Queueing Syst
https://doi.org/10.1007/s11134-018-9571-4

Queuing network models for panel sizing in oncology

Peter T. Vanberkel[1] · **Nelly Litvak**[2,3] · **Martin L. Puterman**[4] · **Scott Tyldesley**[5]

Received: 2 July 2015 / Revised: 28 April 2017

Abstract Motivated by practices and issues at the British Columbia Cancer Agency (BCCA), we develop queuing network models to determine the appropriate number of patients to be managed by a single physician. This is often referred to as a physician's panel size. The key features that distinguish our study of oncology practices from other panel size models are high patient turnover rates, multiple patient and appointment types, and follow-up care. The paper develops stationary and non-stationary queuing network models corresponding to stabilized and developing practices, respectively. These models are used to determine new patient arrival rates that ensure practices operate within certain performance thresholds. Data from the BCCA are used to calibrate and illustrate the implications of these models.

Keywords Queueing networks · Panel sizing · Oncology · Capacity planning

Mathematics Subject Classification 60 · 65 · 90

✉ Peter T. Vanberkel
peter.vanberkel@dal.ca

1 Department of Industrial Engineering, Dalhousie University, Halifax, Canada

2 Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

3 Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

4 Sauder School of Business, University of British Columbia, Vancouver, Canada

5 British Columbia Cancer Agency, Vancouver, Canada

Published online: 30 January 2018

1 Introduction

As a result of treatment advances, some cancers are now regarded as chronic diseases [5]. This means that after an initial diagnosis, a cancer patient may require treatment and follow-up throughout his or her lifetime. Incorporating routine follow-up appointments and patient relapses into capacity planning is thus essential for ensuring sufficient capacity to meet demand in a timely fashion. This paper describes a case study, in which we develop queueing models for capacity planning in an oncology practice. While the research reported here is motivated by challenges faced by oncologists at the British Columbia Cancer Agency (BCCA) in Vancouver, British Columbia, Canada, our models and results apply broadly to chronic disease management.

We now describe typical patient management practices at the BCCA. Upon receiving a cancer diagnosis, a patient consults an oncologist who determines a course of treatment. Treatments typically consist of surgery, chemotherapy, radiotherapy, hormone therapy or a combination thereof. During treatment, a patient has recurring appointments with an oncologist to review treatment progress, side effects, complications, and/or to evaluate psychosocial impact. The appointment frequency varies by patient and treatment, but typically there are between one and twelve appointments per year. Upon completion of treatment, the patient is discharged and then monitored by their general practitioner to identify if a relapse occurs. If a patient relapses, he/she is referred back to the original oncologist for further treatment. Thereafter, when the cancer is no longer active, the patient is again discharged. Patients can continue to be referred, treated and discharged for as long as they live.

If an oncologist sees too many new patients now, his or her practice will become overwhelmed in the future due to follow-up appointments and relapses. Managing this is difficult, as the treatment time, relapse rate and arrival rate of new patients are all variable. Given this, and increasing demand, the number of new patients a single oncologist can accommodate is worthy of investigation. In this paper, we model the demand for oncology appointments over a long-term planning horizon accounting for both new and relapsed patient demand. The objective is to determine the number of patients a physician should manage. This number is commonly referred to as a physician's "panel size" [8].

Our models compute the expected panel size and appointment demand as a function of the new patient arrival rate. We distinguish cases where the panel size is stationary, representing an established practice, and non-stationary, representing a new practice. In both cases, we allow multiple patient types and use queueing networks to model the panel size. In the stationary case, we explore the relationship between the arrival rate and the expected waiting time for the first follow-up appointment. In the non-stationary case, where the expected waiting time is negligible, we explore the relationship between the arrival rate and the probability of needing overtime to meet the demand for appointments.

The paper is organized as follows. Section 2 reviews panel sizing literature and positions this paper. In Sect. 3, we represent the panel size as a total number of customers in a queueing network. In Sects. 4 and 5, two queueing networks are introduced to model the panel size random variables with non-stationary and stationary arrival rates,

respectively. Section 6 provides numerical results for the case study, which exemplify the utility of the models. Section 7 concludes the paper.

2 Panel sizing literature

Panel size models traditionally compute the probability that daily demand for appointments is greater than the number of available slots [7,8]. Usually these models ignore serial patterns in appointment recurrence.

The clinical implications of mis-specifying the panel size are discussed by Murray and Berwick [12] and Murray et al. [13]. A web-based application designed for physician use is available (panelsizer.com). These studies consider a primary care practice and model individual physicians separately. The introduction of advanced access (also referred to as open-access, or walk-in scheduling) in primary care practices is often used as a motive to study panel sizes and is reviewed by Balasubramanian et al. [2] and Murray and Berwick [12].

The impact of patient no-shows on the ideal panel size is investigated using both theoretical queueing models and simulation by Green and Savin [6]. They show that in the presence of no-shows and subsequent rescheduling, smaller panel sizes are required, and that results for typical primary care practices with advanced access can be reliably approximated with queueing models. Multiple patient types are considered in [1,2,14] to reflect the demand patterns associated with different patient demographics. The authors use simulation to determine the number of appointments expected in a given week.

Studies [7,8] treat panel size as a constant and do not allow a backlog in demand to occur. The overflow probabilities are computed under the assumption that all demand for a given day would be filled on that day. A later paper [6] relaxes this assumption and allows a finite backlog to develop. The authors represent the appointment system as a single server queue to track backlog and measure the performance of the practice.

