Health Syst
DOI 10.1057/s41306-017-0032-9

ORIGINAL ARTICLE

Waiting time-based staff capacity and shift planning at blood collection sites

S. P. J. van Brummelen[1,2,3] · N. M. van Dijk[1,3] · K. van den Hurk[2] · W. L. de Kort[2]

Received: 5 January 2017 / Revised: 23 May 2017 / Accepted: 25 May 2017

Abstract Sanquin, the organization responsible for blood collection in the Netherlands, aims to be donor-friendly. An important part of the perception of donor-friendliness is the experience of waiting times. At the same time, Sanquin needs to control the costs for blood collection. A significant step to shorten waiting times is to align walk-in arrivals, and staff capacity and shifts. We suggest a two-step procedure. First, we investigate two methods from queuing theory to compute the minimum number of staff members required for every half hour. Next, these minimum numbers of staff members will be used to determine optimal lengths and starting times of shifts with an Integer Linear Program. Finally, the practical implications of the method are shown with numerical results. These results show that the presented approach can bring significant savings while at the same time guaranteeing a waiting time-based service level for blood donors.

Introduction

Background

The Dutch blood bank, Sanquin, is responsible for the collection and distribution of blood in the Netherlands. Sanquin is a non-profit organization that has a legal monopoly on both these tasks. Blood and plasma are collected at approximately 50 fixed locations and around 100 sites that are visited by a Mobile Blood Collection Center. Combinedly, these sites collect around 450,000 whole blood donations and around 300,000 plasma donations every year. Although plasma donors make an appointment for their donation, whole blood donors can walk in without an appointment. Although this results in random arrivals to the collection site, the arrival process is not as random as one might think. Clear patterns show up in the arrival times of donors, and these are mostly independent of day and location. Peaks in arrival intensity clearly show up early in the morning, around lunch time, and around dinner time.

Donations take place on a voluntary, non-remunerated basis. Next to a limited health check and small gifts or tokens for recognition, there are no incentives to donate blood or plasma. Therefore, the structure of the blood donation system in the Netherlands stresses the need to treat donors well and avoid any discomfort like unnecessary waiting times. However, for financial reasons, it is preferred not to deploy extra staff members to reduce waiting times. For the same reason, Sanquin currently focuses her collection sites and intake sessions on production. For every session and every hour worked by a staff member, the required number of donations has been set in advance, and staff members are scheduled based on these requirements. Waiting times are not consistently taken into account in these scheduling methods. Some managers of collection sites schedule an extra staff member during peak hours to counteract extreme waiting or sojourn times, but most staff members are scheduled for an entire day or intake session.

A lot can be gained, both in leveling work pressure and in decreasing waiting times, by adjusting the number of

✉ S. P. J. van Brummelen
s.p.j.vanbrummelen@utwente.nl;
s.vanbrummelen@sanquin.nl

[1] Centre for Healthcare Operations Improvement and Research, University of Twente, Enschede, The Netherlands

[2] Donor Studies, Sanquin Research, Plesmanlaan 125, 1066 CX Amsterdam, The Netherlands

[3] Stochastic Operations Research, University of Twente, Enschede, The Netherlands

deployed staff members based on the expected arrival pattern of donors. With many part-time employees, as is the case at Sanquin, it seems to be possible to combine short and longer shifts to improve the effectiveness of the staff scheduling. For this purpose we developed a method using queuing theory and an ILP formulation to take advantage of the patterns in arrival intensities. The proposed method determines starting times and durations of all shifts such that the total number of worked hours is minimized, with certain restrictions on shift lengths. At the same time, the method takes a waiting time restriction into account. A number of ways of implementing these waiting time restrictions are possible. We will show two methods: the first is based on a sojourn time percentile (e.g., 95% of donors should spend less than 60 min in the collection site) and the second is based on an average waiting time, calculated by a slightly more complex and realistic queuing model. Finally, we will use some numerical results to show that this method can be implemented without increasing the total number of working hours.

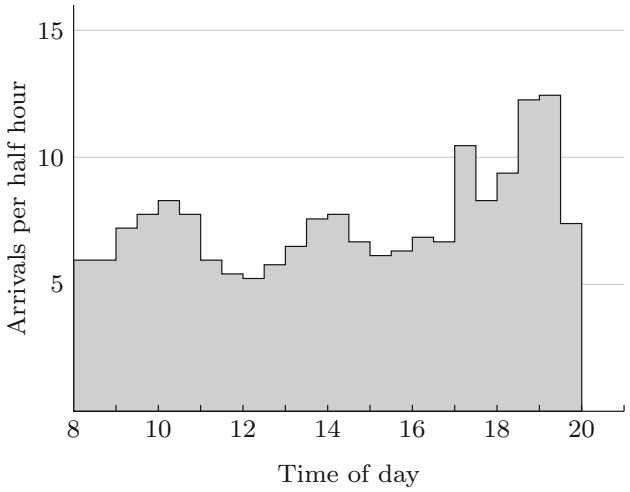

Fig. 1 A typical arrival pattern for a collection site that is opened the whole day

Process description

From a process point of view, there are two differences between whole blood and plasma donations. The first difference regards the arrivals. Before arriving at a collection site, plasma donors make an appointment. Sanquin aims, and mostly succeeds, to spread out plasma donations over the day. For whole blood donations there is no appointment system. To be able to control the number of arrivals of whole blood donors, Sanquin sends out invitations to a selection of whole blood donors by post card once a week. Although donors are encouraged to wait for an invitation and to come at their earliest convenience after receiving the invitation, neither is required. Donors may walk in and donate whenever they like, provided their eligibility to donate at that particular time. Arriving whole blood donors also show clear preferences for certain times of the day, as can be seen in Fig. 1. Some days of the week are more popular than others, but the time preferences do not seem to depend on the day of the week. The peaks in arrivals do depend on the opening hours of the collection site.

