Queueing Information Management of Primary Care Delivery With E-Visits

Aditya Mahadev Prakash, *Member, IEEE*, and Xiang Zhong, *Member, IEEE*

***Abstract*— The goal of this work is to investigate the system configuration and information management of primary care delivery with electronic visits (e-visits). We consider a medical institution employing primary care physicians and other clinicians that offer office visits (in-person) and e-visits (through secure messaging from patient portals), and where different queue-joining behaviors: denoted as the mixed strategy, the duplication strategy and the threshold strategy are adopted by flexible patients based on different system configurations. Different queueing models are developed to capture flexible patients' queue-joining behaviors according to queueing information provision. In particular, we develop the equilibrium behavior of a dual server system where state information is available for one of the servers and the flexible patient exhibits a utility-maximizing behavior, which extends the literature on the analysis of queueing systems with strategic customers. The duplication strategy with deletion offers the least expected waiting time for the patients, and the threshold strategy provides the next best performance which is superior to the mixed strategy, which demonstrates the value of information. *Note to Practitioners*—We present a novel analytical framework for modeling the primary care delivery system and obtain the equilibrium patient flows under different queue-joining behaviors. This framework enables a rigorous analytical investigation of system configurations and their influence on system performance under a reasonable level of abstraction. System efficiency can be improved by taking advantage of patients' heterogeneity in care preferences and time sensitivity. Queue information management and coordination of servers are found to be crucial in achieving the best efficiency, especially with growing flexible customers in the population. The methodology and analysis put forth in this study provide actionable insights into care delivery planners engaged in facilitating e-visits, especially during the COVID pandemic.**

***Index Terms*— Capacity planning, doubly stochastic Poisson process, e-visits, primary care, queueing.**

I. Introduction

With the rapid growth of patient population and the acute shortage of primary care physicians (PCPs), declining access to care is becoming a menace for both patients and care providers in the United States [1]. Due to long appointment delays, limited after-hours care at physicians' offices, and other access barriers, patients seek a significant amount of nonemergent care at emergency departments [2] or conduct self-diagnosis [3], and more patients do not show up for their appointment or do not seek future services [4]. The primary care system risks being mired in excessive costs, adverse patient outcomes, and poor patient retention and this warrants a reform.

With the advent of digital health, an alternative service to office visits, e-visit, is gaining momentum. An e-visit is different from a video conferencing visit that requests the presence of both patients and care providers at the same time with the service quality close to an in-person visit. Instead, e-visits allow patients to answer a series of questions in an electronic patient portal (e.g., MyChart by EPIC) and submit them to a care provider. In this way, care providers are afforded greater flexibility in care delivery due to the asynchronous nature of e-visits. In addition, e-visits do not have to be managed by physicians, but can be offered by nurse practitioners or physician assistants, and the unit cost of a nonphysician provider is typically lower [2]. E-visits and similar telehealth services are growing rapidly during the COVID-19 pandemic as a way to screen, triage, and remotely monitor patients without the infectious risk [5]. Before COVID-19, less than 1% of all physician visits in the United States were conducted via telehealth and that number had spiked to over 50% and tapered off as the first wave of COVID-19 got suppressed. However, the demand (approximately 17%) has stayed high above the prepandemic level [6].

The wide adoption of e-visits and other telehealth services poses significant operational challenges to the primary care providers. Patients can differ in their health conditions, acuity of illness, socioeconomic factors, cultural preferences, and digital literacy in addition to possessing varying idiosyncratic traits [4]. These factors drive patients' preferences for care options. Patients can behave strategically in obtaining their care, depending on their perception of the quality of different care services, their travel burden, and the time-sensitivity of their symptoms. These behaviors also impact the delay in obtaining care experienced by the whole patient population and the effective market share of each service in equilibrium. On the other hand, patients' behaviors are affected by how the services themselves are configured. For instance, clinic closures and limited in-person appointments largely shaped the telehealth visit spike in April 2020 in the United States. In addition, the long appointment delay of an office visit informed by the scheduling system can also nudge the patient to use an e-visit in anticipation of a quick response. The capacity of appointments, the information made available to patients,

Manuscript received July 4, 2021; revised September 17, 2021; accepted September 22, 2021. This article was recommended for publication by Associate Editor J. Song and Editor J. Li upon evaluation of the reviewers' comments. The work of Xiang Zhong was supported in part by NSF under Grant CMMI-2027677. *(Corresponding author: Xiang Zhong.)*

The authors are with the Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: mahadevprakash90@ufl.edu; xiang.zhong@ise.ufl.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2021.3115355.

Digital Object Identifier 10.1109/TASE.2021.3115355

and the scheduling options all affect the strategies certain patients can take, leading to different equilibrium behaviors and patient diversions. As a result, utilizing the heterogeneity of patients and configuring the system accordingly are vital to improving the efficiency of the system.

In this article, we conceptualize the primary care delivery system as a queueing network with two parallel queues offering office visits and e-visits, respectively. This corresponds to the case where e-visits are offered by nonphysician providers, which is common in real practice [2]. To address the heterogeneity in e-visit eligibility and time-sensitivity, patient demands are segmented into *dedicated* office visits, *e-advocates*, and *flexible* encounters, where care options are evaluated on the basis of utility functions that measure the tradeoff between the reward obtained and the cost of waiting for each service. Without loss of generality, we consider three system configurations: 1) the information regarding the queue length (appointment delay or waiting time) is not explicitly offered for either service; then, flexible patients could employ a *mixed strategy* in choosing either service with a positive probability; 2) flexible patients are allowed to submit two requests (a *duplication strategy*), and cancel the redundant request upon receiving one of the services, which does not require the queue length information; and 3) the appointment delay of office visits is transparent to patients, whereas the waiting time of e-visits is not provided; then, flexible patients can adopt a *threshold strategy*: the patient chooses the service with the better expected utility which is based on the observed number of patients waiting for office visits.

These three settings correspond to our observation of the current service system designs. In particular, the threshold strategy is very close to what we have seen in real-world operations for patients who perceive the service quality of office visits adjusted by convenience (e.g., travel cost) higher than e-visits. Patients are time-sensitive, so they will first check the earliest open office visit appointment and decide whether to make the appointment, or use an e-visit instead. For some cases, patients need to overcome some barrier to obtain the queueing time information or simply cannot have it, for example, have to go through a complicated procedure to get the appointment information in the scheduling system, or need to call their clinic scheduler and wait in the phone line to enquire the available time. In this case, we can consider the queue length information to be not explicitly provided to patients, and patients might just make their decision (a mixed strategy) based on their past experience. This also allows them to game the system by submitting requests for both services if there is no penalty on withdrawal of appointments or no-show charged by insurance.

This article makes the following contributions to the literature.

1) We develop stochastic models and computationally efficient analytical methods to characterize the patient flows in equilibrium and compare system performance under different information provision and service settings.
2) Allowing patients to submit duplicate requests and the flexibility to cancel the redundant requests is shown to benefit all classes of patients, providing the withdrawn appointments can be immediately filled. This provides the best-case scenario as the system is work-conserving.
3) Our study provides a lucid demonstration of "the value of information" in flexible service systems, which is context-dependent—the observable office visit queue setting outperforms the nonobservable one only when the e-visit service capacity is below a certain level.
4) We identify the best configurations necessary to maximize patient welfare and explore the possible provider strategies that can be employed under cooperation or competition. In a setting where e-visits and office visits are competitive, interestingly, when absorbing the entire e-eligible market to e-visits is advantageous, hiding queue length information might be preferable to the interest of the medical institution but not patients.
5) In addition, having more flexible patients and increasing their flexibility improves the efficiency of the system universally. This is true for all strategies provided there exists sufficient e-visit capacity.
6) The proposed models enable us to analytically characterize the impact of e-visits on patient access to care, and the solutions developed can serve as a toolset to address a class of service configuration problems for strategic customers facing partially substitutable service options.

The remainder of the article is organized as follows. In Section II, we review the related literature. Section III discusses the assumptions of the model including Section I to the various strategies used by the patients. Section IV develops the analytical model for various strategies. The properties of the system are discussed in Section V and the article is concluded in Section VI. The proofs of all propositions can be found in the Appendix.

II. Literature Review

E-visits have garnered growing attention with the emerging portal technology. The COVID-19 pandemic has boosted the demand for telehealth services in a short period of time, with providers scrambling for service capacity [7], [8]. Such an adoption is expected to be irreversible with continued and increased use of e-visits in the future [9]. However, few studies have analyzed the service design of e-visits, which is instrumental to the sustainable and scalable e-visit implementation. Using queueing-based analysis, Liu and D'Aunno [10] studied the cost efficiency and productivity of involving nurse practitioners in primary care. A quantitative analysis of e-visits using a patient health dynamics model was developed by Bavafa *et al.* [11], which captures the usage of e-visits and nonphysician providers and quantifies physicians' expected earnings and patients' expected health outcomes. Rajan *et al.* [12] investigated the effect of telemedicine on chronic care considering the heterogeneous travel burden. The scheduling policies among e-visits and office visits using a vacation queue model was investigated in [13] and the flexible capacity allocation of e-visits was investigated in [14]. However, no analytical study on the patient flow of e-visits addresses the inherent segmentation of patients based on

their care preferences and investigates the implications of information management from an operational perspective.

In our study, the duplication strategy is modeled based on the analysis of [15]. A further development of queues with duplicated requests was presented in [16], which introduced different routing schemes like the "*M*," "*N*," and the "*W*" shaped network topologies. However, it was assumed in [16] that the jobs can be concurrently serviced by various servers, which is not feasible with patient care.

The other two strategies described in this article build on the foundational work on strategic queuing in [17]. Strategic queueing literature was further strengthened with the analysis of several queuing systems in [18]–[20]. The queueing-game-theoretic tools that have been applied in the healthcare domain include the representative work [21], involving cross-border patient movement for healthcare services; [22] which discusses concierge medicine and its impact on patient access; and [23], where nonurgent emergency visits are modeled as a queuing problem. A review of recent advancements in strategic queueing literature is provided in [24] and [25], and a survey on queueing literature with strategic arrivals is presented in [26].

The threshold strategy model is also closely related to the stream of literature on multiphase queues with nonhomogeneous arrival processes. A comprehensive review of operations research methods for modeling patient flow and outcomes can be referred to [27]. A bibliography of the application of queues with nonhomogeneous arrivals in various domains was presented by Whitt [28]. The development of the stochastic models in this work builds on the research on doubly stochastic Poisson processes and its application to queueing theory [29]–[31].

III. Model Assumptions

Primary care patients typically seek care from their dedicated care providers within the insurer's care network. In this study, we consider a medical institution that employs PCPs and nonphysician providers and offers care services (office and e-visits) to their panel patients. Assumptions on encounter classification, patient demand, service supply, and the strategies employed by patients are outlined below.