The main contribution of our work is the analysis of the trade-off between new patient arrival rates and appointment demand in the presence of an intermittent pattern of patient follow-ups and patient relapses. Furthermore, our analysis incorporates typical characteristics of oncology settings. This extends the existing panel size literature, which is focused solely on general practices.

An essential distinction of our models from the majority of the panel size literature is that we consider the new patient arrival rate as the input parameter. In our experience, this parameter is more meaningful to oncologists than panel size. In fact, we observed that it is more common for oncologists to know their demand for new patient appointments than their total practice size.

3 Queueing model for the panel size

A patient followed by an oncologist can be in the (1) newly referred patient state, (2) active patient state or (3) inactive patient state. Newly referred patients immediately enter the new patient state and have an initial appointment (consultation) with the oncologist. After this initial appointment, the patient enters the active patient state

and has recurring appointments with the oncologist. Patients remain in the active state between appointments and only leave the state when they no longer have cause to see the oncologist. Appointments typically occur simultaneously with treatments, such as chemotherapy, radiotherapy or a combination thereof. We do not model these treatments, only oncologist appointments which typically continue until the treatment is complete and the patient is discharged. We have observed that discharge practices vary widely across oncologists [15] but ignore that fact here for simplicity. At this point, the patient enters the inactive patient state. If an inactive patient's cancer relapses, or they develop cancer-related issues, then the patient is referred back to the oncologist and returns to the active patient state. Finally, a patient exits the system from either the active or inactive states when they die, move or are followed by a different oncologist. For simplicity, we aggregate these categories.

We use queuing networks to model patients of one oncologist as they transition between these states. Each state of a patient is represented by a multiple-server queue. The number of servers in each queue represents the maximal number of patients in that state. The time spent in service in each queue represents the time a patient stays in each state. For example, an oncologist who can manage 50 active patients that stay active for 6 months (on average) is modeled with a queue with 50 servers and an average service time of 6 months. From the number of patients in each state and the given appointment frequency, the number of required appointments is approximated.

All new arrivals immediately enter the new patient state and await their initial appointment. After this appointment, they become active patients of type i with probability p_i. Different patient types $i = 1, 2, \ldots, c$ are grouped on the basis of their follow-up pattern, for example, monthly or bimonthly, reflecting their clinical diagnosis or treatment. After treatment active patients of type i transition to the inactive state with probability $p_{i,\mathrm{ip}}$, and with probability $1 - p_{i,\mathrm{ip}}$, exit the system. Inactive patients can relapse and become active patients of type i with probability p_i'. With probability $1 - \sum_{i=1}^{c} p_i'$, an inactive patient never returns to the active patient state and exits the system.

We assume that the times spent in the active and inactive states, which will correspond to service times in the queue, are independent realizations of D_{np}, $D_{\mathrm{ap},i}$ and D_{ip}, respectively, for new patients (np), active patients (ap) of type $i = 1, \ldots, c$, and inactive patients (ip). The independence assumption is justified, since the decision to discharge a patient depends on the patient's health (or follow-up protocol) and not the system state or previous states visited by the patient. We denote by $X_{\mathrm{np}}(t)$, $X_{\mathrm{ap},i}(t)$ and $X_{\mathrm{ip}}(t)$, respectively, the number of new patients, active patients of type $i = 1, 2, \ldots, c$, and inactive patients at time $t \in [0, \infty)$. The panel size is the sum of the number of patients in each queue, written as

$$\mathrm{PS}(t) = X_{\mathrm{np}}(t) + \sum_{i=1}^{c} X_{\mathrm{ap},i}(t) + X_{\mathrm{ip}}(t). \tag{1}$$

We will omit the argument t for the stationary version of these random variables.

In Sect. 4, we introduce the queueing network for modeling patient states in a new practice with non-stationary arrivals. This is followed by the panel size and per-

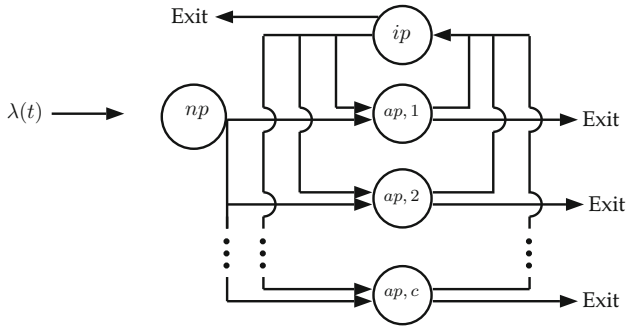

Fig. 1 Non-stationary queueing network model schematic. Each circle represents an infinite server queue

formance metric calculations. In Sect. 5, we do the same, except for an established practice with stationary arrivals.

4 A new practice with non-stationary arrivals

4.1 Queueing model

For the new practice, we model arrivals of new patients as a non-homogeneous Poisson process and each patient state is represented by an infinite server queue. Figure 1 provides a schematic view of the queueing model.

Since the patient population is large, and since patients' cancers develop independently of each other, it is natural to assume that new patients arrive according to a Poisson process. Deviations may occur due to system inefficiencies and calendar effects, which we ignore. A non-homogeneous Poisson process allows the arrival rate to change as the clinic matures. Let $\lambda(t)$ be the instantaneous arrival rate of new patients at time t.