After arrival, the first two phases of the blood collection process are essentially the same for whole blood and plasma donors. When donors enter the site, they possibly enter a queue before registration at the registration desk. Subsequently they are asked to fill out a questionnaire in the waiting area. These steps can be regarded as the first phase of the donation process—the Registration phase. After filling out the questionnaire, donors enter the waiting area, and wait for a staff member to come pick them up for the Testing phase, the second phase in the process. Although the Testing phase is the same for whole blood donors and plasma donors, plasma donors have priority over whole blood donors. In the Testing phase, blood pressure and hemoglobin levels are measured, and the questionnaire is discussed with the donor. Note that the Testing phase can be executed by a general staff member, not necessarily by a physician. When there is no reason to reject the donor for a donation after the Testing phase, donors will be asked to continue to the third and last phase of the process, the Donation phase.

The second major difference between plasma and whole blood donations shows up in the Donation phase. All three elements of the Donation phase—starting the donation, blood or plasma collection, and ending the donation—require more time and equipment for plasma donations compared to whole blood donations. The equipment for plasma donations could technically be used for whole blood donations, but this is never done because of the much higher cost of plasma equipment. Collection sites handle the differences between the donations types differently. Some choose to completely separate the Donation processes of plasma and whole blood donation, and some sites choose to share staff members.

After the Testing phase, the donor is either asked to wait in the waiting area, or on a blood donor chair. In both cases the donor waits for a staff member to either show them to a blood donor chair and then start the donation or straight away start the blood donation. If the collection site uses shared staff between whole blood and plasma donations, the plasma donor is again serviced with priority. After starting the donation, the staff member can help other donors until the donation is done. When ending the donation, the staff member disconnects all equipment, and donors can take some refreshments or directly leave the collection site.

Two types of staff members can be distinguished at collection sites: general staff members and physicians. All tasks described above are handled by general staff members. The physicians have to be present in case of a complication during the donation. The physicians also handle the first interview of a new donor. For these tasks, a collection site always has one physician present. Therefore, the described method only focuses on scheduling general staff members.

We modeled this process by a tandem queue with 3 phases, as visualized in Fig. 2: the Registration phase (phase 1), the Testing phase (phase 2), and the Donation phase (phase 3). The combined three phases have a mean service time of around 20 min for whole blood donors, meaning that a center in theory can handle 3 donors per staff member per hour. To make sure waiting time remains acceptable, Sanquin has committed herself to making sure that, at every collection site, 85% of all whole blood donors spend less than 45 min in the blood donation process. However, this service level has only implicitly been taken into account when scheduling staff members. Staff is scheduled on the basis that every staff member should help 2 donors per hour. As every staff member could help 3 donors per hour if they were working at full capacity, Sanquin reckons that waiting times and breaks have been taken into account by using the lower capacity. This, however, has never been formalized. Our model combines waiting time estimations with staff scheduling.

Theoretical background and modeling

From a theoretical perspective, this paper will combine two different disciplines from the field of Operations Research: Mathematical programming to optimally schedule the staff shifts, and queuing theory to include waiting time targets when scheduling these shifts. This will result in a two-step approach, in line with the terminology used in the extensive review on staff scheduling for service systems by Defraeye and Van Nieuwenhuyse (2016).

When faced with an arrival pattern, such as in Fig. 1, a number of options can be thought of to determine the shifts for staff. The first option is to simply ignore the existence of a pattern, and to ensure that enough staff is available at the peak of the arrival intensity, and scheduling this number of staff members the entire day. This way, excess capacity is available during the remainder of the day. This is common practice at Sanquin. The second option is to break up the day in a few shifts. This way extra staff can be scheduled only for the peak arrival intensity during a shift, thereby reducing excess capacity. If the intervals are made shorter, the over-capacity is reduced even further.

A combination of overlapping shifts and varying starting times could reduce excess capacity, while preserving viable shift lengths. For example, if we would have 3 intervals, and the staff requirements would be 1, 2, 1 respectively, we would be able to cover this with two shifts, both spanning two intervals, one starting the first interval and another starting the second interval. As this eliminates excess capacity, we can guarantee that this is the optimal solution, where the shortest shift length remains two intervals. However, for large instances, such as the one at Sanquin, it is extremely hard to come up with a solution by hand. And, since there is no way to avoid excess capacity completely, there is no way of knowing how good the solution is. Using mathematical programming, this problem can be formulated as an integer linear program (ILP). Using commercially available solvers, ILP models can usually be solved to optimality.

Before the ILP can be used, the required number of staff members first has to be determined. This can be done in a variety of ways. The most simple is the one currently in use at Sanquin. This computation is based on the presumed number of donors a staff member should help in an hour. Currently, Sanquin has set this number to 2.0. This means that for every staff member present in a collection site, the site should collect 2.0 donations per hour. However, even if the staff shifts perfectly match the number of required staff members, waiting times will inevitably occur due to random variations in arrivals and service times. In queuing theory, it is well known that working at full capacity will result in extremely long waiting times. Using queuing theory, the minimum number of staff members can be determined taking waiting times into account. This will increase the required number of working hours, as it will prevent the system to work at full capacity.

Summarizing, there are two competing effects on the total number of working hours. On one hand, the inclusion of flexible staffing could result in a decrease of the number of working hours. The inclusion of waiting times, on the other hand, may require an increase of the number of working hours. This raises the following question: What will happen when both flexible staffing and the inclusion of queuing theory are combined at blood collection sites? The proposed two-step approach in this paper will be employed to answer this question.

