Queueing Systems (2022) 100:505–507
https://doi.org/10.1007/s11134-022-09827-x

Validating state-dependent queues in health care

René Bekker[1]

Received: 6 February 2022 / Accepted: 28 February 2022 / Published online: 11 May 2022

1 Introduction

Demographic projections and lack of medical staff pose an increasing challenge for the sustainability of our health system. Queueing theory may play a primary role in supporting decision making regarding its future organization, as it provides fundamental insight into the balance between demand of care and available capacity. To achieve this, insight in patient flows and systems dynamics is crucial, where proposed queueing models need to be supported by data. This has been recognized in [1], posing various research opportunities at the interface between queueing theory, service engineering, and statistics. However, the focus there is only on Emergency Departments (ED) and Internal Wards (IW) of hospitals. We need a wider view on our complete health system to propose reforms, including ambulatory care, assisted living, and (nursing) home care.

A fundamental property of queues in health is state dependency, arising in arrival processes, service times, and/or capacities. Practical examples of the former are patients diverted to other ED's or wards due to congestion, and patients choosing between health organizations depending on projected access times. Examples of state-dependent services are discharge policies of the 'best patient' in case of congestion, and patients generating additional demand during waiting. Moreover, organizations adapt available capacity based on queue lengths to maintain acceptable access times. Also, during the Covid-19 pandemic, we have experienced modified nurse-to-patient ratio's to keep the health system accessible. Roughly, we may distinguish (semi-)acute (ED, IW, ICU) and non-acute (GP, nursing homes, home care) situations. Models for the former mainly consider short-term dynamics (requiring time-varying elements), whereas the latter tend to focus on long-term design problems. However, all systems seem to be exposed to state dependency.

We present some open challenges for state-dependent queueing models in health. In line with [1], we call for a data-based queueing-science perspective to find appropriate models that explain the complex dynamics of patient flows in health systems. That is,

✉ René Bekker
r.bekker@vu.nl

[1] Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

can we validate queueing models and estimate parameters? And can relatively simple models capture the essential queueing dynamics?

2 Problem statement

As an example, consider an *appointment-based system*. Let X_n denote the virtual queue length at the start of period n, where a period typically represents a single day. Define $A_n^{(x)}$ and $C_n^{(x)}$ as the number of arrivals and potential number of discharges (or appointments) in period n if $X_n = x$. Then, we have the following Lindley-type of recursion

$$X_{n+1} = \max\left(0, X_n + A_n^{(X_n)} - C_n^{(X_n)}\right). \tag{1}$$

The recursion above may also apply to operating theatres, diagnostic facilities (MRI, CT, X-ray, and laboratory), or the midnight count of patients at the IW or ICU (due to the two time-scale phenomenon in hospital care). Note that a special case of the recursion (1) has been analyzed in [2], where the limiting distribution is determined using a Wiener–Hopf factorization. Assume for the moment that A_n is i.i.d. and $C_n \equiv C$. Then, the discrete-time process $(X_n)_{n\geq 0}$ is similar to the waiting time recursion in the D/G/1 queue. Due to Kingman's approximation, we have $\lim_{n\to\infty} \mathbb{E}X_n \approx \mathbb{V}\mathrm{ar}(A)/2(C-\mathbb{E}A)$ iff $\mathbb{E}A < C$. It is unlikely that such state-*in*dependent assumptions hold in practice. Specifically, access times are often in the order of weeks or months (or even years), which can only be achieved in case the utilization $\mathbb{E}A/C$ is close to, but below, 1, requiring an extremely delicate balance.

A possible (partly) explanation for access times in the order of weeks or months are abandonments [6]. They may be incorporated in (1) by taking $A_n^{(x)} = \sum_{i=1}^{A_n} \mathbb{1}\{G_{i,n} > x\}$, with $G_{i,n}$ the patience time of the ith arrival in period n. Note that in [6], the appointment system has been modeled as a GI/D/1 queue (instead of D/G/1) with balking. A key question is whether abandonments can explain a rather stable but long access time. In practice, we also observe capacity adjustments ($C_n^{(x)}$) to counter excessive congestion.

