = \mu_2 = 10$	0.1000	0.1007 ± 0.0014
Case 2	$\mu_0 = \mu_1 = 10, \mu_2 = 5$	0.2000	0.2002 ± 0.0019
Case 3	$\mu_0 = \mu_2 = 10, \mu_1 = 8$	0.1000	0.1005 ± 0.0014
Case 4	$\mu_1 = \mu_2 = 8, \mu_0 = 10$	0.1250	0.1259 ± 0.0015
Case 5	$\mu_0 = 10, \mu_1 = 8, \mu_2 = 5$	0.2000	0.2001 ± 0.0019

Table 2: Expected number of NPs at the collaborative station in the steady state and the staffing in a clinic.

	Parameters	Theoretical Formula	Simulation	Number of CPs needed
Case 1	$\mu_0 = \mu_1 = \mu_2 = 10$	1.0000	1.0067 ± 0.0138	3
Case 2	$\mu_0 = \mu_1 = 10, \mu_2 = 5$	2.0000	2.0023 ± 0.0194	5
Case 3	$\mu_0 = \mu_2 = 10, \mu_1 = 8$	1.0000	1.0053 ± 0.0138	3
Case 4	$\mu_1 = \mu_2 = 8, \mu_0 = 10$	1.2500	1.2585 ± 0.0154	3
Case 5	$\mu_0 = 10, \mu_1 = 8, \mu_2 = 5$	2.0000	2.0013 ± 0.0193	5

Consider the probability that an NP-patient pair has to wait for an available CP in the steady state. We would like to staff CPs so that this performance measure is below some threshold ε. At one clinic location, the objective of staffing is to find the minimum k' such that

$$\mathbb{P}(Y > k') \leq \varepsilon \tag{3}$$

At each replication, we save the number of NPs at the collaborative station when the simulation ends and return the one that is $100 \cdot \varepsilon$ percent largest among all replications. The staffing outputs are presented in Table 2 as well.

3 TELEMEDICINE IN A NETWORK OF MINUTE CLINICS

Now that we have a good understanding of one minute clinic, it is reasonable to ask about multiple minute clinics and the interaction among them. In this subsection, we demonstrate how to use the one minute clinic model to extend our staffing approach to many minute clinics. In the large minute clinic setting, it is important to determine how many doctors are needed for each state. Figure 2 provides a picture of a large network of minute clinics, where each clinic location may have its own parameters. In reality, the minute clinics do not interact with one another except through the availability of the CPs that provide collaborative service. We assume all doctors can work between the clinics as floaters, for example, by working virtually. That is to say, the patients' arrival is still local and all the NPs are dedicated to the local clinic while all the CPs can be accessed within the same network in a telemedicine mode.

In many ways the network setup is optimal from a usage perspective since all minute clinics have access to all CPs and this setup provides complete resource pooling. However, the complete resource pooling perspective complicates the staffing decision problem considerably. On the one hand, drawing on the broader network may reduce the wait until connecting with a CP. Therefore, the remainder of the paper aims at finding an efficient way of staffing CPs in a network and investigating the impact of various types of regulations on service quality, which is measured by the probability of waiting for available CPs, the waiting times, the proportion of CPs' idling time, etc.

3.1 Gaussian Based Approximation for a Network of M/G/1 Queues

As a first approximation, consider a network of N M/G/1 queues in this and the following section, where the inter-arrival times and the service times are exponential, and $m = 1$. Now a subscript i is added in the definitions of variables and parameters above to indicate the i^{th} location, $i = 1,...,N$. Notice that the