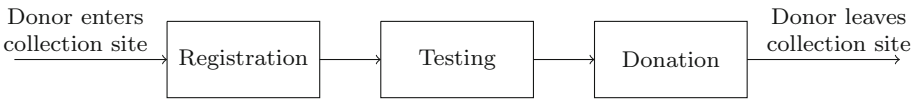

Fig. 2 Model

The paper will be structured as follows. We will start with a literature discussion in section "Literature." A more detailed and technical discussion of the mentioned methods will then be given in section "Methods." Finally, we will provide numerical results for a general approach, in which data from multiple collection sites are combined to give an impression of the average potential of the described method. The paper will be concluded with a discussion.

Literature

To our knowledge, no literature on staff scheduling at blood collection sites currently exists, although a few papers deal with related issues. These papers are discussed in section "Blood collection sites." Due to the absence of papers in the exact topic of this paper, we will divide the literature discussion in two parts. The first part discusses staff scheduling papers with applications outside blood collection sites. The second part discusses papers that deal with logistical challenges at blood collection sites different from staff scheduling. The section will be concluded with the contribution of this paper to the literature.

Staff scheduling

The literature on staff scheduling in general is very extensive, as can be seen in the review by Ernst et al. (2004). Most of the papers in the staff scheduling literature cover the same two basic steps used in our paper: first determining staff requirements within some time intervals, and subsequently determining the optimal shifts to covers these requirements.

The review by Defraeye and Van Nieuwenhuyse (2016) focuses on staff scheduling for non-stationary systems. From a technical point of view, that is exactly what we are trying to achieve for blood collection sites. This review describes a total of 62 papers. Most of these papers use a single queue to calculate their performance indicators. Of the 62 papers, only six use a network of queues. All of these use simulation, while three papers also use an analytical approximation—Izady et al. (2012), Zeltyn et al. (2011), and Fukunaga et al. (2002). From a technical point of view, the paper by Izady and Worthington Izady et al. (2012) is the most closely related to our paper, as they use similar performance indicators: sojourn time percentiles and average waiting times. The other three papers that use a network of queues—Sinreich and Jabali (2007), Ahmed and Alkhamis (2009), and Centeno et al. (2003)—only use simulation for performance evaluation. Two of these papers, the papers by Ahmed and Alkhamis (2009) and by Centeno et al. (2003), use a sojourn time percentile for performance evaluation, like our paper.

None of the papers discussed in the review by Defraeye and Van Nieuwenhuyse (2016) cover blood collection sites or blood banks in general, but eight papers discuss a health care setting. Of those papers, seven are applied to an emergency department—Izady et al. (2012), Zeltyn et al. (2011), Sinreich and Jabali (2007), Ahmed and Alkhamis (2009), Centeno et al. (2003), Defraeye and Van Nieuwenhuyse (2013), Green et al. (2006), and one is applied to ambulance services—Erdogan et al. (2010). The paper by Defraeye and Van Nieuwenhuyse uses a similar performance evaluation—the expected waiting time, but does not use a network of queues, and therefore was not mentioned in the previous paragraph. Although they evaluate the average waiting time, it is not included in their performance goals. The only paper that does use a network of queues and is not applied to health care is the paper by Fukunaga et al. (2002), which is based on a call center. Although call centers are the most frequent application of time-dependent staff scheduling methods in papers, these systems mostly use a single queue for performance evaluation.

As can be seen in the mentioned papers, most recent papers surrounding staff scheduling have used simulation as a tool to estimate waiting times. This has the advantage that it can handle very large and complex systems, but because our systems are limited in size, an analytic model is a faster and a more consistent way to calculate waiting times.

Blood collection sites

The process of blood collection is experienced millions of times per year worldwide, but literature on blood collection sites is rare. This is illustrated and confirmed by the recent review of blood management literature by Baş et al. (2016). It is also mentioned that long waiting times are associated with non-returning donors (e.g. see McKeever et al. (2006)), which stresses the importance of research into the logistics of blood collection sites.

Three papers have used simulation to study service and cost issues for collection sites—Alfonso et al. (2013), Brennan et al. (1992), and Pratt and Grindon (1982). The paper by Alfonso et al. (2013) first describes French blood collection sites as a petri net and then uses simulation to evaluate the petri net description of the collection site. Brennan et al. (1992) employ a simulation model for blood collection sites. Because of concerns regarding long waiting times, they evaluate different set-ups, staff allocations, and work rules for blood collection sites. Pratt and Gridon (1982) use a simulation model to evaluate different donor arrival strategies.

The paper by Testik et al. (2012) also deals with donor arrivals. The paper reports on the application of data

mining techniques to acquire hourly donor arrival rates. They then determine the minimal required number of staff members based on these arrival patterns.

A paper by Bretthauer and Côté (1998) discusses a method to plan resource requirements for general health care systems. Their model is mainly aimed at a high level on planning, answering questions like, 'How many employees do we need to hire?' and 'How many machines do we need?' To illustrate the use of their model, it is applied to a blood collection site.

De Angelis et al. (2003) discuss a similar problem of determining the number of servers for each phase of a health care process. This paper, like those mentioned at the start of this section, is also based on simulation. They include a blood collection site in Rome as a case-study for their method.

In our recent work, van Brummelen et al. (2015), analytic results provided to evaluate waiting and sojourn times for Dutch blood collection sites are shown. The paper deals with both average waiting times: waiting time distributions for separate phases of the blood collection site and a total sojourn time distribution. The results from this paper can be used to justify the independent calculation of waiting times for the phases of the process.

The paper that is most closely related to our paper from a practical point of view is the paper by Blake and Shimla (2014). In this paper, a blood collection site is modeled as a flow shop, and then the results are adjusted for uncertainty by describing every station as an M/M/s queuing model. This is close to how we will model the blood collection site. They calculate the minimum number of staff members required for each of the phases in their setup by setting waiting time restrictions for individual phases.