A. Encounter Classification

Practically, clinics measure patient demands by the volume of encounters and in what follows, we discuss encounter types. The most important driver for e-visits is the disease condition. A patient with a cold or skin rash is eligible for e-visits; however, the same patient could suffer migraine and need a botox injection which requires an on-site visit. For patients who are eligible to avail e-visit services, which we nominate as *e-eligible* encounters. Some patients are tech-savvy, or value the safety of online services, or perceive that the transportation barrier outweighs the benefit of having face-to-face encounters. If eligible, these patients strictly prefer e-visits over office visits, and the corresponding e-encounters are nominated as *e-advocates*. For instance, the retired population living in rural areas or with mobility issues, or patients

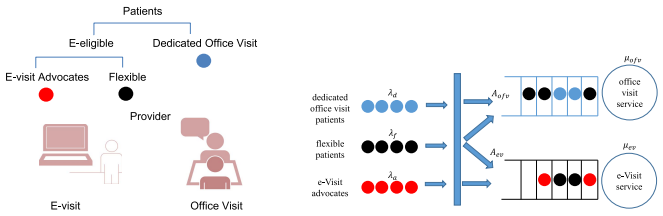

Fig. 1. Encounter classification and patient flow in primary care delivery with e-visits.

with existing conditions and are sensitive to infectious diseases, might strictly prefer e-visits over office visits. The rest of e-eligible patients are open to being served by either channels and we consider these encounters as *flexible*. For patients whose medical conditions are neither e-eligible nor time-sensitive, we nominate their encounters as dedicated office visits. A schematic for patient classification is provided in Fig. 1. We do not differentiate between the terms patient type and encounter type and use them interchangeably.

B. Patient Demand

We assume that the total encounters follow a homogeneous Poisson point process. While the arrival process to a physician's office might not be Poisson due to appointment-based scheduling, the original demand of patients is generally valid to be modeled as a Poisson arrival based on historical appointment data [10], [32], [33]. We denote the Poisson arrival intensities of *potential* encounters of dedicated office visits, flexible patients, and e-advocates as λ_d, λ_f, and λ_a, respectively.

The above patient segmentation can be estimated by analyzing historical patient records and based on qualitative surveys designed to elicit the preferences of patients between office and e-visits. For instance, a survey conducted by Nippon Telegraph and Telephone (NTT) DATA Corporation [34] reported that 76% of patients prioritize access to care over the need for human interactions with their care providers, and 70% of patients are comfortable communicating with their care providers via text, email, or video, in lieu of seeing them in person [35].

C. Service Supply

To describe the service process, we consider a queueing network with two servers and denote μ_{ofv} and μ_{ev} as the office visit and e-visit service rate, respectively. Because office visits are typically appointment-based, we are interested in the appointment delay (can be shifts, days, or weeks) rather than the waiting time in a physician's office (typically measured by minutes). Because the in-clinic waiting time is typically orders of magnitude smaller, we do not consider it in our study for the tactical-level decision support. For e-visits, no appointment is needed, but there is still a delay (can be shifts or days) in getting response due to limited service capacity and service time variability. In this study, we use appointment delay and waiting time interchangeably. The service rates herein act as a proxy for the nominal appointment capacity of a service provider per a scheduling unit (e.g., a clinical session), or the rate at which the physician is able to provide treatment to their

patients. This assumption follows the ones introduced in [36] and [37].

D. Strategies of Flexible Patients

Since the choices of e-advocates and dedicated office visit patients are fixed, we focus on the different strategies a flexible patient can adopt under various circumstances. A strategy is induced by how the system is configured and what information is available to them.

1) Mixed Strategy: Both service queues are not observable; however, patients are aware of the expected waiting time for each service. In equilibrium, the flexible patients can adopt a mixed strategy of joining each queue.

2) Duplication Strategy: The flexible patient is highly time-sensitive but is indifferent to the service quality and convenience (i.e., the total reward) of both services and submits requests to both services. The patient is treated by the server which is available first and the redundant request gets deleted immediately. In a way, the flexible patient is gaming and possibly exploiting the scheduling system. In this case, whether or not the queues are observable is immaterial.

3) Threshold Strategy: The office visit queue is observable but the e-visit queue is not. Here we assume that the e-visit queue is not observable because currently the system (vendor) does not inform patients how many messages are waiting to be responded to. Then, an arriving flexible patient (she) observes the office visit queue and joins the queue if the queue length is below a threshold; otherwise, the patient seeks the e-visit service (a threshold strategy). In practice, a patient can check the office visit appointment delay. Upon being notified that the next available slot is two weeks away, the patient will decide whether she can tolerate the long waiting time, or had better submit an e-visit instead. This asymmetric design is common in other service industries. For instance, customers who arrive at a fast-food restaurant (e.g., Subway) and find a long waiting line may opt to order the food online (e.g., using Tapingo) instead.

We assume patients are risk-neutral, their decisions are irrevocable, and retrials of balking patients and reneging of entering patients are not allowed, following [17]. The only exception is the duplication strategy, where the system deletes the duplicate request upon a copy is served. In addition, we assume the service system is closed, that is, there is no patient loss, and office and e-visit services are provided by the same medical institution. The scenario of patient loss and competitive office and e-visit services are discussed in Sections V-D and V-E, respectively.

IV. Queueing Model

Stochastic models are developed for each flexible patient strategy in this section.

A. Mixed Strategy Model

Flexible patients gain a reward R_{ev} (R_{ofv}) from receiving e-visit (office visit) services and have to pay a unit time cost of C_{ev} (C_{ofv}) for time spent in the system, including both waiting and service time. For ease of exposition, the reward represents the benefit of receiving care, adjusted by service quality and the cost of travel, and we let $C_{\text{ofv}} = C_{\text{ev}} = C$ and define $k = (R_{\text{ofv}} - R_{\text{ev}})/(C)$, which is the waiting cost adjusted reward difference. Flexible patients will determine the strategy such that their utilities are maximized.

Proposition 1: When there is ample office visit capacity, that is, $\mu_{\text{ofv}} > \lambda_d + \lambda_f$, and the waiting cost adjusted reward difference is large enough ($k \geq k_o$), all flexible patients choose office visits. When there is sufficient e-visit capacity, that is, $\mu_{\text{ev}} > \lambda_a + \lambda_f$, and the waiting cost adjusted reward difference is small enough ($k \leq k_e$), all flexible patients choose e-visits. Otherwise, an arriving flexible patient will adopt a mixed strategy, that is, she will join the office visit queue with probability p and the e-visit queue with probability $1-p$, $p \in (0, 1)$.

The effective arrival to the office visit queue is denoted as $A^{\text{mix}}_{\text{ofv}} = \lambda_d + p\lambda_f$ and that to the e-visit queue is $A^{\text{mix}}_{\text{ev}} = \lambda_a + (1-p)\lambda_f$. The mean waiting times in the system are: $W^{\text{mix}}_{\text{ofv}} = (1)/(\mu_{\text{ofv}} - A^{\text{mix}}_{\text{ofv}})$ and $W^{\text{mix}}_{\text{ev}} = (1)/(\mu_{\text{ev}} - A^{\text{mix}}_{\text{ev}})$. In equilibrium, $R_{\text{ofv}} - C_{\text{ofv}} W^{\text{mix}}_{\text{ofv}} = R_{\text{ev}} - C_{\text{ev}} W^{\text{mix}}_{\text{ev}}$, while ensuring that $A^{\text{mix}}_{\text{ofv}} < \mu_{\text{ofv}}$ and $A^{\text{mix}}_{\text{ev}} < \mu_{\text{ev}}$. The formula of p is provided in the Appendix.

If a patient cannot observe the congestion in each queue, she will randomize her option in a way that in equilibrium, she gets the same utility from both services. If the rewards and waiting costs are the same for both services, the patient experiences the same waiting time at both services in equilibrium. Eventually, all three classes of patients experience the same waiting time. This system configuration has the minimum information provision and is easy to implement.

B. Duplication Strategy Model

When a duplication strategy is allowed, the service system can be modeled as one with multitype servers and multiple jobs. We have a set of two servers $\mathcal{S} = \{M_{\text{ofv}}, M_{\text{ev}}\}$ representing office visit and e-visit servers, and job types $\mathcal{C} = \{d, f, a\}$ representing dedicated office visit, flexible, and e-advocate patients, respectively. Server M_{ofv} can serve patients of type $\mathcal{C}(M_{\text{ofv}}) = \{d, f\}$ and server M_{ev} can serve patients of type $\mathcal{C}(M_{\text{ev}}) = \{f, a\}$. Following [15], we derive a merged state space representation using the ordered tuple $(u_{ij}, M_i, t_{\text{j}}, M_{\text{j}})$, where the indices $(i, j) \in \{(\text{ev,ofv}), (\text{ofv,ev}), (\text{ev}, \emptyset), (\text{ofv}, \emptyset), (\emptyset, \emptyset)\}$, and $\emptyset$ denotes an empty set. M_i and M_j stand for the two servers that can serve a customer belonging to $\mathcal{C}(M_i)$ and $\mathcal{C}(M_j)$, respectively. The ordering of the tuple is critical and is read from right to left: server M_j is busy in service of a customer belonging to $\mathcal{C}(M_j)$, followed by t_j number of customers which can be served exclusively by M_j or $\mathcal{C}(M_j) - \mathcal{C}(M_i)$. This is followed by server M_i which is busy serving one customer belonging to a class it can serve, followed by u_{ij} number of customers of unspecified type which can be served either by M_i or M_j or $\mathcal{C}(M_j) \cup \mathcal{C}(M_i)$. It is worth noting that this state space representation is rather succinct and there is no need to specify the states of all entities including the ones that are being served. Using such a state space depiction, all possible

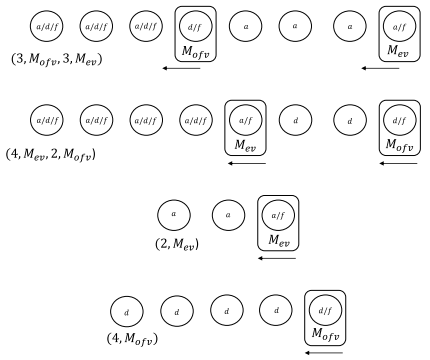

Fig. 2. Possible states of the queueing system where class "f" patients can be served by both servers M_{ofv} and M_{ev}.

transitions between states can be formulated and flow-balance equations can be derived.

A few sample states are illustrated in Fig. 2. For example, the first state $(3, M_{\text{ofv}}, 3, M_{\text{ev}})$ implies that, from the right, the e-visit server is serving either an e-advocate or a flexible patient. Waiting to be served are three e-advocates who can be served only by the e-visit server. In addition, the office visit server is busy and is treating either a dedicated office visit patient or a flexible patient. That is followed by three patients of the unidentified type who can seek services from either server. In total, there are six patients waiting and two patients in service. The steady-state probabilities and waiting times for each class of patients are derived and can be found in the Appendix.

Proposition 2: Dedicated office visit patients and e-advocate patients, depending on their states, have either zero waiting time, an exponential waiting time, or a waiting time that consists of a series of exponential waiting times as in a tandem queue. For flexible patients, if either of the server is free, there is zero waiting time and if both servers are busy, the waiting time is equal to that obtained from an M/M/1 queue with combined service rates and arrival rates.

This system configuration is very efficient since it is work-conserving. In addition, flexible patients experience the least waiting time compared to other classes. However, the system is configured in a way such that redundant requests get cleared and immediately make way for other appointments to take their place. Although patients do not need the queueing information, maintaining such a configuration is a challenging undertaking because one has to keep tracking the duplicated requests and fill in the withdrawn slots, which demands a sophisticated information management system coordinating the two services.