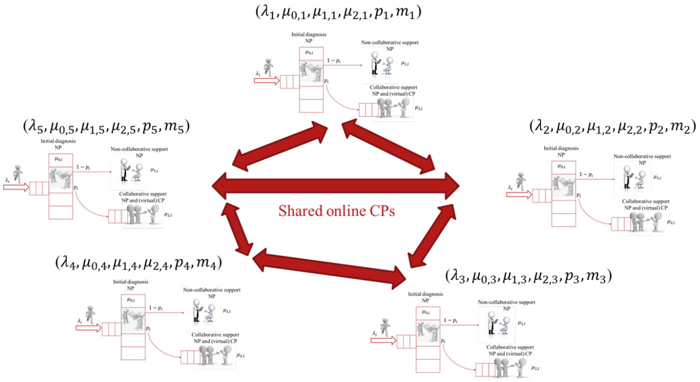

$(\lambda_1, \mu_{0,1}, \mu_{1,1}, \mu_{2,1}, p_1, m_1)$
$(\lambda_2, \mu_{0,2}, \mu_{1,2}, \mu_{2,2}, p_2, m_2)$
$(\lambda_3, \mu_{0,3}, \mu_{1,3}, \mu_{2,3}, p_3, m_3)$
$(\lambda_4, \mu_{0,4}, \mu_{1,4}, \mu_{2,4}, p_4, m_4)$
$(\lambda_5, \mu_{0,5}, \mu_{1,5}, \mu_{2,5}, p_5, m_5)$
Shared online CPs
Initial diagnosis NP
Non-collaborative support NP
Collaborative support NP and (virtual) CP
λ_i
$\mu_{0,i}$
$1 - p_i$
p_i
$\mu_{1,i}$
$\mu_{2,i}$

Figure 2: A network of telemedicine sites. Links represent the access from NPs to online CPs in a different physical location in the same network.

assumption that each location is equipped with one NP and one CP implies that no NP is made to wait for a CP beyond the initial diagnosis. That is to say, each node in the network is independent of the other nodes (and the random variables $X_i(t)$ are independent, so are the $\tilde{X}_i(t)$'s). Recall that $\mathbb{E}[\tilde{X}_i] = \tilde{\rho}_i$ and $\text{Var}(\tilde{X}_i) = \tilde{\rho}_i(1-\tilde{\rho}_i)$ since they are Bernoulli, where $\tilde{\rho}_i = \frac{\lambda_i p_i}{\mu_{2,i}}$.

Fix a quality of service level $\varepsilon > 0$, we seek to find the minimum k' such that

$$\mathbb{P}\left(\sum_{i=1}^{N} \tilde{X}_i > k'\right) \leq \varepsilon. \tag{4}$$

Since each location acts independently, the Lindeberg-Feller central limit theorem (c.f. Theorem 27.2 in Billingsley 2017) implies that

$$Z_N = \frac{\sum_{i=1}^{N} \left(\tilde{X}_i - \tilde{\rho}_i\right)}{\sqrt{\sum_{i=1}^{N} \tilde{\rho}_i(1-\tilde{\rho}_i)}} \tag{5}$$

follows approximately a standard normal distribution. A little algebra in Equation (4) yields

$$\mathbb{P}\left(Z_N > \frac{k' - \sum_{i=1}^{N} \tilde{\rho}_i}{\sqrt{\sum_{i=1}^{N} \tilde{\rho}_i(1-\tilde{\rho}_i)}}\right) \leq \varepsilon. \tag{6}$$

Since CVS-Aetna has over 1000 minute clinics in the United States, using the normal approximation, we have from Equation 6 that

$$\frac{k' - \sum_{i=1}^{N} \tilde{\rho}_i}{\sqrt{\sum_{i=1}^{N} \tilde{\rho}_i(1-\tilde{\rho}_i)}} \geq \Phi^{-1}(1-\varepsilon),$$

where Φ^{-1} is the inverse cdf of the standard normal distribution. That is, in order to meet the $\varepsilon-$constraint, we could seek the minimum k' such that

$$k' \geq \sum_{i=1}^{N} \tilde{\rho}_i + \Phi^{-1}(1-\varepsilon)\sqrt{\sum_{i=1}^{N} \tilde{\rho}_i(1-\tilde{\rho}_i)}. \tag{7}$$

The expression for the number of collaborating physicians is similar to that of the square root staffing rule from Jennings et al. 1996, Wallace and Whitt 2005, Feldman et al. 2008 and Niyirora and Pender 2016. It is our hope that our Gaussian-based approach is close to the results obtained by using stochastic simulation. Before we compare our results, we also present an additional approximation based on the bounds of the Poisson-Binomial distribution.

3.2 Binomial Based Approximation for a Network of M/G/1 Queues

In this section, we provide another approximation for a network of M/G/1 queues that model our telemedicine network. We know from Tang and Tang 2023 that we can bound the tail cdf of the Poisson-Binomial with the Binomial i.e.

$$\mathbb{P}\left(\sum_{i=1}^{N}\tilde{X}_i > k'\right) \leq \mathbb{P}\left(\sum_{i=1}^{N} Y_i > k'\right) = \mathbb{I}_{\bar{\rho}}(N-k',k'+1) \qquad N\bar{\rho} \leq k' \leq N \tag{8}$$

where $\mathbb{I}_{\bar{\rho}}(N-k',k'+1)$ is the incomplete beta function and $\bar{\rho} = \frac{1}{N}\sum_{j=1}^{N}\tilde{\rho}_j$. Thus, by setting

$$\mathbb{I}_{\bar{\rho}}(N-k',k'+1) \quad = \quad \varepsilon$$

and solving for k', then we obtain the following table of staffing values. See Table 3.