Contribution

The main contribution of this paper is the combination of exact methods from two fields of research in Operations Research—queueing theory and Integer Linear Programming—to incorporate waiting time estimation in the determination and planning of staff capacity at blood collection sites. More precisely, we expand the waiting time estimation of Blake and Shimla (2014) to be able to include waiting and sojourn time restrictions on the total blood collection process. Different queuing computations will be used for this purpose. Second, to actually minimize the number of staff working hours, we will use an ILP model to schedule shifts based on the required number of staff members. These required numbers are either based on these waiting time restrictions or on a production standard. The ILP is able to incorporate fluctuating arrivals to the blood collection sites.

We note that the methods to compute the waiting time and the optimal shifts are not new methods from a mathematical perspective. However, the combination of the methods at blood collection sites, or even health care systems in general, has not been reported on before. Additionally, we have shown in recent work—van Brummelen et al. (2015)—that modeling the blood collection site as a tandem queue gives a good approximation of the waiting times. The combination of this queuing model and a small ILP model results in a fast computation of good shift options for practical purposes.

Methods

The first step in our two-step procedure will be based on methods from queuing theory, to determine the minimum number of required staff members (sections "Production standard" and "Queue modeling"). The second step will use an Integer Linear Program (ILP) to schedule shifts, taking the minimum requirements into account (section "The ILP model").

Before discussing the specific methods, it is important to note the difference between:

- A production standard η (used for the production standard method, section "Production standard"). The production standard entails the number of donors that *should* be helped by a staff member every hour.
- A service capacity μ (used for the methods M/M/s and network model, sections "M/M/s" and "Network model," respectively). A service capacity entails the number of donors that *could* be helped by a staff member every hour.

Production standard

Currently, Sanquin uses a production standard when scheduling their staff. This means that for every staff member, a fixed number of donations η have to be completed every hour. Currently, η is set at 2.0. From a utilization standpoint it can be argued that this production standard can be increased, as the average time a staff member is needed during the donation process is less than 30 min. However, we can conclude from basic queuing theory that increasing the production standard would undoubtedly lead to substantially longer waiting times. This argument is also used to not increase the production standard to 2.5 or even 3.0, numbers that imply an average service time closer to the actual average service time of 20 min. Although this argument is valid, the exact implications of increasing the production standard are unknown, as this method does not include waiting time estimation.

To model time-dependent arrivals, an arrival pattern has been included, an example of which is shown in Fig. 1.

This arrival pattern specifies which part of the arrivals is expected in each half hour interval. This means that the minimum number of staff members will also be calculated for every half hour during the opening hours of the system. In the results we will use the arrival pattern observed by van Mechelen and Zonneveld (2013). A uniform and user-specified arrival pattern is also included in the tool for collection sites.

If λ_h is the arrival rate in half hour h, the minimal required number of staff member to be present B_h can be calculated by

$$\underline{B_h \text{ for Production Standard method :}}$$
$$B_h^{(1)} = \lceil \frac{\lambda_h}{\eta} \rceil. \tag{1}$$

This can be used as an input for an ILP model, to allow for the optimization of staff shifts, see section "The ILP model."

Queue modeling

In this section, two methods to calculate the minimum number of required staff members will be discussed. In contrast to the production standard method from section "Production standard," these methods will take waiting time into account. For this purpose, we have considered multiple options to be implemented, of which two will be described in this section. Like the production standard, these methods will determine the minimum number of staff members required for every half hour during the opening hours of the collection site.

For both methods discussed in this section, we will assume exponential distributions for both inter-arrival times and service times. For arrivals, this seems like a natural assumption, as it implies that arrivals are independent, a likely situation because there are no appointments. For services, we have also assumed exponential times. We mention two justifications for this. The first is a lack of reliable data on service times. The second, more important, reason is that exponential service times seem to model waiting times in Dutch blood collection sites closely, as shown by van Brummelen et al. (2015).

M/M/s

As a first simple option, we could model the collection site as a standard M/M/s multi server system, with a service time equal to the sum of the service times of the individual phases of the process. This can be justified if it is assumed that a staff member follows the donor throughout the system. Although this is not applied at Dutch blood collection sites, it is used in blood collection. This practice is commonly referred to as "go with the flow."

Exact formulas are known to calculate the average waiting time, the average delay and even the waiting time and the delay distribution (e.g., Chapter 20 of Winston 2004). As previously mentioned, the official Sanquin policy is that 85% of the donors should spend less than 45 min in the collection site. This means that the 85th percentile of the delay distribution should be lower than 45 min. By using Eq. (2), we can check this, and possibly other, service goal for a given staff level.

The main drawback of this model is that it is not possible to take the system's multiple phases into account. The model simply takes an average occupancy for the entire system. This is a problem because the relation between occupancy and queue lengths is not linear, but has a limit at Infinity if the occupancy becomes 1. In reality, the process steps are interrupted and not all phases take the same time, such that the occupancy will not be the same for each of the phases. If the variations in occupancy are large, the M/M/s model might give approximations for the waiting time that are too optimistic.

If a percentage α of the donors have to have a delay lower than t hours, then the minimal required number of staff members B_h can be calculated using Eq. (2).

$$\underline{B_h \text{ for M/M/s method :}}$$
$$\begin{aligned} B_h^{(2)} = \quad & \text{minimize} \quad s \\ & \text{subject to} \\ & e^{-\mu t}\left(1 + \mathbb{P}(j \geq s)\frac{1 - e^{-\mu t(s-1-s\rho)}}{s-1-s\rho}\right) < 1 - \alpha \end{aligned} \tag{2}$$

Here $\rho = \lambda/(s * \mu)$ with λ and μ the arrival rate and service rate, respectively. λ and μ should use the same time unit as t. $\mathbb{P}(j \geq s)$ represents the probability that there are as many or more donors than there are staff members available. This can easily be calculated using standard M/M/s formulas.