C. Threshold Strategy Model

The threshold strategy model follows the rational queueing literature by Naor [17]. If an arriving flexible patient observes n patients in the office visit queue (including the one in service), the expected utility is $R_{\text{ofv}} - ((n+1)C_{\text{ofv}})/(\mu_{\text{ofv}})$. The patient then compares the utility of requesting an e-visit, $R_{\text{ev}} - C_{\text{ev}}W_{\text{ev}}$, where W_{ev} is the expected waiting time incurred at the e-visit queue, including both waiting and service time. Then, flexible patients play a threshold strategy n_{thr}—an arriving flexible patient joins the office visit queue if she observes $n_{\text{thr}} - 1$ patients or less and joins the virtual e-visit queue (submits an e-visit) if she sees n_{thr} patients or more in the office visit queue. The threshold n_{thr} is determined by $n_{\text{thr}} = \underset{n\in\mathbb{Z}^+}{\text{argmin}}\{R_{\text{ofv}} - ((n+1)C_{\text{ofv}})/(\mu_{\text{ofv}}) < R_{\text{ev}} - C_{\text{ev}}W_{\text{ev}}\}$. It is worth noting that W_{ev} is a function of n_{thr} and e-visit service rate μ_{ev}.

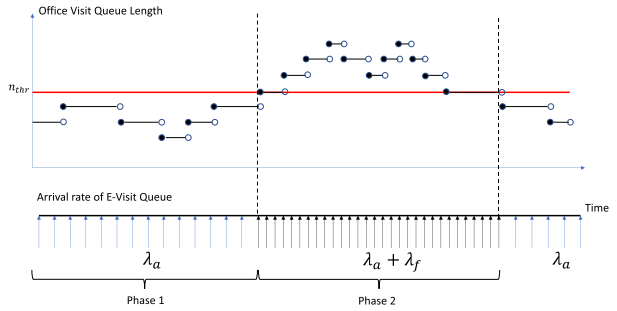

Fig. 3. Arrival process of the e-visit queue. This process is modulated by the queue length of the office visit queue and determined by the threshold n_{thr} chosen by the flexible patient.

Under the threshold strategy, the office visit queue, when viewed in isolation, can be modeled as a single server queue with an arrival rate that varies based on its queue length. Then, viewing the e-visit queue in isolation, the arrival to this queue is modulated by an alternating process. The "off" phase or phase 1 is when the arrival to the e-visit queue is equal to $\lambda_1 = \lambda_a$ and the "on" phase or phase 2 is when the arrival to the e-visit queue is equal to $\lambda_2 = \lambda_a + \lambda_f$. Specifically, phase 1 corresponds to the period of time the office queue length is strictly less than n_{thr} (the threshold), preceding and succeeding this period the queue length switches to a value greater than or equal to n_{thr}. Phase 2 corresponds to the amount of time the office queue length is greater than or equal to n_{thr}, preceding and succeeding this period the system is in phase 1.

Proposition 3: The arrival process to the e-visit queue follows a doubly stochastic Poisson process as depicted in Fig. 3 characterized by nonexponentially distributed sojourn times, the distribution of which is determined by the threshold n_{thr}.

Here, we present a renewal approximation of the doubly stochastic Poisson process. Following [30], the Laplace transform of the probability density function of the synchronous interevent time $\phi^*(s)$ is obtained as

$$\phi^*(s) = \frac{1}{\lambda_1 v_1 + \lambda_2 v_2}\left[\frac{\lambda_1^2 v_1}{s+\lambda_1} + \frac{\lambda_2^2 v_2}{s+\lambda_2}\right] - \frac{(\lambda_1-\lambda_2)^2}{\lambda_1 v_1 + \lambda_2 v_2}\left[\frac{s}{(s+\lambda_1)(s+\lambda_2)}\right]^2 \cdot \frac{\left[1-f_1^*(s+\lambda_1)\right]\left[1-f_2^*(s+\lambda_2)\right]}{1-f_1^*(s+\lambda_1)f_2^*(s+\lambda_2)} \quad (1)$$

where v_1 and v_2 are the expected sojourn times in phases 1 and 2, respectively; $\lambda_1 = \lambda_a$, $\lambda_2 = \lambda_a + \lambda_f$, and $f_1^*(s)$ and $f_2^*(s)$ are the Laplace transforms of the probability density functions of phases 1 and 2 sojourn times, respectively. The derivation can be found in the Appendix.

Corollary 1: The expected waiting time in system can be estimated as $W_{\text{ev}} = (1)/((1-\sigma)\mu_{\text{ev}})$, where $\sigma = \phi(\mu_{\text{ev}}(1-\sigma))$.

Corollary 1 is obtained based on the renewal approximation and the accuracy of the approximation is discussed in the Appendix. Since the distributions of phase durations are exogenously determined by the threshold n_{thr}, which is endogenized by the strategy chosen by flexible patients, we present the following procedure for obtaining the equilibrium threshold.

First, set $n = \lfloor (R_{\text{ofv}}\mu_{\text{ofv}})/(C_{\text{ofv}}) \rfloor - 1$ and calculate the expected utilities for e-visit and office visit services: $U_{\text{ev}} = R_{\text{ev}} - W_{\text{ev}}(n)C_{\text{ev}}$, and $U_{\text{ofv}} = R_{\text{ofv}} - ((n+1)C_{\text{ofv}})/(\mu_{\text{ofv}})$, respectively. Since the higher the threshold, more traffic flows to the office visit queue, and lesser traffic flows to the e-visit queue, it can be trivially seen that U_{ofv} is monotonically decreasing with n and U_{ev} is monotonically increasing with n, and there exists a unique fixed point across which $U_{\text{ev}} - U_{\text{ofv}}$ changes the sign. Therefore, depending on the difference in magnitude between the two utilities, the threshold can be altered from its starting value until the sign of the difference of utilities changes; otherwise, the procedure is terminated at $n = 0$, which implies that e-visits offer a better utility than that can be obtained from an empty office visit queue. This fixed point n is the equilibrium threshold.

The effective arrival to each queue is determined as

$$A_{\text{ev}}^{\text{thr}} = P(N_{\text{ofv}} \geq n_{\text{thr}})\big(\lambda_a + \lambda_f\big) + P(N_{\text{ofv}} < n_{\text{thr}})\lambda_a \quad (2)$$

$$A_{\text{ofv}}^{\text{thr}} = P(N_{\text{ofv}} \geq n_{\text{thr}})\lambda_d + P(N_{\text{ofv}} < n_{\text{thr}})\big(\lambda_f + \lambda_d\big) \quad (3)$$

where N_{ofv} is the number of patients in the office visit queue (including in service) and n_{thr} is the equilibrium threshold.

Remark: In contrast to the mixed strategy, due to the integral nature of the threshold, in equilibrium, the flexible patient may not experience the exact same waiting time at both queues even if the rewards and waiting costs are same for both the queues.

V. Discussions

In this section, we compare different system configurations and the corresponding system performance.

A. System Performance Comparison

Five parameters, the three arrival rates and the two service rates are critical to modeling the system. For system design purposes, in each experimental setting, we vary the e-visit capacity (characterized by e-visit service rate μ_{ev}) and fix the remaining four parameters and investigate how the aggregate patient surplus (determined by average waiting times) vary accordingly. We consider a mean population arrival rate ranging from 12 to 20 patients per day. This roughly reflects the total traffic for a single primary care provider we have observed in practice. The split of these arrival rates across different types is allowed to vary as they could shift in the future or due to externally imposed constraints. For example, the COVID-19 pandemic could force a lot of patients to use e-visits more often. The service rates are chosen to ensure stability of the system. The reward and cost of each service are nominally chosen to reflect the relative importance of each service.

To lay out a clear picture of the system property driven by these parameters, we consider horizontally substitutable services, that is, $R_{\text{ev}} = R_{\text{ofv}}$ and $C_{\text{ev}} = C_{\text{ofv}}$. This assumption is necessary for the implementation of the duplication strategy as patients must be indifferent to which service they receive as long as they receive the quickest possible service. A discussion regarding unequal rewards and time sensitivity is provided in Section V-C.

Since the formulas to evaluate the waiting times of the duplication strategy model and the threshold strategy model are complicated, we present numerical studies to illustrate the performance of each system. Now, we briefly explain our choice of parameters. Each scenario corresponds to a setting where the proportions of each class of patients are varied, along with different office visit capacity provisions. The stability condition for all strategies requires $\mu_{\text{ev}} + \mu_{\text{ofv}} > \Lambda := \lambda_a + \lambda_f + \lambda_d$, $\mu_{\text{ofv}} > \lambda_d$, and $\mu_{\text{ev}} > \lambda_a$, so μ_{ev} is set to be larger than $\max(\Lambda - \mu_{\text{ofv}}, \lambda_a)$.

Fig. 4 provides the equilibrium waiting time of flexible patients under the mixed strategy. In the left panel, it can be seen that with increasing e-visit capacity, the waiting time is monotonically decreasing. There is a sharp drop when the e-visit capacity is increased from its minimum capacity by a moderate amount. As shown in the right panel, the arrival to the e-visit queue is increasing, until there is no longer any flexible patient joining the office visit queue. To divert patients, the e-visit capacity can take a range of values that will lead to different market sizes of e-visits. Based on the cost of providing e-visits, a decision can be made to determine the optimal e-visit capacity such that the cost of staffing shall not overweight the gain in aggregate patient surplus. Finally, beyond a particular e-visit capacity identified in Proposition 1, the expected utility of e-visits exceeds that of office visits and therefore, all flexible patients will choose the e-visit service. This capacity can capture the entire market for e-eligible patients, that is, flexible patients as well as e-advocates. Provision of more e-visit capacity beyond this particular capacity will only lead to a better waiting time for e-eligible patients as can be seen in Fig. 4 (the dark green dotted lines representing expected office visit waiting times are flat).

Fig. 5 depicts the system performance where the flexible patient has the option to submit duplicate requests at both services. The lowest waiting times are witnessed for the class of flexible patients at the expense of the dedicated office visit patients and e-advocates. Specifically, the flexible patient experiences a waiting time equivalent to that experienced in a system whose service rate is the summation of both service rates and the arrival rate is also a combination of the two. It can be seen that the waiting time of e-advocates is close to that of flexible patients with increasing e-visit capacity and had the office visit capacity been increased instead, the waiting time of dedicated office visit patients would be close to that of flexible patients, owing to lesser congestion. When the e-visit capacity is sufficiently large, effectively, almost all flexible patients are served in the e-visit queue but there will always be a nonzero

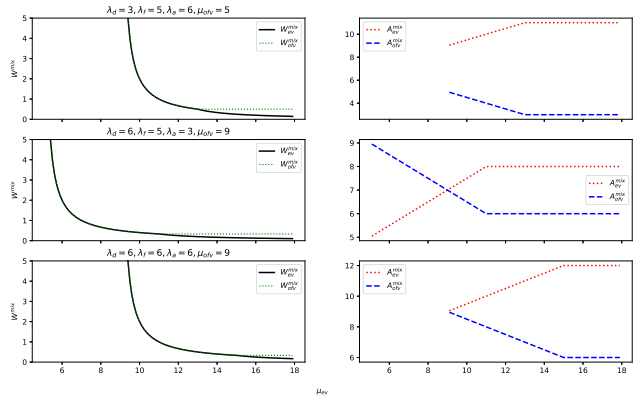

Fig. 4. System performance under the mixed strategy (left to right): 1) the average waiting times for e-visit and office visit queues denoted as $W_{\text{ev}}^{\text{mix}}$ and $W_{\text{ofv}}^{\text{mix}}$, respectively, and 2) the effective arrival rates of e-visit ($A_{\text{ev}}^{\text{mix}}$) and office visit ($A_{\text{ofv}}^{\text{mix}}$) queues for various values of e-visit capacities.