3.3 Numerical Example for a Network of M/G/1 Queues

We provide the parameters that we use for the network simulation of minute clinics in the sequel. To simplify, we consider three different types of clinic stations each with one NP. In particular, **Station 1** = $(\lambda = 1, \mu_0 = 5, \mu_1 = 5, \mu_2 = 2, p = 0.2)$, **Station 2** = $(\lambda = 2, \mu_0 = 10, \mu_1 = 8, \mu_2 = 6, p = 0.4)$, and **Station 3** = $(\lambda = 3, \mu_0 = 20, \mu_1 = 8, \mu_2 = 10, p = 0.5)$. By computing the utilization for each station type, one can notice that these stations are stable. We also consider three 100-station networks each with varying proportions of the three types, **Mix 1** = (20,30,50), **Mix 2** = (50,20,30), and **Mix 3** = (30,50,20).

Assume all clinics are equipped with one CP so that the collaborative service starts whenever needed without waiting. Similarly as a staffing in a single minute clinic, we consider the number of NPs that are at the collaborative station at the end of simulation time for each replication, and output the number at the upper ε quantile over all the replications. The simulation runs for a long time to approximate the steady state. Table 3 summarizes the staffing results in three networks of M/G/1 queues given by the Gaussian and Binomial staffing formula and the simulation to satisfy different levels of probability guarantees. We observe that both of the theoretical approximations are rather accurate in estimating the number of CPs required for the desired quality of service, where the Binomial staffing formula is in general closer to the simulated results since it is non-asymptotic.

Table 3: The number of CPs required in three networks to satisfy the performance guarantees.

ε	(Gaussian for three mixes)	(Binomial for three mixes)	(Simulation for three mixes)
$\varepsilon = .25$	(16,15,15)	(16,14,15)	(16,14,15)
$\varepsilon = .20$	(17,15,16)	(16,15,15)	(16,15,16)
$\varepsilon = .15$	(18,16,17)	(17,16,16)	(17,16,16)
$\varepsilon = .10$	(18,17,17)	(18,16,17)	(18,17,17)
$\varepsilon = .05$	(20,18,19)	(19,18,18)	(19,18,18)
$\varepsilon = .01$	(22,20,21)	(22,20,21)	(22,21,22)
$\varepsilon = .001$	(25,23,23)	(25,23,24)	(24,24,24)

4 MINUTE CLINIC NETWORK WITH CONSTRAINTS

So far, we have explored a simple network model for minute clinic staffing. However, navigating the intricacies of telemedicine introduces many new challenges, primarily centered around the complex interplay of staffing within the confines of state regulation. Particularly in the context of minute clinics, adherence to state-specific regulations, like physician licensing, is a crucial determinant of the provision of telemedicine services.

This regulatory landscape is visually depicted in schematic form in Figure 3, where states with restricted practice are marked in red and exhibit strict regulations, limiting physicians without a state license from providing care in that state. In contrast, states highlighted in green demonstrate full flexibility, which empowers CPs to provide telehealth service within all green states. In other words, the minute clinics located in one of the states with full practice have complete access to CPs in any green state. Reduced practice states are marked in yellow, reducing the flexibility of CPs working out of states. Managing staffing across the entire network becomes inherently complex due to these regulatory constraints, resulting in an increased demand for physicians.

This intricate interplay between regulatory constraints and staffing demands prompts a crucial question: to what extent does this impact staffing requirements?

In order to staff appropriately, we develop an approach that will satisfy all constraints. In this approach, we first staff the restrictive states one by one and find enough agents such that the probabilistic threshold is satisfied. Then we staff all of the unrestricted states as one large pool. To simplify, consider one pool of one restrictive state with N_R clinics and one pool of flexible (full practice) states with N_F clinics. From our central limit theorem approximation, we have the number of physicians needed in the restrictive pool equal to

$$k'_R \approx \sum_{i=1}^{N_R} \tilde{\rho}_i + \Phi^{-1}(1-\varepsilon)\sqrt{\sum_{i=1}^{N_R} \tilde{\rho}_i(1-\tilde{\rho}_i)} \tag{9}$$

and the number of physicians needed in the flexible pool is equal to

$$k'_F \approx \sum_{i=1}^{N_F} \tilde{\rho}_i + \Phi^{-1}(1-\varepsilon)\sqrt{\sum_{i=1}^{N_F} \tilde{\rho}_i(1-\tilde{\rho}_i)}. \tag{10}$$