Network model

The second more complicated, but also more realistic modeling option, could be to use some form of a queuing network. These kinds of models incorporate the fact that the system has multiple phases and multiple servers working at each phase. This allows us to use the full model, as depicted in Fig. 2. Although it is still possible to calculate sojourn time distributions, as shown by van Brummelen et al. (2015), this is a very time-consuming process. Therefore, for network models this paper will only deal with the average waiting time.

The queueing network analyzer (QNA)—Whitt (1983)—will be used to calculate average waiting times in a queuing network. QNA is based on a set of approximative expressions using the coefficients of variation of the external arrivals and coefficients of variation of preceding phases. Due to the serial nature of the system at the Dutch blood bank, the original expressions can be slightly simplified. The expression below describes how the coefficients of variations of departures depend on the coefficients of variation of the arrivals and the parameters of the station in question. Because there is no splitting and superposition of donor flows, the coefficients of variation of the departures are the same as the coefficients of variation of the arrivals at the next station. The description of the parameters and variables used can be found in Table 1.

$$C^2_{a(i+1)} = C^2_{di} = 1 + (1-\rho_i^2)(C^2_{ai} - 1) + \frac{\rho_i^2}{\sqrt{s_i}}(C^2_{si} - 1).$$

Then, if we let $\mathbb{E}_{\mathrm{M/M/s_i}}(W_i)$ denote the expected waiting time for an M/M/s_i queue by

$$\mathbb{E}_{\mathrm{M/M/s_i}}(W_i) = \frac{(s_i\rho_i)^{s_i}}{s_i!\left(\sum_{n=0}^{s_i-1}\frac{(s_i\rho_i)^n}{n!} + \frac{(s_i\rho_i)^{s_i}}{(1-\rho_i)s_i!}\right)(1-\rho_i)^2 s_i}$$

then $\mathbb{E}_{s_i}(W_i)$ can be calculated by using

$$\mathbb{E}_{s_i}(W_i) = \frac{C^2_{ai} + C^2_{si}}{2}\mathbb{E}_{\mathrm{M/M/s_i}}(W_i).$$

Because a donor can only visit a phase once, the expected Delay $\mathbb{E}(T_i)$ can be calculated by

$$\mathbb{E}_{s_i}(T_i) = \mathbb{E}_{s_i}(W_i) + \tau.$$

If the average total delay has to be lower than t min, B_h can be computed by solving Eq. 3.

$$\underline{B_h \text{ for QNA method :}}$$

$$B_h^{(3)} = \begin{array}{ll} \text{minimize} & \sum_{i=1}^{3} s_i \\ \text{subject to} & \sum_{i=1}^{3} \mathbb{E}_{s_i}(T_i) < t \end{array} . \quad (3)$$

This is an integer, non-linear optimization problem, so in general it is very hard to solve. But, since there are only a finite number of configurations of the staff—for a typical blood donor center this could be 1 or 2 staff members at phase 1, 2 to 4 at phase 2, and 3 to 6 at phase 3—, we could solve this by applying brute force, i.e., checking every possible combination of staff members between some lower bound and upper bound. It is possible to do this for every interval that has to be scheduled, and then these numbers can be used as input for the ILP model of section "The ILP model."

QNA also uses coefficients of variation of the inter-arrival times and service times, meaning that these are not required to be exponential. Although this is very useful in most systems, we have decided to set these coefficients to 1 for the blood collection site, resulting in exponential service times, as discussed previously.

Table 1 Parameters and variables used

Parameters	
η	Production standard
τ	Total expected service time
τ_i	Service time at phase i
μ	Service rate ($= 1/\tau$)
μ_i	Service rate at phase i ($= 1/\tau_i$)
λ	Arrival rate
s_i	Number of servers at phase i
ρ_i	Utilization, $= \lambda\tau/s_i$
C^2_{si}	Squared coefficient of variation of services at phase i
C^2_{di}	Squared coefficient of variation of departures at phase i
C^2_{ai}	Squared coefficient of variation of arrivals at phase i
Variables	
W	Total waiting time
W_i	Waiting time at phase i
T	Total sojourn time (delay)
T_i	Delay at phase i

The ILP model

Once the minimum number of staff members $B_h^{(i)}$, determined by Eqs. (1), (2) or (3), has been established, the ILP can now be formulated to determine optimal shifts lengths and starting times. This ILP is given in Box 1. The parameters and variables are explained in Table 2.

All the restrictions have their own implications, which can be interpreted as follows:

1. This restriction ensures that there are at least as many staff members present as the restrictions calculated by any of the three options discussed in sections "Production standard," "M/M/s," and "Network model."
2. This restriction ensures that there are at least as many staff members present as the minimum number required. This is not a value that has been calculated, but some value that has been set as an absolute minimum by the user of the algorithm. This is to ensure that some minimum number of staff members is always present. For example, the M/M/s model is based on a single station, and could require only one staff member. However, all stations have to be manned at all times, requiring at least three staff members.

Table 2 Parameters and variables for the ILP model

Indices	
h, h'	Half hours
t	Shift length
Parameters	
k_t	Cost of a staff member for shift duration t
$B_h^{(i)}$	Required number of staff members present at half hour h (calculated by method i)
mn_h	Minimum number of staff members at half hour h
$q_{t,h,h'}$	1 if $h \leq h' < t + h$ and a shift of length t, starting at half hour h is allowed, 0 otherwise
Variables	
$x_{t,h}$	Starting shifts at half hour h of length t
y_h	Staff members present at half hour h
$z_{t,h,h'}$	Number of breaks at half hour h' of a staff member that has a shift length t and started at half hour h

3. This restriction converts $x_{t,h}$, the starting shifts for staff members, to y_h, the number of staff members present. It also makes sure that shift lengths that are not allowed do not convert to staff members that are working.
4. This restriction ensures that there is enough slack in the schedule to give everyone that is entitled to a break can get a break.
5. This restriction ensures that the solution is integer, i.e., no fractions of staff members.
6. This restriction ensures that the solution is integer, i.e., no fractions of breaks.