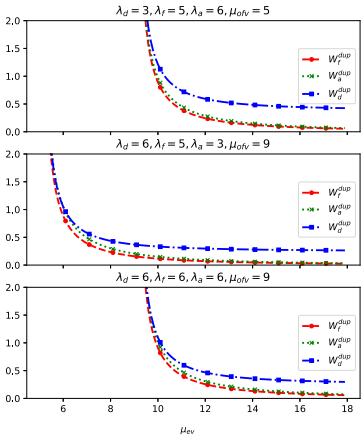

Fig. 5. Average waiting times for dedicated office visits (W_d^{dup}), e-advocates (W_e^{dup}), and flexible patients (W_f^{dup}) under the duplication strategy for various values of e-visit capacities.

probability that a flexible patient will be served by the office visit queue even at high e-visit capacity. Thus, it can be seen that under a duplication strategy, the e-visit server will not capture the entire e-eligible market.

Fig. 6 exhibits the system performance when flexible patients adopt the threshold strategy. First, with the increase in e-visit capacity, we witness a decrease in the threshold beyond which the flexible patient joins the e-visit queue, and such a threshold eventually stabilizes: it becomes zero as there is enough e-visit capacity such that the total time spent in the e-visit queue is smaller than the expected office visit service time. Second, we also observe that the waiting time for the office visit queue is slightly lesser than the e-visit waiting time for lower e-visit capacities. This is because, the flexible patient on arrival observes the conditional waiting time $(n+1)/(\mu_{\text{ofv}})$, which is higher than the expected waiting time for an office visit patient. Overall, the expected waiting time is decreasing with increased e-visit capacity, and the arrival to the e-visit queue is increasing, until there is no longer any flexible patient arrival to the office visit queue. Beyond this e-visit capacity, the behavior of the system under the threshold strategy is the same as that of the mixed strategy. We also see that the waiting times follow a step

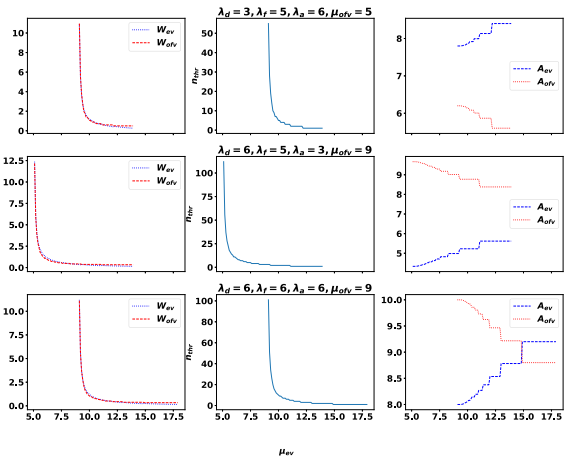

Fig. 6. System performance under the threshold strategy (left to right): 1) the average waiting times for e-visit and office visit queues W^{thr}; 2) the thresholds (n_{thr}) employed by flexible patients in equilibrium; and 3) the effective arrival rates of e-visit ($A_{\text{ev}}^{\text{thr}}$) and office visit ($A_{\text{ofv}}^{\text{thr}}$) queues for various values of e-visit capacities.

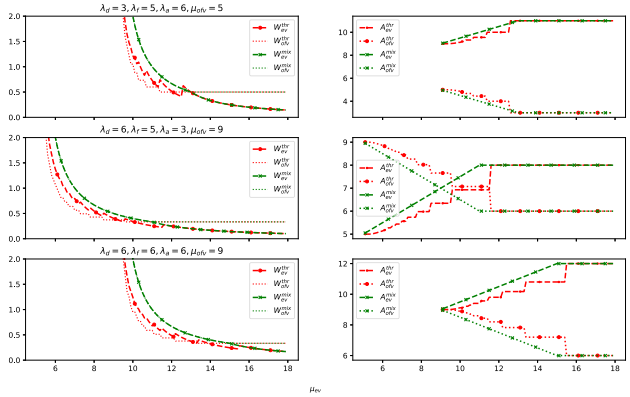

Fig. 7. Comparison of system performance with unobservable and observable office visit queues.

pattern. The expected waiting time at the office visit queue follows the shape of the value of thresholds employed by the flexible patients which is provided in the middle plots. These thresholds also follow a stepwise pattern due to its integral nature. With increasing e-visit capacity, the reduction in office visit waiting time is caused by the decrease in thresholds. Since the flexible patients optimize their strategy, the e-visit waiting time closely follows the office visit waiting time. However, as the earlier remark suggests, the integral nature of the threshold strategy followed does not lead to an exact match in waiting times at both the queues.

To further compare the two system configurations, we plot the measurements in the same figure (Fig. 7). It can be seen in the right panel that, the arrival of patients to the e-visit queue under the mixed strategy is always no less than that under the threshold strategy. This indicates that the information of prospective lower waiting time at the office visit queue can draw opportunists there, and hence, balance the patient flow and lead to an overall lesser waiting time. With more e-visit capacity, however, such information is not necessary, as all flexible patients seek e-visits directly. It is interesting to note that the turning points of the two

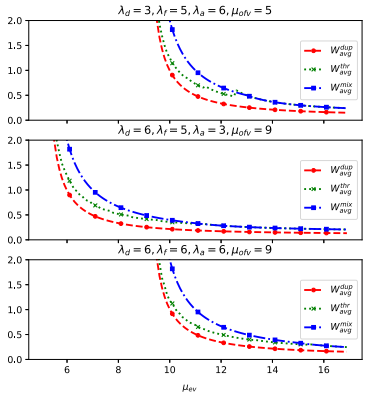

Fig. 8. Aggregate waiting time of the mixed strategy ($W^{\text{mix}}_{\text{avg}}$), the duplication strategy ($W^{\text{dup}}_{\text{avg}}$), and the threshold strategy ($W^{\text{thr}}_{\text{avg}}$) for various values of e-visit capacities.

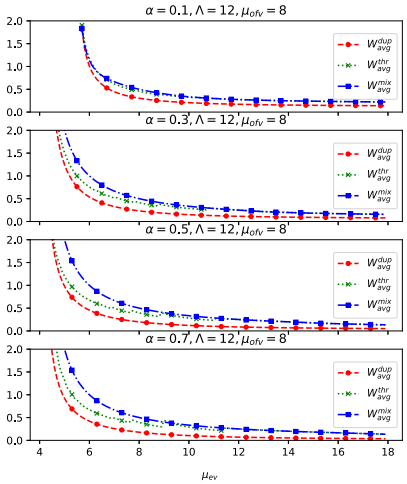

Fig. 9. Comparison of the system performance of the three system configurations with different population mixes.

effective arrival curves are not the same; the e-visit capacity to attract all flexible patients of the threshold strategy is larger than that of the mixed strategy. This is trivial to see as $R_{\text{ev}} - (C_{\text{ev}})/(\mu_{\text{ev}} - (\lambda_a + \lambda_f)) > R_{\text{ofv}} - (C_{\text{ofv}})/(\mu_{\text{ofv}})$ implies $R_{\text{ev}} - (C_{\text{ev}})/(\mu_{\text{ev}} - (\lambda_a + \lambda_f)) > R_{\text{ofv}} - (C_{\text{ofv}})/(\mu_{\text{ofv}} - \lambda_d)$. Under the threshold strategy, flexible patients can join the office visit queue if it is empty or less congested. Such a liberty is not afforded when the queue lengths are unobservable. Next, we look at the left panel of Fig. 7. Although office visit arrivals of the threshold strategy are greater than that of the mixed strategy, the former enjoys lesser waiting times. In the mixed strategy case, the arrival is a Poisson process; with the threshold strategy, the arrival process has a smaller variability than that of Poisson, so even the intensity is higher, it can still yield a smaller queue length.

Lastly, we compare the aggregate patient surplus of different system configurations. Since we assume identical rewards and waiting costs and do not allow patient loss, we only need to compare the expected waiting times for all patients in the system. The average waiting times weighted by their arrival rates are shown in Fig. 8. It is found that the duplication strategy offers the least waiting time for the system compared to other strategies. This is due to the work-conserving nature: the flexible patient gets her service as soon as one server is available and frees up the appointment she holds in the other server, dynamically balancing the workload of the two queues. This policy sets the best efficiency the system can possibly obtain. However, this performance is achieved under the premise that the redundant appointment gets deleted and can be immediately filled with any later request. There are potential difficulties in implementing such a process since it will be burdensome to advance appointments for office visits. Nevertheless, a system that allows redundant requests for flexible patients is most efficient. In this case, any cost of providing queue information to patients is replaced by the cost of jointly managing information between two systems.

The comparison of the mixed strategy and the threshold strategy exemplified the value of information to flexible customers. We summarize that for low e-visit capacity scenarios, the threshold strategy is superior to the mixed strategy in terms of the aggregate patient surplus. This is due to the exploitation of the information regarding the possible low congestion in the office queue by flexible patients who use that to their advantage. However, for maintaining the same market size, the e-visit capacity needed for the threshold strategy case is always no smaller than that of the mixed strategy case. If revenue gains are the same but the unit cost of providing e-visit services is lower than that of office visits, the medical institution would actually be in favor of the mixed strategy. Overall, the impact of information provision diminishes as more e-visit services become available. When the equilibrium threshold becomes zero, the mixed strategy is equivalent to the threshold strategy.

B. Population Composition

We hereby analyze the impact of the population mix of patients. Define α as the proportion of flexible patients in the population. We have observed that due to the COVID pandemic, many patients have altered their attitude and increasingly adopted e-visits. For illustration purposes, we split the inflexible patients equally among dedicated office visits and e-advocates and vary α and observe the waiting times under the three strategies. Fig. 9 elucidates the relationship: the duplication strategy still offers the best performance even when the proportion of flexible patients is low. However, the difference between the waiting times of the threshold strategy and the mixed strategy become more significant as the proportion of flexible patients increases. It articulates that information is helpful when there are more people who can use it. Comparing vertically, that is, keep the same e-visit service rate under different α's, we see in Fig. 10 that in general, the system performs better with a higher proportion of flexible patients. The system benefits from the flexibility of patients, which improves the efficiency of the system. Note that in the bottom figure of the right panel, the waiting time under the threshold strategy when $\alpha = 0.7$ is inferior to $\alpha = 0.5$ till $\mu_{\text{ev}} = 5.3$. This can be attributed to the huge swings in phase arrival rates (high variance) when the proportion of flexible patients is really high and the e-visit capacity is too low. Unlike the mixed strategy and duplication strategy, the monotonicity

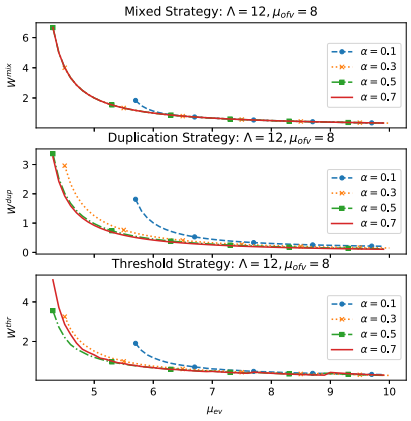

Fig. 10. Impact of change in population mixes on each system.