Finally, if the two pools were both flexible, the number of physicians needed are

$$k'_T \approx \sum_{i=1}^{N_F+N_R} \tilde{\rho}_i + \Phi^{-1}(1-\varepsilon)\sqrt{\sum_{i=1}^{N_F+N_R} \tilde{\rho}_i(1-\tilde{\rho}_i)}. \tag{11}$$

Theorem 1 We have the following bounds on the difference $k'_F + k'_R - k'_T$.

$$0 \le k'_F + k'_R - k'_T \quad \lessapprox \quad (2-\sqrt{2})\Phi^{-1}(1-\varepsilon)\max\left(\sqrt{\sum_{i=1}^{N_F} \tilde{\rho}_i(1-\tilde{\rho}_i)}, \sqrt{\sum_{i=1}^{N_R} \tilde{\rho}_i(1-\tilde{\rho}_i)}\right). \tag{12}$$

Proof. First, using the triangle inequality for square roots for any $x \ge 0$ and $y \ge 0$, $\sqrt{x}+\sqrt{y} \ge \sqrt{x+y}$. Next, note that without loss of generality $x \ge y$ that

$$\begin{aligned} \sqrt{x}+\sqrt{y}-\sqrt{x+y} &\le \max_{0\le r\le 1} \sqrt{x}+\sqrt{xr}-\sqrt{x+xr} && (13)\\ &= \max_{0\le r\le 1}\left(1+\sqrt{r}-\sqrt{1+r}\right)\sqrt{x} && (14)\\ &= (2-\sqrt{2})\sqrt{x}, && (15)\end{aligned}$$

where the last step follows because the function $f(r) = 1+\sqrt{r}-\sqrt{1+r} = 1-\frac{1}{\sqrt{r}+\sqrt{1+r}}$ is increasing in r, where the maximum is attained when $r=1$ or $x=y$. The remainder of the result follows from substituting $x = \sum_{i=1}^{N_R} \tilde{\rho}_i(1-\tilde{\rho}_i)$ and $y = \sum_{i=1}^{N_F} \tilde{\rho}_i(1-\tilde{\rho}_i)$, so that the maximum is attained when $N_R = N_F$. □

Note that in the case where there are M different types, for instance, $M-1$ restrictive states and 1 pool of flexible states, then

$$\sum_{j=1}^{M} k'_j - k'_T \le \left(M-\sqrt{M}\right)\Phi^{-1}(1-\varepsilon)\sqrt{\max_{1\le j\le M}\left(\sum_{i=1}^{N_j} \tilde{\rho}_i(1-\tilde{\rho}_i)\right)}. \tag{16}$$

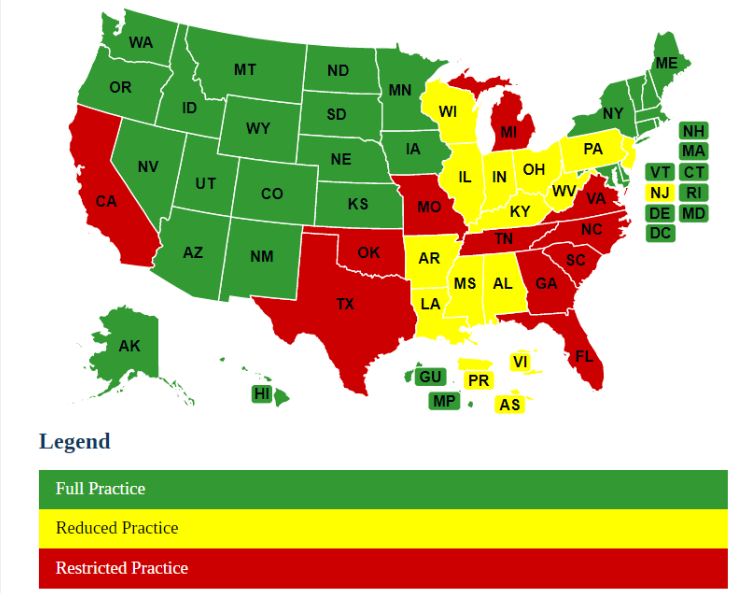


Figure 3: A map of locations and constraints.

Table 4: The number of CPs required in three networks to satisfy the performance guarantees under partial restriction.