The costs k_t can be seen in Table 3. The costs are set such that the model will always select one longer shift rather than a combination of two sequential shorter shifts, by making a longer shift slightly cheaper than the combined cost of two shorter shifts. The difference is small enough that longer shifts will not be selected if a combination of two shorter shifts results in less working hours.

Given the calculated minimum staff levels and the ILP model, we will use commercially available packages to compute the optimal solution. We have used AIMMS 4.5.2 to build the ILP model and its restrictions and use CPLEX 12.6.1 to solve the ILP. Even for the biggest Sanquin cases—collection sessions of 12 h, the solver reaches the optimal solution within a second.

Table 3 The costs associated with the various shift lengths

Shift duration	Costs
2	2
3	3
4	3.99
5	4.99
6	5.98
7	6.98
8	7.97
9	8.97

Results

Current situation (base scenario)

The exact method that Sanquin uses to schedule staff has not been formalized. Based on discussions with employees and team leaders, we may conclude that the method that is closest to reality—which will therefore be used as a base scenario in this section—is the production standard method that has been presented in section "Production standard". A production standard of 2.0 is used to determine the minimum required number of staff members.

Staff members are scheduled for an entire session, except for long sessions, which are split into two shifts, but these two shifts usually have the same number of assigned staff members. This means that Sanquin will usually staff the number of employees that are required during peak hours for the entire day. Employees will get a shift length equal to either the total or half of the session length plus some additional time before opening and after closing the collection site. This extra time is required to set up and shut down equipment, respectively.[1] In Table 5, this method of shift planning will be called "session shifts." As it is closest to the current situation, it will be referred to and used as the base scenario, indicated with * in Table 5.

Alternative scenarios

Table 5 shows the three different methods to calculate the minimum number of required staff members, B_h, that were presented in this paper: production standard, M/M/s, and network modeling. The last two are accompanied by a waiting time restriction. For M/M/s this is the probability that the delay time, i.e., the total time spent in the system, will exceed a certain threshold. For the network model this

[1] As this extra time is required, it is included in all scenarios for employees that work the first or last shift.

is a restriction on the total mean waiting time. These restrictions should hold for every half hour, meaning that a busy period with long waiting times cannot be compensated for by a quiet period with very short waiting times. Note that the individual restrictions of the M/M/s and network models are not linked. For example, we do not claim that an expected waiting time below 5 min implies that less than 12 % of donors spend longer than 45 min at the blood collection center.

Table 5 also includes a distinction between scenarios that only allow session shifts, as explained in section "Current situation (base scenario)" and scenarios that allow "flexible shifts." Flexible shifts, in this case, allows for shifts that start at any half hour during the day (e.g., 9.00, 9.30, 10.00, etc.) and last a whole number of hours between 3 and 9 h. Finally, Table 5 includes results for a production standard/service capacity of 2.0, 2.5, and 3.0. It is important again to note the difference between the production standard (used for the production standard method) and the service capacity (used for the M/M/s model and network model). This means that a service capacity of 3.0 seems reasonable, as the total process has a service time of approximately 20 min, but a production standard of 3.0 results in extremely long waiting times.

To get an impression of the results that can be achieved by the proposed combination of queuing and ILP, 35 instances will be used for every scenario. Table 5 shows the average result of all these instances for every scenario. The instances are a combination of 5 arrival rates for all of the 7 session types that Sanquin distinguishes. These 7 session types are shown in Table 4. The average donor arrival rates per hour that were used range from 12 to 20, with increments of 2.

A few scenarios for the M/M/s method are shown to be not possible (NP). In these cases, the tail of the service time distribution exceeds the required probabilities. This means that even without any waiting time, the delay time restriction still cannot be met due to the assumed stochasticity of the exponential service times. This cannot happen with the network model, as it places a restriction on the waiting time. The waiting time can be arbitrarily close to 0 if enough staff is added.

Table 4 Session types at Sanquin and their opening hours

Session name	Opening hours
O1	8.00–11.00
O2	8.00–12.00
OM	8.00–15.30
MA	12.30–20.00
A1	16.00–20.00
A2	17.00–20.00
OMA	8.00–20.00

As a first observation, it can be seen that a waiting time restriction increases the required staff hours. By just introducing a waiting time restriction, while still assuming a service capacity of 2.0, staff hours increase by up to 43.3 % if waiting times are only allowed to be 5 min. However, it is safe to assume a higher service capacity for these queuing methods. Even a very safe service capacity increase to 2.5 decreases the increase of staff hours to an increase of at most 17.9 %, and is even able to completely negate the increase for the M/M/s method for the $\mathbb{P}(T > 60 \text{ min}) < 0.15$ case—the lower waiting time restrictions are still impossible. When increasing the service capacity to a realistic 3.0, all but one scenario show a decrease in the number of staff hours.

The second main observation is, as expected, that flexible staffing results in significant savings on staff hours. By just introducing flexible staffing, i.e., comparing session shifts and flexible scenarios with the same further settings, savings are around 20 %, ranging from 20.4 % for the network model with a waiting time restriction of 5 min and a service capacity of 2.5 to 26.5 % for the production standard method with a production standard of 2.0.