The inactive patient queue is naturally assumed to have an infinite number of servers because these patients do not consume resources. Our assumption that the new and active patient queues have an infinite number of servers is appropriate because new practices are typically underutilized and can accommodate demand without delay. It also represents the practice of oncologists (such as co-author ST) who adjust their schedules (for example, use some non-clinical time or overtime) in order to see all patients within the required time period. For active patients, without loss of generality, we use separate queues for different patient types $1, 2, \ldots, c$; see Fig. 1.

Let $J = \{\text{np}, (\text{ap}, 1), (\text{ap}, 2), \ldots, (\text{ap}, c), \text{ip}\}$ be the set of infinite server queues. The transition probabilities $r'_{k,j}$ are given by

$$r'_{k,j} = \begin{cases} p_i & \text{if } k = \text{np},\ j = (\text{ap}, i), i = 1, 2, \ldots, c, \\ p_{i,\text{ip}} & \text{if } k = (\text{ap}, i),\ j = \text{ip}, i = 1, 2, \ldots, c, \\ p'_i & \text{if } k = \text{ip},\ j = (\text{ap}, i), i = 1, 2, \ldots, c, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

4.2 Panel size

In a network of infinite server queues with non-stationary Poisson arrivals, the number of customers in queue j at time t has a Poisson distribution with mean given by (see, for example, [4,11])

$$\mathbb{E}[X_j(t)] = \mathbb{E}\left[\lambda_j^{'+}\left(t - D_j^e\right)\right]\mathbb{E}[D_j], \quad j \in J, \tag{3}$$

where D_j^e is the excess (or residual) service time and $\lambda_j^{'+}(t)$ is the aggregate arrival rate to queue j, which is the minimum nonnegative solution to the system of equations

$$\lambda_j^{'+}(t) = \lambda(t)\mathbf{1}\{j = \text{np}\} + \sum_{k \in J} \mathbb{E}\left[\lambda_k^{'+}(t - D_k)\right] r'_{k,j}, \quad j \in J,$$

where $\mathbf{1}\{A\}$ is the indicator of A. After solving (3), the expected panel size follows from (1).

4.3 Appointment demand and clinic performance

To determine the number of appointments, we consider discrete periods of length δ indexed by $n = 1, 2, \ldots$. In our case study, we will take δ to be a month, but it can also be a week or two weeks, as required by the application.

We assume that each new patient receives exactly one consultation before exiting the new patient state. To reflect the urgency of new patient referrals, we assume that all who arrive during δ are served during δ. Denote by N_n the number of *new* patient appointments in period n, then

$$\mathbb{E}[N_n] = \int_{n\delta}^{(n+1)\delta} \lambda(t)\mathrm{d}t. \tag{4}$$

We will next compute the required number of *follow-up* appointments during period n. Recall that the patient of type i has an appointment scheduled every i periods. Without further specifying the scheduling mechanism, we assume that the probability that a type i patient has an appointment in period n is $1/i$. This is an idealized assumption, which ignores the non-stationarity. However, from practice we expect that a clinic will attempt to divide appointments evenly between time periods; thus, our assumption represents reality reasonably well. Now, let the random variable F_n denote the number of follow-up appointments in period n. Then we have

$$\mathbb{E}[F_n] = \frac{1}{\delta}\sum_{i=1}^{c}\int_{n\delta}^{(n+1)\delta} \frac{1}{i}\,\mathbb{E}[X_{\text{ap},i}(t)]\mathrm{d}t. \tag{5}$$

Following other authors [7,8], we use overtime frequency as a performance measure. Assume that the capacity of an oncologist is divided into slots. Denote by G_n the total demand for appointment slots in period n. Then

$$\mathbb{E}[G_n] = a\mathbb{E}[N_n] + b\mathbb{E}[F_n], \tag{6}$$

where a and b are the average number of slots required, respectively, for new patient and follow-up appointments. Recall that the number of patients in each queue is Poisson distributed, and $X_{\text{ap},i}(t)$, $i = 1, 2, \ldots, c$, are independent by the definition of the transition matrix (2). It follows that F_n and N_n are also Poisson distributed; see (4) and (5) for their means.

Denote by g the maximal number of appointment slots during regular working hours in a planning period of length δ. The probability of overtime is $P(G_n > g)$, which can now be easily computed under different assumptions, for example, G_n may be a sum of two compound Poisson random variables or have a Poisson distribution with parameter (6).

5 An established practice with stationary arrivals

5.1 Queueing model

In the stationary setting, $\text{PS}(t)$, $X_{\text{np}}(t)$, $X_{\text{ap},i}(t)$, $X_{\text{ip}}(t)$, N_n, F_n and G_n are time invariant, and denoted by PS, X_{np}, $X_{\text{ap},i}$, X_{ip}, N, F and G, respectively. Further, $\lambda(t) = \lambda\ \forall t$ and $\Lambda = \lambda\delta$ is the mean number of arrivals during a period of length δ. N, F, G denote the number of appointments and slots during a period of length δ. The arrival process does not need to be Poisson for the methods described in this section.

Unlike the non-stationary setting, here we model the system using three queues (np, ap, ip) with multiple patient types; see Fig. 2. The number of servers in the ap-queue is denoted by m_{ap} and represents the number of active patients that an oncologist can manage. This may not be a parameter that an oncologist can readily provide, and therefore, we will show how it can be determined by the model for a given arrival rate and a desired mean waiting time; these parameters are usually known in practice.