α	ε	(Gauss 3 mixes) (Bin) (Sim) k'_R	(Gauss 3 mixes) (Bin) (Sim) k'_F	Difference
.1	.10	(3,3,3) (3,3,3) (3,3,3)	(17,15,16) (16,15,16) (16,15,16)	(2,1,2) (1,2,2) (1,1,2)
.1	.05	(4,3,3) (3,3,3) (3,3,3)	(18,17,17) (18,16,17) (17,16,17)	(2,2,1) (2,1,2) (1,1,2)
.1	.01	(4,4,4) (4,4,4) (4,4,4)	(20,19,19) (20,19,19) (19,19,19)	(2,3,2) (2,3,2) (1,2,1)
.1	.001	(5,5,5) (5,5,5) (5,5,5)	(23,21,22) (23,21,22) (24,22,25)	(3,3,4) (3,3,3) (5,3,6)
.2	.10	(5,5,5) (5,4,4) (5,4,5)	(15,14,14) (15,14,14) (15,14,14)	(2,2,2) (2,2,1) (2,1,2)
.2	.05	(6,5,5) (5,5,5) (5,5,5)	(16,15,16) (16,15,15) (16,15,15)	(2,2,2) (2,2,2) (2,2,2)
.2	.01	(7,6,6) (7,6,6) (7,6,6)	(18,17,18) (18,17,18) (19,18,17)	(3,3,3) (3,3,3) (4,3,1)
.2	.001	(8,7,8) (8,8,8) (8,7,8)	(21,19,20) (21,20,20) (21,20,20)	(4,3,5) (4,5,4) (5,3,4)
.5	.10	(10,10,10) (10,9,9) (10,9,9)	(10,10,10) (10,9,9) (10,9,9)	(2,3,3) (2,2,1) (2,1,1)
.5	.05	(11,10,11) (11,10,10) (11,10,10)	(11,10,11) (11,10,10) (11,10,10)	(2,2,3) (3,2,2) (3,2,2)
.5	.01	(13,12,12) (13,12,12) (12,12,12)	(13,12,12) (13,12,12) (12,12,12)	(4,4,3) (4,4,3) (2,3,2)
.5	.001	(15,14,14) (15,14,14) (14,15,14)	(15,14,14) (15,14,14) (14,15,14)	(5,5,5) (5,5,4) (4,6,4)

With the same parameter configuration as Table 3, Table 4 compares the number of CPs needed in the restrictive setting to the case where it is not restrictive for the three mixes of M/G/1 queues. α is the fraction of the number of restricted clinics in the networks. The difference is computed correspondingly by $k'_R + k'_F - k'_T$, where k'_T is obtained from Table 3. Clearly, more physicians are needed when the physicians are restricted by state laws. From Table 4, we make three important observations. The first is that as ε decreases to zero, the gap between the restrictive and non-restrictive grows. This can be predicted from Theorem 1. The second observation is that the gap between the restrictive and non-restrictive increases as the fraction of the restrictive gets closer to 1/2, i.e., the number of clinics in the restrictive state and in the flexible pool is approximately equal. This also is predictable from Theorem 1. Finally, the last observation is that the difference between the restrictive and non-restrictive is smaller in the Binomial approximation than in the Gaussian approximation. This is most likely the case since the Binomial approximation is non-asymptotic while the Gaussian approximation is best when the number of stations is very large.

5 AN EXCEL-BASED TOOL FOR PRACTITIONERS

In this section, we describe a Microsoft-Excel-based PivotTable to compute the numbers of CPs that are required to guarantee the probability that the number of NP-patient pairs at the collaborative station exceeds the number of CPs at the steady state is within 5%. Notice this performance guarantee is in the sense of the state average.

The simulation can deal with any reasonable distribution of the inter-arrival times and service times, yet this example assumes exponential distributions for illustration. We embed the staffing results in a table with a drop-down menu, where all the parameters can be selected in different combinations. We run the simulation with all combinations of the following parameters: arrival rate $\lambda = 1.5, 2, 2.5$, initial diagnosis rate $\mu_0 = 12$, non-collaborative service rate $\mu_1 = 5$, collaborative service rate $\mu_2 = 4, 6$ and the probability of going to collaborative station $p = 0.2, 0.3, 0.4$. All the rates are measured in the unit per hour. For example, arrival rate $\lambda = 1.5$ assumes patients arrive at the initial diagnosis station every 40 minutes. Full practice or green states as depicted in Figure 3 are the states where CPs have full flexibility to engage in virtual collaborative service across all green states. By selecting this label true, only states with full practice are listed in the tool.