The benefits of flexible shifts are again shown in Fig. 3. This shows the effect of additional shift length options. It is based on an average of the 9 scenarios for the network model from Table 5, and results are expressed as a percentage of the session shift option. The first data point is the number of hours that are needed to staff the collection site if only 9 h shifts are allowed, the second data point adds shifts of 8 h, etc. Only the data from OMA sessions (see Table 4) were taken into account, because the other sessions are not opened for 9 h, making the 9 h shifts redundant in these sessions. If only 9 h shifts are allowed, flexible shifts are worse than session shifts. This has to do with the fact that two 9 h shifts cover more that the total session, while one is not enough. A combination of 8 and 9 h shifts still shows the same effect, but it is significantly

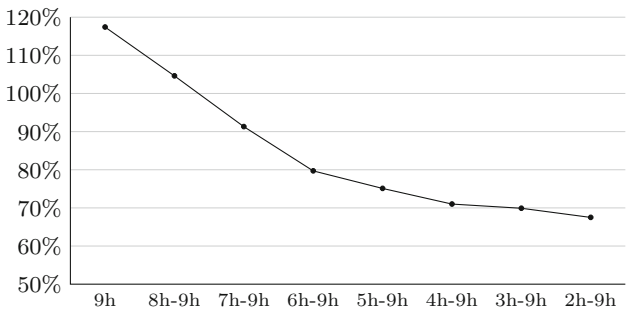

Fig. 3 Effects of adding extra shift length possibilities on number of working hours for the OMA session (see Table 4). Required number of staff members based on the network model. Results are in number of working hours as a percentage of session shifts and are based on an average of the 9 different scenarios for the network model included in Table 5

Table 5 Average changes in staff hours based on all session types and multiple collection site sizes compared to the current situation (*)

Method	Possible shifts	Waiting time restriction	Service capacity[a]		
			2.0	2.5	3.0 (%)
Production standard	Session shifts	N/A	*	–18.6%	–32.0
	Flexible shifts	N/A	–26.2%	–40.1%	–49.5
M/M/s	Session shifts	$\mathbb{P}(T > 45 \text{ min}) < 0.12$	NP	NP	–10.2
		$\mathbb{P}(T > 45 \text{ min}) < 0.15$	NP	NP	–17.8
		$\mathbb{P}(T > 60 \text{ min}) < 0.15$	27.4%	–7.9%	–23.9
	Flexible shifts	$\mathbb{P}(T > 45 \text{ min}) < 0.12$	NP	NP	–29.5
		$\mathbb{P}(T > 45 \text{ min}) < 0.15$	NP	NP	–36.3
		$\mathbb{P}(T > 60 \text{ min}) < 0.15$	–1.8%	–29.5%	–42.0
Network model	Session shifts	$\mathbb{E}(W) < 5$ min	43.3%	17.9%	3.2
		$\mathbb{E}(W) < 10$ min	31.6%	8.4%	–6.6
		$\mathbb{E}(W) < 15$ min	26.6%	3.9%	–11.5
	Flexible shifts	$\mathbb{E}(W) < 5$ min	12.6%	–6.1%	–18.5
		$\mathbb{E}(W) < 10$ min	2.8%	–15.4%	–27.4
		$\mathbb{E}(W) < 15$ min	–2.0%	–19.8%	–31.6

The methods in the first column are based on sections "Production standard," "M/M/s," and "Network model," respectively. In case a result shows NP, it is not possible to meet the waiting time restriction with this service capacity, irrespective of the capacity used

[a] Note that in the case of a production standard method the production standard is equal to the service capacity

Box 1 The ILP model

Minimize

$$\sum_t \sum_h x_{t,h} \cdot k_t$$

Subject to:

(1)	$y_h - \sum_{t=12}^{18} \sum_{h'} z_{t,h',h} \geq B_h^{(i)}$	$\forall(h)$	$B_h^{(i)} from equation(i)$
(2)	$y_h - \sum_{t=12}^{18} \sum_{h'} z_{t,h',h} \geq mn_h$	$\forall(h)$	
(3)	$\sum_t \sum_h x_{t,h} \cdot q_{t,h,h'} = y_{h'}$	$\forall(h')$	
(4)	$x_{t,h} \leq \sum_{h'=h+1}^{h+t-1} z_{t,h,h'}$	$\forall(h),\ t \geq 12$	
(5)	$x_{t,h} \in \mathbb{N}$	$\forall(t,h)$	
(6)	$z_{t,h,h'} \in \mathbb{N}$	$\forall(t,h,h')$	

reduced. Also note that the marginal effect decreases, the additional effect of adding 6 h shifts is much larger that the additional effect of adding 2 h shifts. This means that a large portion of the beneficial effects of flexible shifts can already be achieved without very short shifts.

Finally, a combination of flexible staffing and a waiting time restriction almost exclusively results in savings of staff hours. Even for a safe service capacity assumption of 2.5, savings are substantial for all included waiting time requirements. For a realistic service capacity assumption of 3.0, savings are at least 18.5 % compared to the current situation, and savings go as high as 42.0 %, while still guaranteeing that 85 % of all donors spend at most 60 min at the collection site.

Discussion

With the presented combined approach, substantial savings on personnel are a possibility—assuming the results can be followed exactly with regard to employment contracts and assumptions on the current scenario turn out to be correct. At the same time, by aligning employee shifts and arrival patterns, it is possible to include waiting or delay time

restrictions. Generally, three observations can be obtained from the results:

1. By including waiting time restrictions, an increase in staff working hours will be required.
2. By using flexible shift planning, substantial savings on working hours by staff can be obtained.
3. By combining flexible shift planning and waiting time restrictions, no extra staff is needed, and generally a small saving on staff hours remains a possibility.

Most of these savings originate from a more flexible way of scheduling the shifts of staff members, in which shorter shifts are made possible. In the flexible staffing in our "Results" section, we allowed for all shifts lengths from 3 to 9 h, but other shift possibilities and restrictions can easily be incorporated, depending on specific requirements from certain blood collection sites.

If we recall from section "Process description" that the production standard of 2.0 was set to include waiting time, it is worthwhile to compare the production standard of 2.0 with some of the waiting time restrictions with the realistic service capacity of 3.0, while maintaining the session shift assumption. We can then see that the result closest to a 0% increase is for an expected waiting time restriction of 5 min, with all other restrictions resulting in a decrease of the number of staff hours. This means that in most cases, the 2.0 production standard is probably quite low, and an increase would most likely be possible in most cases. However, it is important to note that this might not hold for all collection sites. Especially small centers might still see a sharp increase in waiting and delay times if the production standard would be increased.