The set of patients states is again J, and the transition probabilities $r_{k,j}$ are defined as follows:

$$r_{k,j} = \begin{cases} p_i & \text{if } k = \text{np},\ j = (\text{ap}, i),\ i = 1, 2, \ldots, c, \\ p_{i,\text{ip}} & \text{if } k = (\text{ap}, i),\ j = \text{ip},\ i = 1, 2, \ldots, c, \\ p'_i & \text{if } k = \text{ip},\ j = (\text{ap}, i),\ i = 1, 2, \ldots, c, \\ 0 & \text{otherwise.} \end{cases}$$

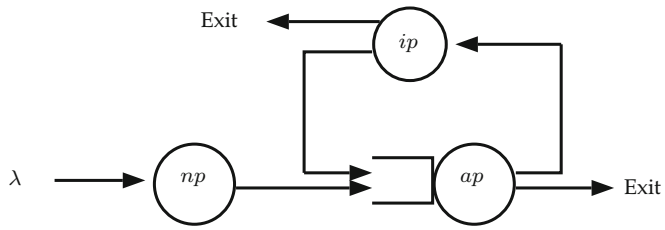

Fig. 2 Representation of the stationary queueing network model. Queues np and ip have an infinite number of servers and queue ap has m_{ap} servers

5.2 Panel size

Each new patient receives one new patient appointment before exiting the new patient state. We again assume that all new patients who arrive during δ are served during δ. Hence, in a period δ,

$$\mathbb{E}[X_{\text{np}}] = \mathbb{E}[N] = \Lambda. \tag{7}$$

For the expected number of patients in the inactive patient queue, we use Little's Law to obtain

$$\mathbb{E}[X_{\text{ip}}] = \mathbb{E}[D_{\text{ip}}] \sum_{i=1}^{c} \Lambda_{\text{ip},i}, \tag{8}$$

where $\Lambda_{\text{ip},i}$ is the aggregate mean arrival rate (the number of arrivals per time period of length δ) to queue ip. This and the aggregate mean arrival rates to each queue for each patient type are determined from the balance flow equations (see, for example, [3]):

$$\begin{aligned}
\Lambda_{\text{np},i} &= \Lambda p_i, \\
\Lambda_{\text{ap},i} &= \Lambda_{\text{np},i} + \sum_{k=1}^{c} \Lambda_{\text{ip},k} r_{\text{ip},(\text{ap},i)}, \\
\Lambda_{\text{ip},i} &= \Lambda_{\text{ap},i} r_{(\text{ap},i),\text{ip}}.
\end{aligned}$$

The expected waiting time in the active patient queue ($\mathbb{E}[W]$) is the time that a patient has to wait until the oncologist has capacity for an additional active patient. This delay is akin to waiting for a first follow-up appointment. Note that $W = 0$ in the non-stationary case due to the infinite server capacity assumption.

Subsequent follow-up appointments are assumed to happen every i periods without delay. Our case study data indicate that $> 96\%$ of follow-up patients have at least two months between appointments, while a two-month waiting time for the first follow-up appointment is considered to be long. Hence, the great majority of existing patients maintain a higher priority even when the waiting time for the first follow-up appointment is two months. Reference [10] considers a related problem of evaluating the impact of scheduling rules and add-on policies on waiting times for first appointments when patients are triaged and assigned urgency levels prior to their first appointment.

Using Little's Law, the expected number of patients of type i in the active patient queue is given by

$$\mathbb{E}[X_{\text{ap},i}] = \Lambda_{\text{ap},i}(\mathbb{E}[D_{\text{ap},i}] + \mathbb{E}[W]). \tag{9}$$

Now the expected panel size can be computed using (1).

5.3 Waiting time in the active patient queue

For the number of customers in a queueing system as in Sect. 5.1, no exact results are known in the literature. Therefore, we use an approximate decomposition method

for multi-class open queueing networks [3]. For this approximation, it is sufficient to assume that the number of arrivals in different time periods, as well as service times, are independent and identically distributed with known first and second moments. This is convenient because these data are typically available from hospital information systems.

We employ Sakasegawa's [16] $G/G/m_{\text{ap}}$ two-moments approximation, which has been shown to perform well [9,17,18], particularly in high load situations [19] as in our case. For our system, the approximation is given by the formula

$$\mathbb{E}[W] \approx \frac{\text{SCV}_{\text{ap},a} + \text{SCV}_{\text{ap},s}}{2} \frac{\rho_{\text{ap}}^{\left(\sqrt{2(m_{\text{ap}}+1)}-1\right)}}{m_{\text{ap}}(1-\rho_{\text{ap}})} \mathbb{E}\left[D_{\text{ap}}^{+}\right], \tag{10}$$

where $\text{SCV}_{\text{ap},a}$ and $\text{SCV}_{\text{ap},s}$ are, respectively, the squared coefficient of variation of the interarrival time and service time distribution functions, respectively, at queue ap. $\mathbb{E}[D_{\text{ap}}^{+}]$ is the mean aggregate service time in queue ap. In the Appendix, we provide the details of how the approximation is computed.

For stability, we assume $\sum_{i=1}^{c} \Lambda_{\text{ap},i} \mathbb{E}[D_{\text{ap},i}] < m_{\text{ap}}$ and it follows that

$$m_{\text{ap}} = \frac{\sum_{i=1}^{c} \Lambda_{\text{ap},i}\, \mathbb{E}[D_{\text{ap},i}]}{\rho_{\text{ap}}}, \tag{11}$$

where $\rho_{\text{ap}} \in (0, 1)$ is the utilization of m_{ap}.