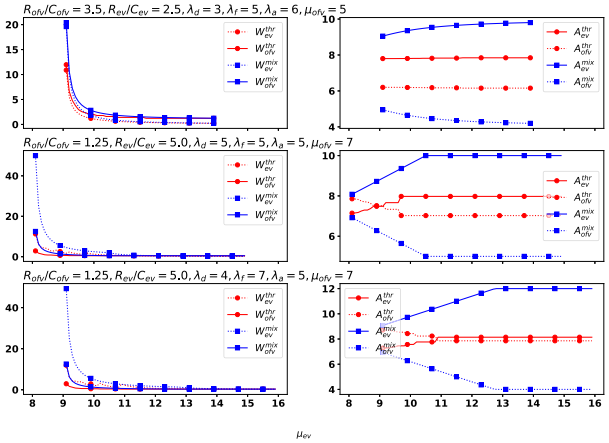

Fig. 11. Comparison of mixed and threshold strategy when there is vertical differentiation among the services.

of waiting times with respect to α is lost for the threshold strategy.

C. Vertical Differentiation

Here, we look at the system performance when flexible patients value the two services differently, that is, $R_{\text{ev}} \neq R_{\text{ofv}}$ and/or $C_{\text{ev}} \neq C_{\text{ofv}}$. If the office visit is valued higher than the e-visit service, the patient is willing to wait longer for the office visit service. Therefore, the equilibrium average waiting times in the two queues differ. Fig. 11 compares the mixed and threshold strategy under vertical differentiation. Three cases are provided with different benefit-to-cost ratios for office and e-visit services. The first plot has the office visit offering a higher benefit, whereas the next two plots have the e-visit offering a superior benefit. The flattened arrival rate curves in the first plot indicate that the change in arrival rates are not very sensitive to the addition of e-visit service capacity when the office visit offers a very high benefit-to-cost ratio over e-visits. In contrast, when the benefit-to-cost ratio of e-visits is high, the flexible patients can be sensitive to the e-visit service capacity and will be attracted to choose e-visit services. It is interesting to observe that for the mixed strategy, nearly the whole flexible segment will be attracted to e-visits as the e-visit service capacity becomes abundant, whereas for the threshold strategy, the diversion is moderate, which helps balance the flow and avoids overcrowding. The flexible patients might be overly optimistic when making decisions based on expectation solely. This conservative behavior induced by allowing for an observable office visit queue is also witnessed in the case when the benefit-to-cost ratio of the office visit is greater. This again justifies the value of information, especially when the e-visit capacity is not large enough to guarantee the expected utility (see Fig. 11), where the expected waiting time of the mixed strategy (blue curves) is greater than that of the threshold strategy (red curves) when μ_{ev} is not large enough.

Essentially, the benefit-to-cost ratio measures the level of flexibility of those flexible patients. Patients are more flexible when $(R_{\text{ev}})/(C_{\text{ev}})$ and $(R_{\text{ofv}})/(C_{\text{ofv}})$ are closer to each other, and less flexible when one significantly dominates the other. We do not consider the duplication strategy under the vertically differentiated service system since flexible patients would not submit duplicate requests for services valued differently.

D. Patient Loss

It can be seen in Fig. 6 that for really low e-visit capacity, the waiting time is extremely high. In traditional office visit appointment systems, it is observed that longer delays lead to more no-shows of patients. This indicates that there is a threshold to the utility sought by flexible patients. We can relax the assumption of a closed system and assume that flexible patients will not seek a service from the medical institution unless the net utility is positive. They may either seek services elsewhere or choose to self-treat. We can therefore impose a condition of net positive utility for flexible patients, which will affect the equilibrium traffic.

If the arriving flexible patient observes n patients in the office visit queue, the expected utility is $R_{\text{ofv}} - ((n+1)C_{\text{ofv}})/(\mu_{\text{ofv}})$. In the absence of e-visits, flexible patients balk from the office visit queue if this utility is negative and the patient is lost from the system. The threshold n_{thr} beyond which the patient utility is negative is determined by $n_{\text{thr}} = \underset{n \in \mathbb{Z}^+}{\text{argmin}} \{R_{\text{ofv}} - ((n+1)C_{\text{ofv}})/(\mu_{\text{ofv}}) < 0\}$.

In the presence of e-visits, if the office visit threshold n_{thr} is exceeded, a flexible patient who balks will adopt a mixed strategy with probability x of requesting the e-visit service as long as their expected utility from e-visits is positive. The utility of requesting an e-visit is $R_{\text{ev}} - C_{\text{ev}} W_{\text{ev}}(x, n_{\text{thr}})$, where $W_{\text{ev}}(x, n_{\text{thr}})$ is the expected waiting time in the system incurred at the e-visit queue. For a fixed e-visit capacity and threshold n_{thr}, $W_{\text{ev}}(x, n_{\text{thr}})$ is increasing in x. The probability x_{thr} is determined as $x_{\text{thr}} = \underset{x \in [0,1]}{\text{argmin}} \{R_{\text{ev}} - W_{\text{ev}}(x, n_{\text{thr}})C_{\text{ev}} \leq 0\}$.

The probability $x_{\text{thr}} = 1$ implies that there is no flexible patient loss for the system. Without patient loss, the positive utility selection and the best utility selection are equivalent. If $x_{\text{thr}} = 1$, it means that the e-visit service has enough capacity to absorb flexible patients. Now, the flexible patient can maximize her utility by balking from the office visit queue at a threshold smaller than n_{thr}. This new threshold $\hat{n}$ can be determined by $\hat{n} = \underset{n \in \mathbb{Z}^+}{\text{argmin}} \{R_{\text{ofv}} - ((n+1)C_{\text{ofv}})/(\mu_{\text{ofv}}) < R_{\text{ev}} - C_{\text{ev}} W_{\text{ev}}(1, n)\}$.

The procedure for positive utility selection is outlined below. First, determine $n = \lfloor (R_{\text{ofv}}\mu_{\text{ofv}})/(C_{\text{ofv}}) \rfloor$ and calculate the

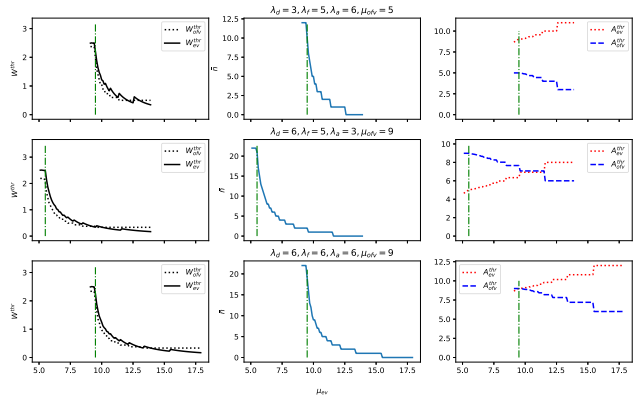

Fig. 12. System performance under the positive utility strategy by flexible patients: (left to right) 1) the average waiting times for e-visit and office visit services (W^{thr}); 2) the equilibrium thresholds (n_{thr}) used by the flexible patients; and 3) the patient diversion due to e-visits for various values of e-visit capacities. In this scenario, there is no patient loss beyond a certain e-visit capacity marked in green dashed vertical lines.

expected utilities for e-visit and office visit, respectively, by assuming $x = 1$. Then, two scenarios can unfold. If the utility of joining the e-visit queue is negative, there should be patient loss. Using a binary search, we determine x (mixed strategy probability) for which the expected e-visit utility is zero. If the utility of joining the e-visit queue is nonnegative, flexible patients can balk from the office visit queue for a shorter threshold. We can adjust the threshold n till the e-visit utility just outweighs that of office visits (similar to the best utility selection procedure).

Fig. 12 depicts the system performance where flexible patients always seek a positive utility from the system. The experimental setting is the same as that of Fig. 6. The positive utility selection results in the initial plateau region in the waiting time (the left panel in Fig. 12), where patient loss is incurred till a positive utility is achieved from increasing the e-visit capacity. With additional capacity, the arrival to e-visit queue ramps up till there is no patient loss from the system as shown in the middle panel of the arrival rate plots. Post the plateau region, the performance of the system is similar to that of the best utility seeking case.

If the medical institution would like to determine the e-visit capacity, it should consider providing at least the capacity that ensures no patient loss. With the additional provision of e-visit capacity, the medical institution might incur costs without receiving additional revenue. If the services are not provided by the same medical institution, the e-visit provider might want to set the capacity that captures the entire e-eligible market which is greater than the capacity that ensures no patient loss. If the cost of having additional e-visit capacity is inelastic to the demand, that market capturing e-visit capacity can be installed by the e-visit provider.

E. Competitive E-Visits and Office Visits

Here, we briefly discuss the comparison of the strategies when e-visits and office visits are not provided by the same entity and therefore compete for patients. If the e-visit provider does not cooperate with the office visit provider on information sharing between the two systems, the duplicate request will not be deleted and will tentatively be wasted. Then, the dedicated office visit patients and e-advocates witness an expected waiting time of $(1)/(\mu_{\text{ofv}} - \lambda_d - \lambda_f)$ and $(1)/(\mu_{\text{ev}} - \lambda_a - \lambda_f)$, respectively, and flexible patients experience an expected waiting time which is between $(1)/(\mu_{\text{ev}} + \mu_{\text{ofv}} - \lambda_a - \lambda_f - \lambda_d)$ and $(1)/(\mu_{\text{ev}} + \mu_{\text{ofv}} - \lambda_a - 2\lambda_f - \lambda_d)$. Therefore, the aggregate patient surplus will be worse off under the duplication strategy comparing to the mixed strategy.

If both providers have complete information, anyone who enjoys a lower cost to provide information will offer queue length information. For instance, if it is not conventional for patients to obtain their expected waiting time for e-visits, it is a dominant strategy for the office visit provider to reveal information because comparing to the mixed strategy, they will attract more patients given more arrivals means more profits for the provider. Notably, in order to provide information for office visit services, clinics typically have receptionists or call center representatives and maintain a scheduling system, such as the appointment system in patient portals. On the other hand, the e-visit provider might be able to provide low-resolution information to their customers at no cost, and an example of the current attempt is to announce to patients that messages will typically be responded within 24–72 h.

VI. Takeaways and Conclusion

Access barriers to primary care services lead to adverse societal consequences. It was estimated that around 14%–27% of emergency room visits could be addressed at primary care facilities, a significant amount of which were attributable to long waiting times for appointments and limited after-hours care at physicians' offices [2]. Fragmented care (e.g., seeking primary care in urgent care settings) and self-diagnosis are prone to error and dangerous if inappropriate decisions are made. The emergence of e-visits fills the gap in care continuum. With a growing impetus on patient centered care, the healthcare system will witness a major transition from traditional care delivery modules to virtual ones. The COVID-19 pandemic has underscored the benefits of e-services as being "contactless" and travel-free, which has significantly stimulated the implementation of e-visits and telehealth services and propelled a bulk of patient population to adopt this novel platform. This might lead to a transformation in patient expectations and continued demand for the service, and with this transition, it is vital to ensure that services are configured to the highest performance standard. We have shown that provision of queue length information has a significant influence on how patients decide their service selection strategy, and we expect that with technological innovation, the cost of information management, and service coordination can be brought down significantly, offering more potential for improving care delivery efficiency and patient experience.