Clearly, the more advanced approximate results for queueing networks could benefit the method described in this paper. However, these results would most likely require more computational time, which might affect the applicability of the method in practice.

If the method were to be applied at Sanquin collections sites, two issues could cause a difference between the model and the results from reality after implementation. First, as the method does not take actual, individual employment contracts into account, and does not assign employees to shifts, realistic savings would probably be a bit lower. However, the savings would still be substantial enough to implement the system, especially for very long sessions. A second possible discrepancy might be caused by the waiting time estimation. Although a large difference between the results from the model and reality is unlikely, especially for the network model, some differences might still occur. Although some methods exist that would likely lead to smaller differences between reality and the model, these methods come with their own downsides. Simulation could be an alternative method to give a more realistic approximation. However, this will most likely slow down calculations and would eliminate the possibility for an exact answer. Most importantly, though, it would not be generic and would require adapting the simulation model to each individual collection site.

Since we can combine significant savings with waiting time guarantees and fast calculations—individual cases are solved in a matter of (milli)seconds, Sanquin investigated practical consequences of implementing the proposed approach with favorable results. The next step will be to actually apply the approach.

Acknowledgements We wish to thank the anonymous reviewers for accurately reading the manuscript and their helpful comments.

References

Ahmed, M.A., and T.M. Alkhamis. 2009. Simulation optimization for an emergency department healthcare unit in Kuwait. *European Journal of Operational Research* 198 (3): 936–942.

Alfonso, E., X. Xie, V. Augusto, and O. Garraud. 2013. Modelling and simulation of blood collection systems: Improvement of human resources allocation for better cost-effectiveness and reduction of candidate donor abandonment. *Vox Sanguinis* 104 (3): 225–233.

De Angelis, V., G. Felici, and P. Impelluso. 2003. Integrating simulation and optimisation in health care centre management. *European Journal of Operational Research* 150 (1): 101–114.

Baş, S., G. Carello, E. Lanzarone, Z. Ocak, and S. Yalndağ. 2016. *Management of blood donation system: Literature review and research perspectives*, 121–132. Berlin: Springer.

Blake, J.T., and S. Shimla. 2014. Determining staffing requirements for blood donor clinics: The Canadian blood services experience. *Transfusion* 54 (3 Pt 2): 814–820.

Brennan, J.E., B.L. Golden, and H.K. Rappoport. 1992. Go with the flow: Improving red cross bloodmobiles using simulation analysis. *Interfaces* 22 (5): 1–13.

Bretthauer, K.M., and M.J. Côté. 1998. A model for planning resource requirements in health care organizations. *Decision Sciences* 29 (1): 243–270.

Centeno, M.A., R. Giachetti, R. Linn, and A.M. Ismail. 2003. Emergency departments II: A simulation-ILP based tool for scheduling ER staff. In *Proceedings of the 35th Conference on Winter Simulation: Driving Innovation, Winter Simulation Conference*, 1930–1938.

Defraeye, M., and I. Van Nieuwenhuyse. 2013. Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. *Decision Support Systems* 54 (4): 1558–1567.

Defraeye, M., and I. Van Nieuwenhuyse. 2016. Staffing and scheduling under nonstationary demand for service: A literature review. *Omega* 58: 4–25.

Erdogan, G., E. Erkut, A. Ingolfsson, and G. Laporte. 2010. Scheduling ambulance crews for maximum coverage. *Journal of the Operational Research Society* 61 (4): 543–550.

Ernst, A.T., H. Jiang, M. Krishnamoorthy, and D. Sier. 2004. Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research* 153 (1): 3–27.

Fukunaga, A., E. Hamilton, J. Fama, D. Andre, O. Matan, and I. Nourbakhsh. 2002. Staff scheduling for inbound call centers and customer contact centers. *AI Magazine* 23 (4): 30–40.

Green, L.V., J. Soares, J.F. Giglio, and R.A. Green. 2006. Using queueing theory to increase the effectiveness of emergency

department provider staffing. *Academic Emergency Medicine* 13 (1): 61–68.

Izady, N., and D. Worthington. 2012. Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research* 219 (3): 531–540.

McKeever, T., M.R. Sweeney, and A. Staines. 2006. An investigation of the impact of prolonged waiting times on blood donors in Ireland. *Vox Sanguinis* 90 (2): 113–8.

Pratt, M.L., and A.J. Grindon. 1982. Computer simulation analysis of blood donor queueing problems. *Transfusion* 22 (3): 234–7.

Sinreich, D., and O. Jabali. 2007. Staggered work shifts: A way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Management Science* 10 (3): 293–308.

Testik, M.C., B.Y. Ozkaya, S. Aksu, and O.I. Ozcebe. 2012. Discovering blood donor arrival patterns using data mining: A method to investigate service quality at blood centers. *Journal of Medical Systems* 36 (2): 579–594.

van Brummelen, S.P.J., W.L. de Kort, and N.M. van Dijk. 2015. Waiting time computation for blood collection sites. *Operations Research for Health Care* 7: 70–80.

Van Mechelen, I. S., and P. L. M. Zonneveld. 2013. Capaciteitplanning en wachttijdbepaling, M.Sc. Thesis.

Whitt, W. 1983. The queueing network analyzer. *Bell System Technical Journal* 62 (9): 2779–2815.

Winston, W. 2004. *Operations research: Applications and algorithms*. San Francisco: Thomson Brooks/Cole.

Zeltyn, S., Y.N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, T. Lauterman, D. Schwartz, K. Moskovitch, S. Tzafrir, and F. Basis. 2011. Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation* 21 (4): 1–25.