If an oncologist knows his/her capacity for active patients (m_{ap}), then $\mathbb{E}[W]$ can be obtained directly from (10) and the panel size from the stationary version of (1), by summing up (7)–(9). In cases in which oncologists do not know m_{ap}, but know their new patient arrival rate Λ, the waiting time can be directly used as input to compute the number of active patients in (9). This will be described in the next section.

5.4 Appointment demand and clinic performance

The total number of appointments expected in the stationary setting follows from the stationary version of (6), where $\mathbb{E}[N]$ is defined in (7). As in the non-stationary case, an active patient of type i has an appointment in time period n with probability $1/i$. It follows that

$$\mathbb{E}[F] = \sum_{i=1}^{c} \frac{\Lambda_{\text{ap},i}\, \mathbb{E}[D_{\text{ap},i}]}{i}.$$

Now an algorithmic approach can be used to find the appropriate arrival rate, panel size and m_{ap} as follows: For an initial Λ and desired waiting time, we compute $\mathbb{E}[G]$, m_{ap} and the panel size. If $\mathbb{E}[G]$ is higher than desired by the oncologist (g), then Λ must be reduced. Following this approach, the new patient arrival rate (and panel size) can be computed, which leads to the desired appointment frequency and waiting time.

6 Application

In this section, we use numerical examples to illustrate how our models can be used to address relevant practice design issues. Data are derived from the practice of a specific oncologist (co-author ST). He is a clinical scientist with a clinical practice at the BCCA in Vancouver, Canada.

We start with the analysis of an established practice in Sect. 6.1 and continue with a new practice in Sect. 6.2. The time discretization for both settings is one month ($\delta = 1$). To allow for a comparison of the scenarios, the patient mix parameters (p_i, p'_i, $p_{i,\mathrm{ip}}$) and the time spent by patients in each state (D_{np}, $D_{\mathrm{ap},i}$ and D_{ip}) are the same in both scenarios.

Ten years of data extracted from the BCCA information management system were used for this analysis. The data contained approximately 45,000 prostate cancer patient appointments. Prostate cancer patients tend to have long survivorship and therefore require more long-term follow-up than patients with other cancers. It follows that results for practices that treat different cancer modalities and different case mixes may see significantly different return rates, survival rates, etc. For example, lung, head and neck cancers would result in a significantly different patient treatment patterns than seen here because of their survival rates.

Summary statistics from the dataset indicate seven different patient types: types $i = 1, 2, 3, 4, 5, 6$ and 12. Patients with a time between appointments of approximately 7, 8, 9, 10 and 11 months were rare. The fraction of patients of each type was, respectively: 0.033, 0.098, 0.082, 0.164, 0.279, 0.197 and 0.148 for new patients and 0.125, 0.000, 0.063, 0.063, 0.188, 0.250 and 0.313 for relapsed patients (i.e., those transitioning from the inactive to the active state). The time spent as active patients was on average 11.5, 17.2, 19.9, 27.0, 47.1, 56.1 and 31.5 months for each patient type, with standard deviation 11.1, 31.4, 17.1, 12.5, 29.4, 31.1 and 4.8. Approximately 77% of inactive patients return again as active patients; the mean time in the inactive state was approximately 20 months, with a standard deviation of 6 months. Not all returning patients may have experienced an actual cancer relapse, other reasons for follow-up include adjuvant treatments or side effect management.

6.1 Results for an established practice with stationary arrivals

Over the period of study, established oncologist ST saw on average 8.6 new patients and 275.1 follow-up patients per month. The average waiting time for the first follow-up appointment had a target of one month, which was typically achieved.

With $\Lambda = 8.6$ and $\mathbb{E}[W] = 1$ as input for the stationary model, the panel size, partitioned by patient type i and state, is shown in Table 1. The results show that in an average month oncologist ST will have 2017.5 patients in his panel, the majority (71%) will be active patients who require follow-up appointments. Of the 1430.1 patients in the active state, 37.6 are waiting for their first appointments. This follows since the number of patients receiving follow-up appointments is $\sum_{i=1}^{c} \Lambda_{\mathrm{ap},i} \mathbb{E}[D_{\mathrm{ap},i}] = 1392.6$, leaving $1430.1 - 1392.6 = 37.6$ patients waiting. We find 578.7 inactive patients on average in his practice.

Table 1 Expected panel size for an established practice with stationary arrivals, $\Lambda = 8.6$ and $\mathbb{E}[W] = 1$

i	1	2	3	4	5	6	12	Total
New patients								8.6
Active patients	48.8	15.3	52.6	90.2	376.7	510.4	335.8	1430.1
Inactive patients	60.1	13	38.7	49.5	120.5	137.6	159	578.7
Total								2017.5

Table 2 Active patient slots needed (m_{ap}) for a given Λ and $\mathbb{E}[W]$

Λ	$\mathbb{E}[W] = 1$	$\mathbb{E}[W] = 2$	$\mathbb{E}[W] = 3$	$\mathbb{E}[W] = 6$
6	990.2	983.9	980.9	977.1
7	1152.8	1146.1	1143.1	1139.1
8	1315.2	1308.4	1305.2	1301.1
9	1477.7	1470.5	1467.3	1463.1
10	1640.0	1632.7	1629.3	1625.1

Table 3 Appointment demand for an established practice with stationary arrivals, $\Lambda = 8.6$ and $\mathbb{E}[W] = 1$

i	1	2	3	4	5	6	12	Totals
New patients								8.6
Active patients	44.9	7.2	16.7	21.7	73.7	83.5	27.1	275.1

Our analysis shows that to achieve a target waiting time of $\mathbb{E}[W] = 1$ with an average monthly arrival rate of new patients of $\Lambda = 8.6$, $m_{\text{ap}} = 1392.6$ active patients are required. The sensitivity of m_{ap} to changes in Λ and $\mathbb{E}[W]$ is displayed in Table 2. We see that m_{ap} is very sensitive to the arrival rate of new patients and relatively insensitive to the mean waiting time. Increasing m_{ap} by a small amount can greatly reduce average waiting times. The table also demonstrates that small increases in Λ require significant increases in m_{ap}.