In closing, for the appropriate service design and successful implementation of e-visits, understanding patient needs, addressing physician concerns, and removing operational barriers are instrumental. The overall benefits of e-visits need to be assessed to incorporate e-visits' impact on the entire medical care spending and patient outcomes, which demands multidisciplinary research endeavors.

Future work will involve the exploration of systems with general service distributions to improve the generalizability of the models, and the determination of the optimal staffing, scheduling, and capacity assignment among physicians and nonphysician providers, which could be flexible according to patient flow dynamics. It is also imperative to explore payment structures that accommodate technologically mediated interactions between providers and patients and design the service contract between patients and the medical institution that enables the delivery of the best outcomes for all stakeholders. The contract could feature pricing (copayment) of e-visits, insurers' reimbursement policies, as well as information sharing.

APPENDIX: PROOFS

Proof of Proposition 1: If $\mu_{\text{ofv}} > \lambda_d + \lambda_f$ and $R_{\text{ofv}} - (C_{\text{ofv}})/(\mu_{\text{ofv}} - (\lambda_d + \lambda_f)) \geq R_{\text{ev}} - (C_{\text{ev}})/(\mu_{\text{ev}} - \lambda_a)$, then $p = 1$. And, if $\mu_{\text{ev}} > \lambda_a + \lambda_f$ and $R_{\text{ofv}} - (C_{\text{ofv}})/(\mu_{\text{ofv}} - \lambda_d) \leq R_{\text{ev}} - (C_{\text{ev}})/(\mu_{\text{ev}} - (\lambda_a + \lambda_f))$, then, $p = 0$. To simplify the analysis, let $C_{\text{ofv}} = C_{\text{ev}} = C$ and define $k = (R_{\text{ofv}} - R_{\text{ev}})/(C)$, which is the waiting cost adjusted reward difference.

To find the probability p with which the flexible patients choose the office visit queue at equilibrium, we equate the expressions for utilities at office visit and e-visit queue: $R_{\text{ofv}} - \text{CW}_{\text{ofv}} = R_{\text{ev}} - \text{CW}_{\text{ev}}$, where $W_i = (1)/(\mu_i - A_i), i \in \{\text{ev}, \text{ofv}\}$. We have to ensure that the value of p obtained satisfies $0 \leq p \leq 1$ and that $A_i \leq \mu_i$ for $i \in \{\text{ev}, \text{ofv}\}$. When $p = 1$, we get $k = (R_{\text{ofv}} - R_{\text{ev}})/(C) = k_o := (\mu_{\text{ev}} - \mu_{\text{ofv}} + \lambda_d + \lambda_f - \lambda_a)/((\mu_{\text{ofv}} - (\lambda_d + \lambda_f))(\mu_{\text{ev}} - \lambda_a))$ and when we have $p = 0$, we get the value of $k_e := (\mu_{\text{ev}} - \mu_{\text{ofv}} + \lambda_d - \lambda_f - \lambda_a)/((\mu_{\text{ofv}} - \lambda_d)(\mu_{\text{ev}} - \lambda_a - \lambda_f))$.

It can be concluded that when there is ample office visit capacity, and the waiting cost adjusted reward difference is large enough, that is, $k \geq k_o$, all flexible patients choose office visits. Similarly, when there is enough e-visit capacity, and the waiting cost adjusted reward difference is small enough, that is, $k \leq k_e$, all flexible patients choose e-visits. Otherwise, flexible patients choose a mixed strategy, that is, having probability $p \in (0, 1)$ to choose office visits. The stability condition implies that $p < (\mu_{\text{ofv}} - \lambda_d)/(\lambda_f)$ and $p > (\lambda_a + \lambda_f - \mu_{\text{ev}})/(\lambda_f)$. Then, in equilibrium

$$R_{\text{ofv}} - \frac{C_{\text{ofv}}}{\mu_{\text{ofv}} - \left(\lambda_d + p\lambda_f\right)} = R_{\text{ev}} - \frac{C_{\text{ev}}}{\mu_{\text{ev}} - \left(\lambda_a + (1-p)\lambda_f\right)}. \tag{A.1}$$

We get the solution for p as

$$p = \frac{k\left(\mu_{\text{ev}} - \mu_{\text{ofv}} + \lambda_d + \lambda_f - \lambda_a\right) - 2}{2k\lambda_f} + \frac{\sqrt{4 + k^2\left(\mu_{\text{ofv}} + \mu_{\text{ev}} - \lambda_a - \lambda_d - \lambda_f\right)^2}}{2k\lambda_f}. \tag{A.2}$$

Proof of Proposition 2: There are two servers indexed by ofv and ev with service rates μ_{ofv} and μ_{ev} and three classes of customers indexed by d, f, and a with arrival rates λ_d, λ_a, and λ_f, where ofv serves classes d and f, ev serves classes f and a. Class f is, therefore, the redundant class of customers.

To illustrate the state representation, let us consider a particular state space $(u_{\text{ofv,ev}}, M_{\text{ev}}, t_{\text{ofv}}, M_{\text{ofv}})$. Reading from right to left, it indicates that the office visit server is busy servicing a customer of type d or f followed by t_{ofv} customers of type d, followed by a busy e-visit server servicing a customer of type f or a, and it is followed by $u_{\text{ofv,ev}}$ customers of type f, d, or a. The state-space can also be of the form $(u_{\text{ofv,ev}}, M_{\text{ofv}}, t_{\text{ev}}, M_{\text{ev}})$, where t_{ev} customers are of type a or f and the $u_{\text{ofv,ev}}$ customers belong to any of the three classes. The indices on u and t in the state-space representation (u_{ij}, M_i, t_j, M_j) are dropped since they can be inferred from the ordering of the tuple and can be treated as constants. Then, the detailed balanced equations can be derived using this state-space configuration.

Using Theorem 2 of [15], we can obtain the steady-state probabilities $\Pi()$ as

$$\Pi(u, M_{\text{ofv}}, t, M_{\text{ev}}) = \left(\frac{\lambda_d + \lambda_a + \lambda_f}{\mu_{\text{ofv}} + \mu_{\text{ev}}}\right)^u \left(\frac{\lambda_a}{\mu_{\text{ev}}}\right)^t \cdot \left(\frac{\lambda_d + \lambda_f}{\mu_{\text{ofv}} + \mu_{\text{ev}}}\right)\left(\frac{\left(\lambda_a + \lambda_f\right)\left(\lambda_d + \lambda_a + \lambda_f\right)}{\left(\lambda_d + \lambda_a + 2\lambda_f\right)\mu_{\text{ev}}}\right)\Pi(0) \quad \text{for } u, \quad t \geq 0 \tag{A.3}$$

$$\Pi(u, M_{\text{ev}}, t, M_{\text{ofv}}) = \left(\frac{\lambda_d + \lambda_a + \lambda_f}{\mu_{\text{ofv}} + \mu_{\text{ev}}}\right)^u \left(\frac{\lambda_d}{\mu_{\text{ofv}}}\right)^t \cdot \left(\frac{\lambda_a + \lambda_f}{\mu_{\text{ofv}} + \mu_{\text{ev}}}\right)\left(\frac{\left(\lambda_d + \lambda_f\right)\left(\lambda_d + \lambda_a + \lambda_f\right)}{\left(\lambda_d + \lambda_a + 2\lambda_f\right)\mu_{\text{ofv}}}\right)\Pi(0) \quad \text{for } u, \quad t \geq 0 \tag{A.4}$$

$$\Pi(u, M_{\text{ofv}}) = \left(\frac{\lambda_d}{\mu_{\text{ofv}}}\right)^u \left(\frac{\left(\lambda_d + \lambda_f\right)\left(\lambda_d + \lambda_a + \lambda_f\right)}{\left(\lambda_d + \lambda_a + 2\lambda_f\right)\mu_{\text{ofv}}}\right)\Pi(0) \quad \text{for } u \geq 0 \tag{A.5}$$

$$\Pi(u, M_{\text{ev}}) = \left(\frac{\lambda_a}{\mu_{\text{ev}}}\right)^u \left(\frac{\left(\lambda_a + \lambda_f\right)\left(\lambda_d + \lambda_a + \lambda_f\right)}{\left(\lambda_d + \lambda_a + 2\lambda_f\right)\mu_{\text{ev}}}\right)\Pi(0) \quad \text{for } u \geq 0. \tag{A.6}$$

Here, $\Pi(0)$ represents the empty state.

Summing the above equations over all possible values of u and t and adding $\Pi(0)$ and equating to 1 gives the value of $\Pi(0)$. Using distributional Little's Law and [15, eq. (33)], we obtain the Laplace transformation for waiting time under duplication strategy as

$$\begin{aligned} E\left[e^{-sW_d}\right] &= \Pi(\cdot, M_{\text{ofv}}).\frac{\mu_{\text{ofv}} - \lambda_d}{\mu_{\text{ofv}} - \lambda_d + s} + \Pi(\cdot, M_{\text{ev}}).1 \\ &\quad + \Pi(\cdot, M_{\text{ofv}}, \cdot, M_{\text{ev}}).\frac{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right)}{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_d\right) + s} \\ &\quad + \Pi(\cdot, M_{\text{ev}}, \cdot, M_{\text{ofv}}).\frac{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_d\right)}{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right) + s} \\ &\quad \cdot \frac{\mu_{\text{ofv}} - \lambda_d}{\mu_{\text{ofv}} - \lambda_d + s} \end{aligned} \tag{A.7}$$

$$
\begin{aligned}
&E\left[e^{-sW_f}\right] \\
&= \Pi(\cdot, M_{\text{ev}}).1 + \Pi(\cdot, M_{\text{ofv}}).1 \\
&\quad + \Pi(\cdot, M_{\text{ev}}, \cdot, M_{\text{ofv}}).\frac{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right)}{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right) + s} \\
&\quad + \Pi(\cdot, M_{\text{ofv}}, \cdot, M_{\text{ev}}).\frac{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right)}{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right) + s}
\end{aligned} \tag{A.8}
$$

$$
\begin{aligned}
&E\left[e^{-sW_a}\right] \\
&= \Pi(\cdot, M_{\text{ofv}}).1 + \Pi(\cdot, M_{\text{ev}}).\frac{\mu_{\text{ev}} - \lambda_a}{\mu_{\text{ev}} - \lambda_a + s} \\
&\quad + \Pi(\cdot, M_{\text{ev}}, \cdot, M_{\text{ofv}}).\frac{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right)}{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right) + s} \\
&\quad + \Pi(\cdot, M_{\text{ofv}}, \cdot, M_{\text{ev}}).\frac{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right)}{\mu_{\text{ofv}} + \mu_{\text{ev}} - \left(\lambda_d + \lambda_f + \lambda_a\right) + s} \\
&\quad \cdot \frac{\mu_{\text{ev}} - \lambda_a}{\mu_{\text{ev}} - \lambda_a + s}
\end{aligned} \tag{A.9}
$$

where

$$
\begin{aligned}
\Pi(\cdot, M_{\text{ofv}}) &= \sum_{u=0}^{\infty} \Pi(u, M_{\text{ofv}}) \\
&= \frac{(\lambda_d + \lambda_f)(\lambda_d + \lambda_a + \lambda_f)}{(\lambda_d + \lambda_a + 2\lambda_f)\mu_{\text{ofv}}} \frac{1}{1 - \frac{\lambda_d}{\mu_{\text{ofv}}}} \Pi(0)
\end{aligned} \tag{A.10}
$$