For *inpatient systems*, such as hospitals and nursing homes, it is more natural to model these as multi-server queues. To address the issues above, we may assume an arrival rate $\lambda(x)$, a service rate $\mu(x)$, and $c(x)$ servers when the backlog equals x.

2.1 Data problems

Health-care data are not widely available. Publicly available data,[1] such as in [1] for inpatient care, would greatly facilitate data-driven queueing analysis.

The available data are also not always complete. For instance, only the admission epoch and not the arrival epoch is typically registered. Moreover, data concerning queue lengths or access times are only occasionally observed and stored. Also, statistical challenges arise as the state of the system changes during a (state-dependent) service time. Such estimates become even more involved due to time dependency, e.g., early discharging on Fridays. Let T_n and D_n denote the epochs at which customer n is entering and leaving service, and let $Y_n = \min(X_n, C_n)$ be the number of customers in service.

[1] This is often involved, or not possible, due to confidentiality and proprietary.

2.2 Queueing-science problems

Given T_n, D_n and Y_n, and X_n for a subset of customers,

– Estimate the distributional properties of $A_n^{(x)}$ and $C_n^{(x)}$ (or $\lambda(x)$, $\mu(x)$ and $c(x)$)
– Validate which model most likely explains the observed behavior of Y_n and X_n.

In most cases, customers may choose between different health organizations offering a comparable service (see Sect. 1). From a modeling perspective, it is convenient to approximate the complete system by a single queue. The central idea in [5] is to focus on a single queue (the single-queue approximation, SQA) and let a state-dependent arrival rate $\lambda(x)$ capture the interaction with the other queues. In contrast with [5], customers typically have different preference profiles for each of the health organizations (e.g., due to medical specialists or geographical location) next to short waiting times. Hence, $\lambda(x)$ should be able to capture the impact of the (non-work-conserving) allocation policy.

2.3 Queueing problems

For customers with different preference profiles for each of the multi-server queues,

– Determine an (asymptotically) optimal preference-based allocation policy
– Propose an SQA for a class of allocation policies and provide theoretical support.

3 Discussion

Modeling health systems is challenging due to the huge influence of human actions on system dynamics. State-dependent queues naturally arise from such human actions (e.g., [4]). The ultimate challenge is to provide sufficiently simple models capturing the essential elements, as this may provide the building blocks in modeling the complete health system. The application domain of call centers may serve as inspiration. For instance, from [3], it follows that the simple Erlang-A model is remarkably useful, despite the fact that service and patience times are not really exponential. A crucial difference with call centers, however, are the alternative choices for health providers.

One possible and well-established option for tractability is to focus on asymptotics. For instance, in [6], appointment systems are analyzed using diffusion approximations for GI/G/1 queues with balking, leading to Ornstein–Uhlenbeck processes in the limit. The mean reverting behavior of these processes is a noticeable phenomenon in practice.

References

1. Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y., Yom-Tov, G.: On patient flow in hospitals: a data-based queueing-science perspective. Stoch. Syst. **5**(1), 146–194 (2015)
2. Boxma, O., Lotov, V.: On a class of one-dimensional random walks. Markov Process. Relat. Fields **2**, 349–362 (1996)
3. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L.: Statistical analysis of a telephone call center: a queueing-science perspective. J. Am. Stat. Assoc. **100**(469), 36–50 (2005)
4. Delasay, M., Ingolfsson, A., Kolfal, B., Schultz, K.: Load effect on service times. Eur. J. Oper. Res. **279**(3), 673–686 (2019)
5. Gupta, V., Harchol-Balter, M., Sigman, K., Whitt, W.: Analysis of join-the-shortest-queue routing for web server farms. Perform. Eval. **64**(9–12), 1062–1081 (2007)
6. Zacharias, C., Armony, M.: Joint panel sizing and appointment scheduling in outpatient care. Manag. Sci. **63**(11), 3978–3997 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.