Table 3 shows the appointment demand, partitioned by patient type i, corresponding to $\Lambda = 8.6$ and $\mathbb{E}[W] = 1$. The results show that in an average month oncologist ST will see 275.1 follow-up patients. The 8.6 new patients require on average $a = 3$ appointment slots and the follow-up patients require $b = 1$ appointment slots meaning the average monthly demand for appointment slots is $\mathbb{E}[G] = 301$. This agrees closely with the observed offering of 294.8 appointment slots in an average month. The small error is to be expected due to a certain degree of variability in observed data.

6.2 Results for a new practice with non-stationary arrivals

In this section, we consider a new practice with an arbitrarily large supply of new patients and compare the results to the stationary setting. The performance indicator here is the probability of overtime $\mathbb{P}(G_n > g)$, where G_n is determined as in Sect. 4.

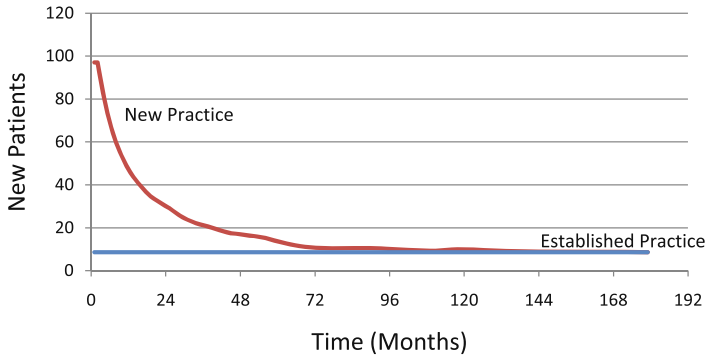

Fig. 3 Capacity to accommodate new patients

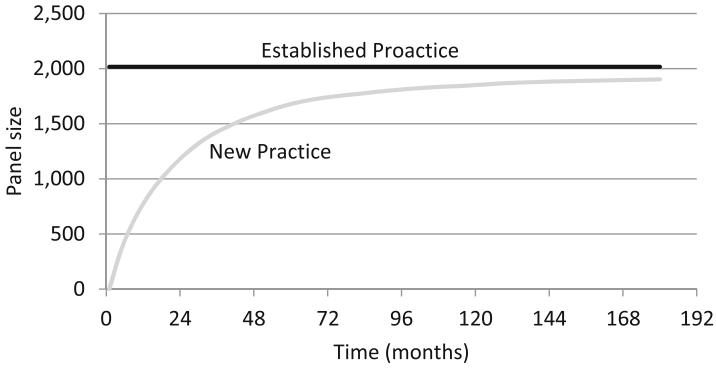

Fig. 4 Evolution of the oncologist's panel size

We use this to compute the maximum arrival rate $\lambda(t)$ of new patients such that $\mathbb{P}(G_n > g) \leq \varphi$ for a specified threshold φ. We take G_n to be Poisson distributed with expectation (6), $\varphi = 0.2$ and set $g = 301$ as in the previous section.

Figure 3 shows the mean arrival rates for patients. We see that a significant number of new patients can be accommodated by the new oncologist in the first few months of practice. However, that amount decreases rapidly during the first 4 years, after which the oncologist can accommodate only slightly more than in the stationary setting. The panel size corresponding to $\lambda(t)$ also approaches what is expected from the stationary model, as illustrated in Fig. 4. Figure 5 shows that in the first year, the majority of the oncologist's time is spent seeing new patients; however, this changes rapidly.

We conclude that in the early years of a new practice a significant number of new patients are needed in order to sufficiently utilize the oncologist's time. Yet care must be taken to eventually limit the number of new patients; otherwise, the practice will be overpopulated and the demand for follow-up appointments will become overwhelming.

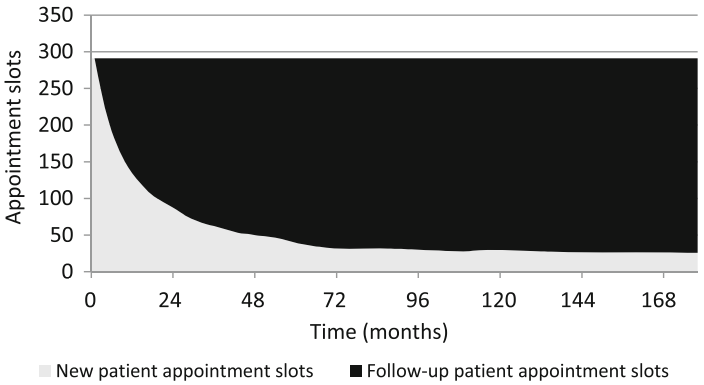

Fig. 5 Evolution and breakdown of appointment slots in a new practice

7 Discussion

This paper uses stationary and non-stationary queuing networks to extend panel size models to settings where there are multiple patient and appointment types. In it, the panel size varies over time to accommodate patient arrivals, turnover and recurrence. The methodology illustrated the trade-offs between demand, patient waiting times and the new patient arrival rate. Understanding this balance is crucial for achieving a well-functioning practice and provides insight into how capacity changes affect performance.