$$
\begin{aligned}
\Pi(\cdot, M_{\text{ev}}) &= \sum_{u=0}^{\infty} \Pi(u, M_{\text{ev}}) \\
&= \frac{(\lambda_a + \lambda_f)(\lambda_d + \lambda_a + \lambda_f)}{(\lambda_d + \lambda_a + 2\lambda_f)\mu_{\text{ev}}} \frac{1}{1 - \frac{\lambda_a}{\mu_{\text{ev}}}} \Pi(0)
\end{aligned} \tag{A.11}
$$

$$
\begin{aligned}
\Pi(\cdot, M_{\text{ev}}, \cdot, M_{\text{ofv}}) &= \sum_{u=0}^{\infty} \sum_{t=0}^{\infty} \Pi(u, M_{\text{ev}}, t, M_{\text{ofv}}) \\
&= \frac{\lambda_a + \lambda_f}{\mu_{\text{ofv}} + \mu_{\text{ev}}} \frac{(\lambda_d + \lambda_f)(\lambda_d + \lambda_a + \lambda_f)}{(\lambda_d + \lambda_a + 2\lambda_f)\mu_{\text{ofv}}} \\
&\quad \cdot \frac{1}{1 - \frac{\lambda_d}{\mu_{\text{ofv}}}} \frac{1}{1 - \frac{\lambda_d + \lambda_a + \lambda_f}{\mu_{\text{ofv}} + \mu_{\text{ev}}}} \Pi(0)
\end{aligned} \tag{A.12}
$$

$$
\begin{aligned}
\Pi(\cdot, M_{\text{ofv}}, \cdot, M_{\text{ev}}) &= \sum_{u=0}^{\infty} \sum_{t=0}^{\infty} \Pi(u, M_{\text{ofv}}, t, M_{\text{ev}}) \\
&= \frac{\lambda_d + \lambda_f}{\mu_{\text{ofv}} + \mu_{\text{ev}}} \frac{(\lambda_a + \lambda_f)(\lambda_d + \lambda_a + \lambda_f)}{(\lambda_d + \lambda_a + 2\lambda_f)\mu_{\text{ev}}} \\
&\quad \cdot \frac{1}{1 - \frac{\lambda_a}{\mu_{\text{ev}}}} \frac{1}{1 - \frac{\lambda_d + \lambda_a + \lambda_f}{\mu_{\text{ofv}} + \mu_{\text{ev}}}} \Pi(0)
\end{aligned} \tag{A.13}
$$

representing the steady-state probabilities of the system being in these states.

Proof of Proposition 3: We first look at the office visit queue in isolation. Let $\lambda = \lambda_d + \lambda_f$ be the total potential demand for office visits and denote $\rho = (\lambda/\mu_{\text{ofv}})$ and $\rho_d = (\lambda_d/\mu_{\text{ofv}})$. Under the stability condition $\rho_d < 1$, the average number of patients and the average time in system are given by $L_{\text{ofv}} = \sum_{i=1}^{\infty} i p_i$, where $p_0 = [(1-\rho^n)/(1-\rho) + (\rho^n)/(1-\rho_d)]^{-1}$, $p_k = \rho^k p_0$, $0 < k \leq n$, and $p_k = \rho^n \rho_d^{k-n} p_0$, $k > n$; and $W_{\text{ofv}} = (L_{\text{ofv}})/(A_{\text{ofv}})$. The effective arrival rate of the office visit queue $A_{\text{ofv}} = (\lambda(1-\rho^n)/(1-\rho) + \lambda_d(\rho^n)/(1-\rho_d))p_0$. In addition, $P(N_{\text{ofv}} < n) = (1-\rho^n)/(1-\rho)p_0$ and $P(N_{\text{ofv}} \geq n) = 1 - P(N_{\text{ofv}} < n)$. We then determine the distribution of each phase that modulates the e-visit queue.

Phase 1 Sojourn Time: Phase 1 indicates the duration of time during which the queue length is less than the threshold n_{thr}. We are interested in the distribution of the duration of Phase 1. Let the random variable T_n be the time till the office visit queue is equal to n_{thr}, when there are n customers in the system at the beginning, where $0 \leq n < n_{\text{thr}}$

$$
\begin{aligned}
T_n &= Y + T_{n+1}, \quad n = 0 \\
T_n &= X + \left\{ \begin{array}{l} T_{n-1}, \text{ with probability } \dfrac{\mu_{\text{ofv}}}{\lambda_d + \lambda_f + \mu_{\text{ofv}}} \\ T_{n+1}, \text{ with probability } \dfrac{\lambda_d + \lambda_f}{\lambda_d + \lambda_f + \mu_{\text{ofv}}} \end{array} \right\} \\
&\quad 0 < n < n_{\text{thr}} \\
T_{n_{\text{thr}}} &= 0.
\end{aligned}
$$

Here, X is an exponential random variable with parameter $\lambda_d + \lambda_f + \mu_{\text{ofv}}$ and Y is an exponential random variable with parameter $\lambda_d + \lambda_f$.

Once the equilibrium is attained, we are interested in the distribution of time duration which starts from the moment the office visit queue length becomes $n_{\text{thr}} - 1$ till the moment it reaches the state n_{thr}. Hence, we need to characterize the distribution for $T_{n_{\text{thr}}-1}$.

Taking the Laplace transform, and letting $\lambda = \lambda_d + \lambda_f$, we get

$$
\bar{T}_0(s) = \frac{\lambda}{\lambda + s} \bar{T}_1(s) \tag{A.14}
$$

$$
\bar{T}_n(s) = \frac{1}{\lambda + \mu_{\text{ofv}} + s} \left[\mu_{\text{ofv}} \bar{T}_{n-1}(s) + \lambda \bar{T}_{n+1}(s) \right]. \tag{A.15}
$$

These recurrence relations can be solved by having $\bar{T}_n(s) = c_1 x_1^n(s) + c_2 x_2^n(s)$, where x_1, x_2 are the roots of the equation

$$
x = \frac{\mu_{\text{ofv}} + \lambda x^2}{\lambda + \mu_{\text{ofv}} + s}. \tag{A.16}
$$

Using the boundary conditions, we have

$$
\bar{T}_{n_{\text{thr}}}(s) = c_1 x_1^{n_{\text{thr}}}(s) + c_2 x_2^{n_{\text{thr}}}(s) = 1 \tag{A.17}
$$

$$
\bar{T}_0(s) = c_1 + c_2 = \frac{\lambda}{\lambda + s}(c_1 x_1(s) + c_2 x_2(s)) = \frac{\lambda}{\lambda + s} \bar{T}_1(s). \tag{A.18}
$$

Solving the equations simultaneously, we get

$$
c_1 = \frac{-(\lambda x_2 - (\lambda + s))}{x_2^{n_{\text{thr}}}(\lambda x_1 - (\lambda + s)) - x_1^{n_{\text{thr}}}(\lambda x_2 - (\lambda + s))} \tag{A.19}
$$

$$
c_2 = \frac{(\lambda x_1 - (\lambda + s))}{x_2^{n_{\text{thr}}}(\lambda x_1 - (\lambda + s)) - x_1^{n_{\text{thr}}}(\lambda x_2 - (\lambda + s))}. \tag{A.20}
$$

The desired phase sojourn time distribution is given by

$$
\begin{aligned}
\bar{T}_{n_{\text{thr}}-1}(s) &= c_1 x_1^{n_{\text{thr}}-1} + c_2 x_2^{n_{\text{thr}}-1} && \text{(A.21)} \\
&= \frac{(\lambda x_1 - (\lambda + s)) x_2^{n_{\text{thr}}-1} - (\lambda x_2 - (\lambda + s)) x_1^{n_{\text{thr}}-1}}{x_2^{n_{\text{thr}}}(\lambda x_1 - (\lambda + s)) - x_1^{n_{\text{thr}}}(\lambda x_2 - (\lambda + s))} \\
&= \lambda \frac{(a-b)(c+b)^{n_{\text{thr}}-1} - (a+b)(c-b)^{n_{\text{thr}}-1}}{(a-b)(c+b)^{n_{\text{thr}}} - (a+b)(c-b)^{n_{\text{thr}}}} \\
&:= f_1^*(s) && \text{(A.22)}
\end{aligned}
$$

where $a = (\mu_{\text{ofv}} - (\lambda + s))/(2)$, $c = (\lambda + \mu_{\text{ofv}} + s)/(2)$ and $b = (((\lambda + \mu_{\text{ofv}} + s)^2 - 4\lambda\mu_{\text{ofv}})^{1/2})/(2)$. The mean time is derived as $v_1 := (\mu_{\text{ofv}}^n - \lambda^n)/((\mu_{\text{ofv}} - \lambda)\lambda^n)$.

Phase 2 Sojourn Time: Phase 2 indicates that the office visit queue length is n_{thr} or greater. Let $n^* = n - n_{\text{thr}} + 1$, where n is the queue length and $n \geq n_{\text{thr}}$. Let the random variable C_{n^*} be the time till $n^* = 0$ which implies the time till the queue length becomes $n_{\text{thr}} - 1$. We are interested in the distribution of the time duration starting from when the queue length is n_{thr} till it becomes $n_{\text{thr}} - 1$. Hence, we seek the distribution for C_1

$$C_{n^*} = X + \begin{cases} C_{n^*-1}, & \text{with probability} \dfrac{\mu_{\text{ofv}}}{\lambda_d + \mu_{\text{ofv}}} \\ C_{n^*+1}, & \text{with probability} \dfrac{\lambda_d}{\lambda_d + \mu_{\text{ofv}}} \end{cases}$$

where X is an exponential random variable with parameter $\lambda_d + \mu_{\text{ofv}}$. Let $\mu_{\text{ofv}} = \mu$. Taking the Laplace transform, we get

$$\bar{C}_{n^*}(s) = \frac{1}{\lambda_d + \mu_{\text{ofv}} + s}\left[\mu_{\text{ofv}}\bar{C}_{n^*-1}(s) + \lambda_d\bar{C}_{n^*+1}(s)\right]. \quad \text{(A.23)}$$

The solution takes the format

$$\bar{C}_{n^*}(s) = c_1 x_1^{n^*}(s) + c_2 x_2^{n^*}(s) \quad \text{(A.24)}$$

where $x_1(s)$ and $x_2(s)$ are the roots of the equation: $(\Lambda + \mu + s)x = \lambda_d x^2 + \mu$ satisfying $0 \leq x_1(s) \leq 1 \leq x_2(s)$. $\bar{C}_{n^*}(s) \leq 1$ implies $c_2 = 0$, and $\bar{C}_0(s) = 1$ implies $c_1 = 1$. The required distribution for the phase sojourn time is

$$\bar{C}_1(s) = \frac{1}{2\lambda_d}\left(\lambda_d + \mu_{\text{ofv}} + s - \sqrt{(\lambda_d + \mu_{\text{ofv}} + s)^2 - 4\lambda_d\mu_{\text{ofv}}}\right). \quad \text{(A.25)}$$

The average phase sojourn time is given by $v_2 := (1)/(\mu_{\text{ofv}} - \lambda_d)$.