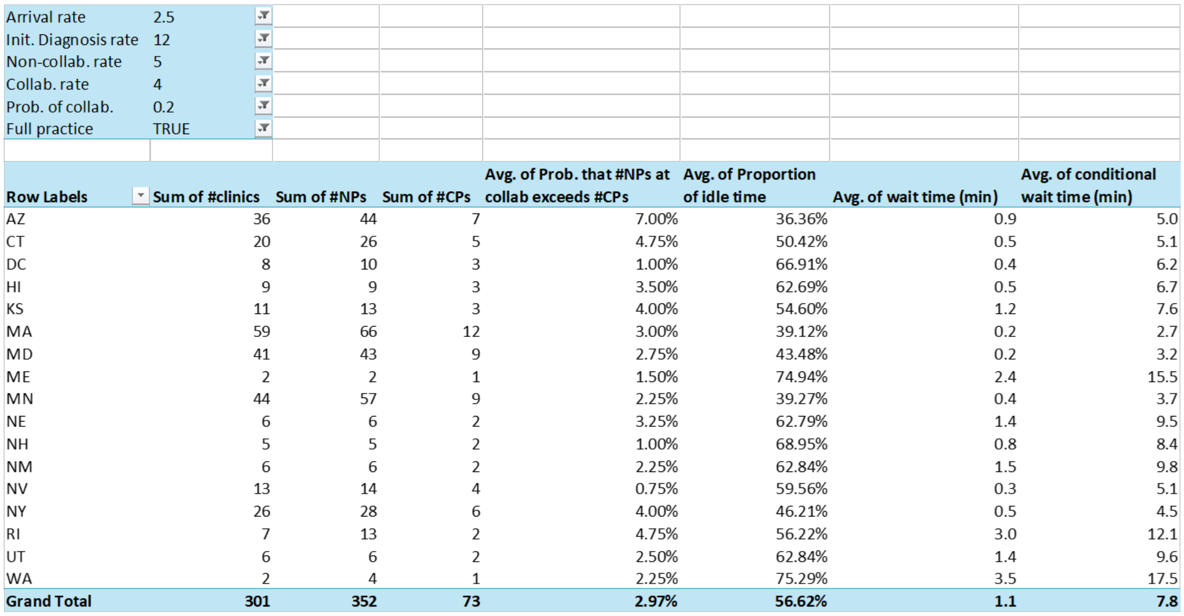
After specifying the parameters of interest, the statistics of the states that correspond to this choice of parameters are listed. The number of CPs needed to satisfy the performance guarantee is first calculated by the simulation per the same principles illustrated in previous sections. The performance metric that we care about, namely, the probability that the number of NPs at the collaborative station exceeds the number of CPs calculated is further demonstrated by generating another set of test samples. Other statistical outputs including the average proportion of idling time of the CPs and the average waiting times of patients for a CP are also shown to provide a better understanding of the staffing results. As opposed to the average among all patients, the output ‘conditional wait time’ is averaged among those who have to wait for collaborative service, i.e., the waiting time for collaboration is strictly greater than zero.

Figure 4 shows an example of a set of parameter combinations, where the performance metric that we care about is indeed within 5% for most states. For those this metric exceeds the 5% threshold, for example, Arizona, it is still within a reasonable range; so are the average waiting times for the availability of CPs. Recall that the green states can be combined in the sense that the staffing of physicians is across these states, namely, they can be considered as one single network. Consequently, under the same parameter configuration, only 47 CPs are required to sufficiently satisfy the performance guarantee as opposed to a simple summation of that for all green states, which yields 73, as if they were restrictive regulatory. The proportion of idling time of CPs is also reduced by staffing across all the green states and therefore the labor utilization is increased. Another observation is that the waiting times of patients for collaborative service are also reduced thanks to the flexible regulation, which indicates an improvement in service quality as well. See Figure 5 for a comparison.

6 CONCLUSION

In this paper, we study how to staff collaborating physicians to satisfy the quality of service measure, the probability of waiting for a collaborating physician. Managerially, our results offer guidance on how the manager of a network of minute clinics should staff the number of collaborating physicians in the presence of state regulation. Qualitatively, these insights may also be relevant in other staffing situations where there are restrictions to some of the servers.