The main challenge in a case study such as presented in this paper is to tailor the model to the specific application. Moreover, the model's input parameters have to be chosen in such a way that their value is available from the data. For example, in our case, the panel size is usually not known, but the arrival rate of new patients is typically available.

Besides oncology, the methods described in this paper can be used in other settings, particularly for physicians treating patients with chronic diseases. Future queueing models could account for transitions between patient types as they become healthier (or sicker), appointment frequency depending on acuity (or availability) and/or sequencing oncologist appointments with treatments. When the queueing model is tailored to the specific setting, the approaches of this paper can be used to support capacity planning, to determine operational confines of a practice and to improve the start-up period for new physicians.

Acknowledgements The authors would like to acknowledge staff from the CIHR Team in Operations Research for Improved Cancer Care at the BC Cancer Agency for their support of this project and their assistance with data collection and Daniel Ding for his valuable feedback on our manuscript.

Appendix: Aggregation for the multi-class open queueing network

Here we describe the technique used to aggregate the parameters in the multi-class open queueing network in Sect. 5. See [3] for more details on the method.

Let $J' = \{\text{np}, \text{ap}, \text{ip}\}$ be the set of queues in the network and let $j \in J'$. We start by computing the first and second moments of the aggregate service time (D_{ap}^{+}) for patients in the active patient queue. This amounts to a weighted average of the service time of each patient type as follows:

$$\mathbb{E}\left[D_{\text{ap}}^{+}\right] = \frac{1}{\sum_{i=1}^{c} \Lambda_{\text{ap},i}} \sum_{i=1}^{c} \Lambda_{\text{ap},i} \mathbb{E}[D_{\text{ap},i}],$$

$$V\left[\left(D_{\text{ap}}^{+}\right)\right] = \frac{1}{\sum_{i=1}^{c} \Lambda_{\text{ap},i}} \sum_{i=1}^{c} \Lambda_{\text{ap},i} V[(D_{\text{ap},i})],$$

where $V[X]$ is the variance of random variable X and $\Lambda_{j,i}$ is the arrival rate of patient type i to queue $j \in J'$.

From these aggregate values, the squared coefficient of variance for the service time in the active patient queue ($\text{SCV}_{\text{ap,s}}$, note that the subscript ap indicates the active patient queue and the subscript s indicates service) can be obtained as follows:

$$\text{SCV}_{\text{ap,s}} = \frac{1}{\Lambda_{\text{ap}}^{+}\left(\mathbb{E}[D_{\text{ap}}^{+}]\right)^2} \times \sum_{i=1}^{c} \left(\Lambda_{\text{ap},i} \mathbb{E}[D_{\text{ap},i}]^2 \left(\left(\frac{\sqrt{V[D_{\text{ap},i}]}}{(\mathbb{E}[D_{\text{ap},i}])} \right)^2 + 1 \right) - 1 \right). \tag{12}$$

The aggregate mean arrival rate and the aggregate routing probabilities are, respectively, $\Lambda_j^{+} = \sum_{i=1}^{c} \Lambda_{j,i}$, $j \in J'$, and $r_{\text{ap,np}}^{+} = 1$, $r_{\text{ap,ip}}^{+} = 1/\Lambda_{\text{ap}}^{+} \sum_{i=1}^{c} \Lambda_{\text{ap},i} r_{(\text{ap},i),\text{ip}}$, $r_{\text{ip,ap}}^{+} = 1/\Lambda_{\text{ip}}^{+} \sum_{i=1}^{c} \sum_{k=1}^{c} \Lambda_{\text{ip},i} r_{\text{ip},(\text{ap},k)}$. Since all new patients go to queue np, the external (new) patient arrival rate $\Lambda_0^{+} = \Lambda$ and $r_{0,j}^{+} = \mathbf{1}\{j = \text{np}\}$.

At this point, the c patient classes are aggregated into a single class and we now consider the network to be a single class open queueing network with the aggregate parameters described above. To analyze the single class open queueing network, we next determine the SCV for the interarrival times to each queue ($\text{SCV}_{j,a}$) as follows:

$$\begin{aligned} \text{SCV}_{\text{np},a} &= \alpha_{\text{np}}, \\ \text{SCV}_{\text{ap},a} &= \alpha_{\text{ap}} + \text{SCV}_{\text{np},a} \beta_{\text{np,ap}} + \text{SCV}_{\text{ip},a} \beta_{\text{ip,ap}}, \\ \text{SCV}_{\text{ip},a} &= \alpha_{\text{ip}} + \text{SCV}_{\text{ap},a} \beta_{\text{ap,ip}}, \end{aligned} \tag{13}$$

where α_j and $\beta_{i,j}$ are constants depending on the input data:

$$\alpha_j = 1 + w_j \left(I_j \left(\frac{\sqrt{V[\Lambda]}}{\Lambda} \right)^2 - 1 \right) + w_j \left(\sum_{k \in J'} \frac{\Lambda_k^{+} r_{k,j}^{+}}{\Lambda_j^{+}} \left(\left(1 - r_{k,j}^{+}\right) + r_{k,j}^{+} \rho_k^2 x_k \right) \right),$$