Proof of Corollary 1: Here, we present the renewal approximation. The system is approximated by a G/M/1 queue. Let the steady-state probability of e-visit queue being in state n be of the product form $a_n = (1 - \sigma)\sigma^n$, we then try to obtain the root of σ in the equation

$$\sigma = \phi^*(\mu_{\text{ev}}(1 - \sigma)) \quad \text{(A.26)}$$

where $\phi^*(s)$ is defined in (1) in the main text. This can be expanded as

$$\sigma = \left(\frac{\lambda_a^2\left[\left(\lambda_d + \lambda_f\right)^{n_{\text{thr}}-1} + \mu_{\text{ofv}}^{n_{\text{thr}}-1}\right]}{2\left(\lambda_d + \lambda_f\right)^{n_{\text{thr}}}[\mu_{\text{ev}}(1 - \sigma) + \lambda_a]} + \frac{\left(\lambda_a + \lambda_f\right)^2}{(\mu_{\text{ofv}} - \lambda_d)\left[\mu_{\text{ev}}(1 - \sigma) + \lambda_f + \lambda_a\right]}\right) \cdot \frac{1}{\lambda_a\frac{\left(\lambda_d+\lambda_f\right)^{n_{\text{thr}}-1} + \mu_{\text{ofv}}^{n_{\text{thr}}-1}}{2\left(\lambda_d+\lambda_f\right)^{n_{\text{thr}}}} + \frac{\lambda_a+\lambda_f}{\mu_{\text{ofv}}-\lambda_d}} - \frac{\lambda_f^2}{\lambda_a\frac{\left(\lambda_d+\lambda_f\right)^{n_{\text{thr}}-1} + \mu_{\text{ofv}}^{n_{\text{thr}}-1}}{2\left(\lambda_d+\lambda_f\right)^{n_{\text{thr}}}} + \frac{\lambda_a+\lambda_f}{\mu_{\text{ofv}}-\lambda_d}} \cdot \left[\frac{\mu_{\text{ev}}(1 - \sigma)}{(\mu_{\text{ev}}(1 - \sigma) + \lambda_a)\left(\mu_{\text{ev}}(1 - \sigma) + \lambda_a + \lambda_f\right)}\right]^2 \cdot \frac{1 + \left(\bar{c} - \bar{b}\right)\psi - \Lambda\psi - \frac{\bar{c}-\bar{b}}{\Lambda}}{1 - \left(\bar{c} - \bar{b}\right)\psi} \quad \text{(A.27)}$$

where

$$\bar{c} = \frac{\Lambda + \mu_{\text{ofv}} + \mu_{\text{ev}}(1 - \sigma)}{2}$$

$$\bar{b} = \frac{\sqrt{(\Lambda + \mu_{\text{ofv}} + \mu_{\text{ev}}(1 - \sigma))^2 - 4\Lambda\mu_{\text{ofv}}}}{2}$$

$$\psi = \frac{(a - b)(c + b)^{n_{\text{thr}}-1} - (a + b)(c - b)^{n_{\text{thr}}-1}}{(a - b)(c + b)^{n_{\text{thr}}} - (a + b)(c - b)^{n_{\text{thr}}}}$$

$$\Lambda = \lambda_a + \lambda_f + \lambda_d.$$

Then, the waiting time W_{ev} as a function of n_{thr} is given by $W_{\text{ev}} = (1)/((1 - \sigma)\mu_{\text{ev}})$.

REFERENCES

[1] L. Brandenburg, P. Gabow, G. Steele, J. Toussiant, and B. J. Tyson, "Innovation and best practices in health care scheduling," Inst. Med., Washington, DC, USA, Tech. Rep., 2015.

[2] R. M. Weinick, R. M. Burns, and A. Mehrotra, "Many emergency department visits could be managed at urgent care centers and retail clinics," *Health Affairs*, vol. 29, no. 9, pp. 1630–1636, Sep. 2010.

[3] V. Hynes, "The trend toward self-diagnosis," *Can. Med. Assoc. J.*, vol. 185, no. 3, pp. E149–E150, Feb. 2013.

[4] C. Jung, R. Padman, G. Shevchik, and S. Paone, "Who are portal users vs. early E-visit adopters? A preliminary analysis," in *Proc. AMIA Annu. Symp.*, 2011, p. 1070.

[5] SanfordHealth. (2020). *At Home Virtual Care | Sanford Health*. [Online]. Available: https://www.sanfordhealth.org

[6] *Telehealth Adoption Tracker*, Chartis Group, Portsmouth, U.K., 2021.

[7] P. Olson, "Telemedicine, once a hard sell, can't keep up with demand," *Wall Street J.*, Apr. 2020. [Online]. Available: https://www.wsj.com/articles/telemedicine-once-a-hard-sell-cant-keep-up-with-demand-11585734425

[8] S. Romanick-Schmiedl and G. Raghu, "Telemedicine—Maintaining quality during times of transition," *Nature Rev. Disease Primers*, vol. 6, no. 1, pp. 1–2, 2020.

[9] A. Villareal. (2020). *Telemedicine: The Good, the Bad, the Pleasantly Surprising*. [Online]. Available: https://www.aafp.org

[10] N. Liu and T. DÁunno, "The productivity and cost-efficiency of models for involving nurse practitioners in primary care: A perspective from queueing analysis," *Health Services Res.*, vol. 47, no. 2, pp. 594–613, Apr. 2012.

[11] H. Bavafa, S. Savin, and C. Terwiesch. (Jun. 11, 2021). *Customizing Primary Care Delivery Using E-Visits*. [Online]. Available: https://ssrn.com/abstract=2363685

[12] B. Rajan, T. Tezcan, and A. Seidmann, "Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care," *Manage. Sci.*, vol. 65, no. 3, pp. 1236–1267, Mar. 2019.

[13] X. Zhong, J. Li, P. A. Bain, and A. J. Musa, "Electronic visits in primary care: Modeling, analysis, and scheduling policies," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 3, pp. 1451–1466, Jul. 2017.

[14] X. Zhong, "A queueing approach for appointment capacity planning in primary care clinics with electronic visits," *IISE Trans.*, vol. 50, no. 11, pp. 970–988, Nov. 2018.

[15] J. Visschers, I. Adan, and G. Weiss, "A product form solution to a system with multi-type jobs and multi-type servers," *Queueing Syst.*, vol. 70, no. 3, pp. 269–298, Mar. 2012.

[16] K. Gardner, S. Zbarsky, S. Doroudi, M. Harchol-Balter, and E. Hyytia, "Reducing latency via redundant requests: Exact analysis," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 43, no. 1, pp. 347–360, 2015.

[17] P. Naor, "The regulation of queue size by levying tolls," *Econometrica*, vol. 37, no. 1, pp. 15–24, Jan. 1969.

[18] R. Hassin and M. Haviv, *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, vol. 59. Berlin, Germany: Springer, 2003.

[19] S. Stidham, Jr., *Optimal Design of Queueing Systems*. Boca Raton, FL, USA: CRC Press, 2009.

[20] R. Hassin, *Rational Queueing*. Boca Raton, FL, USA: CRC Press, 2016.

[21] D. A. Andritsos and C. S. Tang, "The impact of cross-border patient movement on the delivery of healthcare services," *Int. J. Prod. Econ.*, vol. 145, no. 2, pp. 702–712, Oct. 2013.

[22] S. Gavirneni and V. Kulkarni, "Concierge medicine: Applying rational economics to health care queuing," *Cornell Hospitality Quart.*, vol. 55, no. 3, pp. 314–325, Aug. 2014.

[23] S. Sharma, Y. Xu, M. K. Gupta, and C. Courcoubetis. (Oct. 1, 2019). *Non-Urgent Visits and Emergency Department Congestion: Patients' Choice and Incentive Mechanisms*. [Online]. Available: https://ssrn.com/abstract=3480940

[24] A. Economou and V. Kulkarni, "Editorial introduction to the special issue on 'strategic queueing: Game-theoretic models in queueing theory'—Part 1," *Queueing Syst.*, vol. 96, pp. 201–203, 2020, doi: 10.1007/s11134-020-09680-w.

[25] A. Economou and V. Kulkarni, "Editorial introduction to the special issue on 'strategic queueing: Game-theoretic models in queueing theory'—Part 2," *Queueing Syst.*, vol. 97, pp. 221–222, 2021, doi: 10.1007/s11134-021-09699-7.

[26] M. Haviv and L. Ravner, "A survey of queueing systems with strategic timing of arrivals," 2020, *arXiv:2006.12053*. [Online]. Available: http://arxiv.org/abs/2006.12053

[27] R. Palmer, N. J. Fulop, and M. Utley, "A systematic literature review of operational research methods for modelling patient flow and outcomes within community healthcare and other settings," *Health Syst.*, vol. 7, no. 1, pp. 29–50, 2017.

[28] W. Whitt. (2016). *Queues With Time-Varying Arrival Rates: A Bibliography*. [Online]. Available: http://www.columbia.edu/ww2040/TV_bibliography_091016.pdf

[29] J. F. C. Kingman, "On doubly stochastic Poisson processes," *Proc. Math. Cambridge Phil. Soc.*, vol. 60, pp. 923–930, Oct. 1964.

[30] A. Lawrance and P. Lewis, "Some models for stationary series of univariate events," in *Stochastic Point Processes: Statistical Analysis, Theory, and Applications*. London, U.K.: Wiley, 1972, pp. 199–256.

[31] V. N. Bhat, "Renewal approximations of the doubly stochastic Poisson processes," *Microelectron. Rel.*, vol. 33, no. 13, pp. 1991–1996, Oct. 1993.

[32] L. V. Green, J. Soares, J. F. Giglio, and R. A. Green, "Using queueing theory to increase the effectiveness of emergency department provider staffing," *Acad. Emergency Med.*, vol. 13, no. 1, pp. 61–68, Jan. 2006.

[33] L. V. Green and S. Savin, "Reducing delays for medical appointments: A queueing approach," *Oper. Res.*, vol. 56, no. 6, pp. 1526–1538, 2008.

[34] *Trends in Telehealth*, NTT DATA, Tokyo, Japan, 2014.

[35] *The Promise of Telehealth for Hospitals, Health Systems and Their Communities*, Amer. Hospital Assoc., Chicago, IL, USA, 2015.

[36] N. Liu, S. R. Finkelstein, and L. Poghosyan, "A new model for nurse practitioner utilization in primary care: Increased efficiency and implications," *Health Care Manage. Rev.*, vol. 39, no. 1, pp. 10–20, 2014.

[37] L. V. Green and N. Liu, "A study of New York City obstetrics units demonstrates the potential for reducing hospital inpatient capacity," *Med. Care Res. Rev.*, vol. 72, no. 2, pp. 168–186, Apr. 2015.

Aditya Mahadev Prakash (Member, IEEE) received the bachelor's degree from IIT Kharagpur, Kharagpur, India, in 2013, the Post-Graduate Diploma degree in management from the Indian Institute of Management, Bengaluru, India, in 2015, and the Ph.D. degree from the Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA, in 2020.

His research interests include the analysis, modeling, design and optimization of healthcare systems and data analytics.

Xiang Zhong (Member, IEEE) received the B.S. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2011, and the M.S. degree in statistics and the Ph.D. degree in industrial engineering from the University of Wisconsin-Madison, Madison, WI, USA, in 2014 and 2016, respectively.

She is currently an Assistant Professor with the Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA. Her research interests include stochastic modeling and control, and data analytics with the application in healthcare, service, and production systems.

Dr. Zhong is a member of IISE and INFORMS.