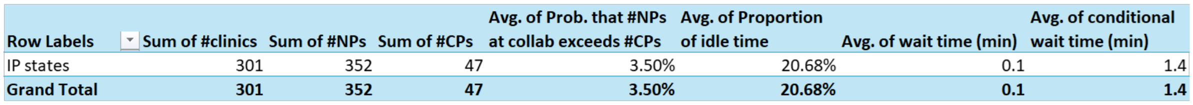
Our work also highlights new challenges that must be addressed in future work. First, our model assumes independent Poisson arrivals to each of the minute clinics. It would be interesting to extend this to either Poisson streams that are dependent or even non-Poisson streams like Hawkes processes or



Arrival rate	2.5
Init. Diagnosis rate	12
Non-collab. rate	5
Collab. rate	4
Prob. of collab.	0.2
Full practice	TRUE

Row Labels	Sum of #clinics	Sum of #NPs	Sum of #CPs	Avg. of Prob. that #NPs at collab exceeds #CPs	Avg. of Proportion of idle time	Avg. of wait time (min)	Avg. of conditional wait time (min)
AZ	36	44	7	7.00%	36.36%	0.9	5.0
CT	20	26	5	4.75%	50.42%	0.5	5.1
DC	8	10	3	1.00%	66.91%	0.4	6.2
HI	9	9	3	3.50%	62.69%	0.5	6.7
KS	11	13	3	4.00%	54.60%	1.2	7.6
MA	59	66	12	3.00%	39.12%	0.2	2.7
MD	41	43	9	2.75%	43.48%	0.2	3.2
ME	2	2	1	1.50%	74.94%	2.4	15.5
MN	44	57	9	2.25%	39.27%	0.4	3.7
NE	6	6	2	3.25%	62.79%	1.4	9.5
NH	5	5	2	1.00%	68.95%	0.8	8.4
NM	6	6	2	2.25%	62.84%	1.5	9.8
NV	13	14	4	0.75%	59.56%	0.3	5.1
NY	26	28	6	4.00%	46.21%	0.5	4.5
RI	7	13	2	4.75%	56.22%	3.0	12.1
UT	6	6	2	2.50%	62.84%	1.4	9.6
WA	2	4	1	2.25%	75.29%	3.5	17.5
Grand Total	**301**	**352**	**73**	**2.97%**	**56.62%**	**1.1**	**7.8**

Figure 4: Table of physician staffing results of IP states separately.



Row Labels	Sum of #clinics	Sum of #NPs	Sum of #CPs	Avg. of Prob. that #NPs at collab exceeds #CPs	Avg. of Proportion of idle time	Avg. of wait time (min)	Avg. of conditional wait time (min)
IP states	301	352	47	3.50%	20.68%	0.1	1.4
Grand Total	**301**	**352**	**47**	**3.50%**	**20.68%**	**0.1**	**1.4**

Figure 5: Table of physician staffing results of IP state combined.

ephemeral point processes, see for example Daw and Pender 2018, Koops et al. 2018, Daw and Pender 2022, and Daw et al. 2020. Moreover, there is interest in extending the theoretical approximations of the single server minute clinic model to a multi-server model and comparing those with the simulation outputs. Although most of the CVS-Aetna minute clinics are single NP due to the space limitations of the CVS stores, CVS-Aetna is continuing to expand their locations to accommodate the demand from patients. Finally, another extension worth pursuing would be to make the arrival process time-varying and depend on the spatial geography, see for example Liu and Whitt 2014, Pender 2016, Whitt 2018, and Besbes et al. 2022. This is because we know minute clinics experience more demand during the winter months, especially after long holidays since there is more interaction between people or more spreading of diseases and viruses.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Nairobi Mattindale, Senior Medical Director at CVS Health, and Dr. Sonny Morton, Regional Medical Director at MinuteClinic, for their invaluable expertise and understanding of CVS MinuteClinics. Their insights have greatly enhanced and shaped this work.

REFERENCES

Besbes, O., F. Castro, and I. Lobel. 2022. “Spatial capacity planning”. *Operations Research* 70(2):1271–1291.

Billingsley, P. 2017. *Probability and measure*. John Wiley & Sons.

Daw, A., A. Castellanos, G. B. Yom-Tov, J. Pender and L. Gruendlinger. 2020. “The co-production of service: Modeling service times in contact centers using Hawkes processes”. *arXiv preprint arXiv:2004.07861*.

Daw, A. and J. Pender. 2018. “Queues driven by Hawkes processes”. *Stochastic Systems* 8(3):192–229.

Daw, A. and J. Pender. 2022. “An ephemerally self-exciting point process”. *Advances in Applied Probability* 54(2):340–403.

Dowie, R., H. Mistry, T. A. Young, R. C. Franklin and H. M. Gardiner. 2008. “Cost implications of introducing a telecardiology service to support fetal ultrasound screening”. *Journal of telemedicine and telecare* 14(8):421–426.

Feldman, Z., A. Mandelbaum, W. A. Massey, and W. Whitt. 2008. “Staffing of time-varying queues to achieve time-stable performance”. *Management Science* 54(2):324–338.