$$\beta_{k,j} = w_j r_{k,j}^{+} \frac{\Lambda_k^{+} r_{k,j}^{+}}{\Lambda_j^{+}} \left(1 - \rho_k^2\right), \quad k \in J',$$

where $I_{\text{np}} = 1$, $I_{\text{ap}} = I_{\text{ip}} = 0$ and

$$
\begin{aligned}
w_j &= \left(1 + 4(1-\rho_j)^2(v_j - 1)\right)^{-1}, \\
v_j &= \left(\sum_{k \in J' \cup \{0\}} \left(\frac{\Lambda_k^+ r_{k,j}^+}{\Lambda_j^+}\right)^2\right)^{-1}, \\
\rho_j &= \frac{\Lambda_j^+ \mathbb{E}[D_j^+]}{m_j}, \\
x_j &= 1 + m_j^{-0.5}(\max[\text{SCV}_{j,s}, 0.2] - 1), \quad j \in J'.
\end{aligned}
$$

Since m_{np} and m_{ip} are large, then $x_{\text{np}} \approx 1$, $x_{\text{ip}} \approx 1$, and $\rho_{\text{np}} < 1$, $\rho_{\text{ip}} < 1$.

Using the SCV for the arrival process (13) and the SCV for the service time (12), the expected waiting time in the active patient queue is approximated as follows (see [16]):

$$
\mathbb{E}[W] \approx \frac{\text{SCV}_{\text{ap},a} + \text{SCV}_{\text{ap},s}}{2} \frac{\rho_{\text{ap}}^{\left(\sqrt{2(m_{\text{ap}}+1)}-1\right)}}{m_{\text{ap}}(1-\rho_{\text{ap}})} \mathbb{E}\left[D_{\text{ap}}^+\right].
$$

References

1. Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., Stahl, J.: Improving clinical access and continuity through physician panel redesign. J. Gen. Int. Med. **25**(10), 1109–1115 (2010)
2. Balasubramanian, H., Denton, B., Lin, M.: Handbook of Healthcare Delivery Systems: Managing Physician Panels in Primary Care, Chap. 10, pp. 1–23. CRC Press, Taylor and Francis, New York (2010)
3. Bitran, G.R., Morabito, R.: Open queueing networks: optimization and performance evaluation models for discrete manufacturing systems. Prod. Oper. Manag. **5**(2), 163–193 (1996)
4. Boucherie, R.J., Taylor, P.: Transient product from distributions in queueing networks. Discrete Event Dyn. Syst. **3**(4), 375–396 (1993)
5. Cella, D.F., Tulsky, D.S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., Silberman, M., Yellen, S.B., Winicour, P., Brannon, J.: The functional assessment of cancer therapy scale: development and validation of the general measure. J. Clin. Oncol. **11**(3), 570–579 (1993)
6. Green, L.V., Savin, S.: Reducing delays for medical appointments: a queueing approach. Oper. Res. **56**(6), 1526–1538 (2008)
7. Green, L.V., Savin, S., Murray, M.: Providing timely access to care: what is the right patient panel size? Jt. Comm. J. Qual. Patient Saf. **33**(4), 211–218 (2007)
8. Hall, R.W.: Patient Flow: Reducing Delay in Healthcare Delivery, 1st edn. Springer, New York (2006)
9. Kimura, T.: Approximations for multi-server queues: system interpolations. Queueing Syst. **17**(3–4), 347–382 (1994)
10. Ma, C., Puterman, M., Saure, A., Taylor, M., Tyldesley, S.: Capacity planning and appointment scheduling for new patient oncology consults. Health Care Manag. Sci. **19**, 1–15 (2015)
11. Massey, W.A., Whitt, W.: Networks of infinite-server queues with nonstationary Poisson input. Queueing Syst. **13**(1), 183–250 (1993)
12. Murray, M., Berwick, D.M.: Advanced access: reducing waiting and delays in primary care. J. Am. Med. Assoc. **289**(8), 1035–1040 (2003)
13. Murray, M., Davies, M., Boushon, B.: Panel size: how many patients can one doctor manage? Fam. Pract. Manag. **14**(4), 44 (2007)

14. Ozen, A., Balasubramanian, H.: The impact of case mix on timely access to appointments in a primary care group practice. Health Care Manag. Sci. **16**(2), 101–118 (2013)
15. Poznanski, D.: Prostate cancer follow-up practices of genitourinary radiation oncologists at the BC cancer agency Vancouver centre. Master's thesis, University of British Columbia (2010)
16. Sakasegawa, H.: An approximation formula $l_q \approx \alpha\rho^{\beta}/(1-\rho)$. Ann. Inst. Stat. Math. **29**(1), 67–75 (1977)
17. Whitt, W.: Approximations for the GI/G/m queue. Prod. Oper. Manag. **2**(2), 114–161 (1993)
18. Wu, K., McGinnis, L.: Performance evaluation for general queueing networks in manufacturing systems: characterizing the trade-off between queue time and utilization. Eur. J. Oper. Res. **221**(2), 328–339 (2012)
19. van Eeden, K., Moeke, D., Bekker, R.: Care on demand in nursing homes: a queueing theoretic approach. Health Care Manag. Sci. **19**(3), 227–240 (2016)