Haddad, T. C., R. N. Blegen, J. E. Prigge, D. L. Cox, G. S. Anthony, M. A. Leak *et al*. 2021. “A scalable framework for telehealth: the Mayo Clinic Center for Connected Care response to the COVID-19 pandemic”. *Telemedicine Reports* 2(1):78–87.

Jennings, O. B., A. Mandelbaum, W. A. Massey, and W. Whitt. 1996. “Server staffing to meet time-varying demand”. *Management Science* 42(10):1383–1394.

Koops, D. T., M. Saxena, O. J. Boxma, and M. Mandjes. 2018. “Infinite-server queues with Hawkes input”. *Journal of Applied Probability* 55(3):920–943.

Labiris, G., C. Tsitlakidis, and D. Niakas. 2005. “Retrospective economic evaluation of the hellenic áir force teleconsultation project”. *Journal of Medical Systems* 29(5):493–500.

Liu, Y. and W. Whitt. 2014. “Stabilizing performance in networks of queues with time-varying arrival rates”. *Probability in the Engineering and Informational Sciences* 28(4):419–449.

Locke, D. E., R. Khayoun, A. L. Shandera-Ochsner, A. Cuc, J. Eilertsen, M. Caselli *et al*. 2021. “Innovation inspired by COVID: a virtual treatment program for patients with mild cognitive impairment at Mayo Clinic”. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* 5(5):820–826.

Niyirora, J. and J. Pender. 2016. “Optimal staffing in nonstationary service centers with constraints”. *Naval Research Logistics (NRL)* 63(8):615–630.

Pender, J. 2016. “Risk measures and their application to staffing nonstationary service systems”. *European Journal of Operational Research* 254(1):113–126.

Rajan, B., A. Seidmann, and E. R. Dorsey. 2013. “The competitive business impact of using telemedicine for the treatment of patients with chronic conditions”. *Journal of Management Information Systems* 30(2):127–158.

Rajan, B., T. Tezcan, and A. Seidmann. 2019. “Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care”. *Management Science* 65(3):1236–1267.

Sun, S., S. F. Lu, and H. Rui. 2020. “Does telemedicine reduce emergency room congestion? Evidence from New York State”. *Information Systems Research* 31(3):972–986.

Tang, W. and F. Tang. 2023. “The Poisson binomial distribution—Old & new”. *Statistical Science* 38(1):108–119.

Tarakci, H., Z. Ozdemir, and M. Sharafali. 2009. “On the staffing policy and technology investment in a specialty hospital offering telemedicine”. *Decision Support Systems* 46(2):468–480.

Theodore, B. R., J. Whittington, C. Towle, D. J. Tauben, B. Endicott-Popovsky, A. Cahana *et al*. 2015. “Transaction cost analysis of in-clinic versus telehealth consultations for chronic pain: preliminary evidence for rapid and affordable access to interdisciplinary collaborative consultation”. *Pain Medicine* 16(6):1045–1056.

Wallace, R. B. and W. Whitt. 2005. “A staffing algorithm for call centers with skill-based routing”. *Manufacturing & Service Operations Management* 7(4):276–294.

Whitt, W. 2018. “Time-varying queues”. *Queueing models and service management* 1(2).

Zhou, C., Y. Hao, Y. Lan, and W. Li. 2021. “To Introduce or Not? Strategic Analysis of Hospital Operations with Telemedicine”. *European journal of operational research* 304(1):292–307.

Zychlinski, N., G. Mendelson, and A. Daw. 2023. “The Hybrid Hospital: Balancing On-Site and Remote Hospitalization”. https://rgdoi.net/10.13140/RG.2.2.27593.11369, accessed 20th August 2024.

AUTHOR BIOGRAPHIES

SHUWEN LU is a PhD student in Systems Engineering at Cornell University. Her research interests are in operations research, applied probability, and machine learning. Her e-mail address is sl3243@cornell.edu.

MARK E. LEWIS is the Maxwell M. Upson Professor in Operations Research and Information Engineering at Cornell University. His research interests are in operations research, MDPs, and applied probability. His e-mail address is mel47@cornell.edu. His website is https://people.orie.cornell.edu/melewis/.

JAMOL PENDER is an Associate Professor in Operations Research and Information Engineering (ORIE) at Cornell University. He earned his PhD in the Department of Operations Research and Financial Engineering (ORFE) at Princeton University. His research interests include queueing theory, stochastic simulation, dynamical systems, and applied probability. His e-mail address is jjp274@cornell.edu. His website is https://blogs.cornell.edu/jamolpender/.