Noname manuscript No.
(will be inserted by the editor)

EMS call center models with and without function differentiation: a comparison

Martin van Buuren · Geert Jan Kommer · Rob van der Mei · Sandjai Bhulai

Received: date / Accepted: date

Abstract In pre-hospital health care the call center plays an important role in the coordination of emergency medical services (EMS). An EMS call center handles inbound requests for EMS and dispatches an ambulance if necessary. The time needed for triage and dispatch is part of the total response time to the request, which, in turn, is a key performance indicator for the quality of EMS. Call center agents should perform the triage efficiently, so that entering calls have short waiting times, and the dispatch of ambulances must be adequate and swift to get a fast EMS response. This paper presents and compares three discrete event simulation models for EMS call centers: the first has two different call center agent classes between whom communication tasks are split, while the second has one class of call center agents that share all tasks. The third model is a combination of both. The models provide new insight into the EMS call center processes and can be used to address strategic issues, such as capacity and workforce planning. The analysis and simulations of urgent communication and decision processes in this paper are valuable to other emergency call centers[1].

Keywords EMS · Call Center · Simulation · Ambulance · Decision Support Systems · Function Differentiation

1 Introduction

In most countries a request for emergency medical services (EMS) is done by calling a nation-wide emergency telephone number. The request is answered by an agent of a call center. Depending on the country's system, this call center handles the request, or the request is passed through to a regional call center; this may be a general

Martin van Buuren
Centrum Wiskunde & Informatica, Department of Statistics, 123 Science Park, 1098XG, Amsterdam, The Netherlands
Tel.: +31-20-5924365
E-mail: m.van.buuren@cwi.nl

Geert Jan Kommer
National Institute for Public Health and the Environment, Department for Quality of Care and Health Economics, 9 Anthonie van Leeuwenhoeklaan, 3721MA, Bilthoven, The Netherlands
Tel.: +31-30-2742927

Rob van der Mei
Centrum Wiskunde & Informatica, Department of Statistics, 123 Science Park, 1098XG, Amsterdam, The Netherlands
Tel.: +31-20-5924129

Sandjai Bhulai
Vrije Universiteit Amsterdam, Department of Mathematics, 1081A De Boelelaan, 1081HV, Amsterdam, The Netherlands
Tel.: +31-20-5987679

[1] A partial and preliminary version of this paper was presented at the 2015 Winter Simulation Conference [26].

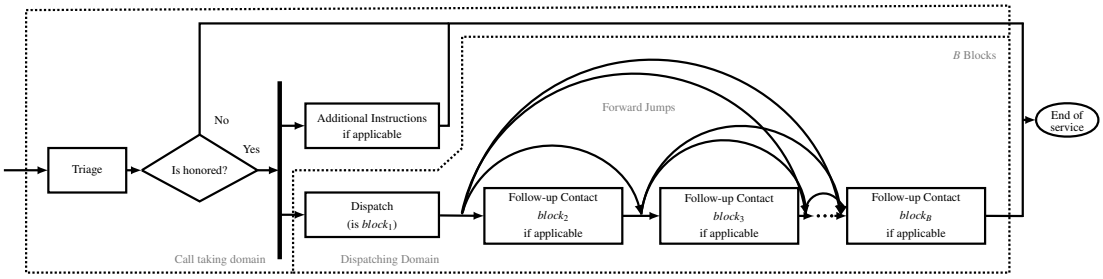

Fig. 1 High level description of the EMS call center, with the call taking and dispatching domains.

call center for emergency services or a specific medical or EMS call center. This paper focuses on EMS call centers.

The call center has the task to perform the triage and dispatch adequately, that is, the need and urgency level for EMS is to be determined properly, so that an ambulance is sent to incidents only in the case the patient really needs this service. The number of ambulances is limited and it is important to have an ambulance available for dispatch when needed. If the request for EMS is honored, an ambulance is sent to the incident and care is provided. If the patient needs hospital care, transportation to a hospital is provided.

The total time needed for taking the request, performing the triage, and dispatching an ambulance is called the call center time, which is part of the total response time. In the case of life-threatening situations short response times are important. Hence, it is essential to have short waiting, triage, and dispatch times.

The three discrete event simulation (DES) models for an EMS call center developed in this paper provide insight into these crucial variables. They use the policy of the EMS call center and number of call center agents as an input. The models simulate the communication processes at the EMS call center in high detail, and they can be used to evaluate the EMS call center's performance. Results are shown for all three models for realistic call volumes and durations, where the input contains call record data provided by a real EMS call center.

We proceed in this section with a high level model overview of general aspects of EMS call centers. Next, in Section 1.2, we provide a literature overview where we position our models. We end this section by stating the contribution of our models.

In Section 2 we describe our models from a queuing theoretical perspective. Section 3 shows realistic parameters that where obtained from a real EMS call center database and expert guesses, used to generate the results of Section 4. In the results section we also provide new insights, which are concluded and discussed in Section 5. Finally, Section 6 contains recommendations for future research.

1.1 High level model overview

Three Call Center Agent Classes Most EMS call centers have two classes of call center agents, which work in cooperation. The first class are the call takers. They handle the inbound requests and perform communication with the caller, also referred to as the applicant. Dispatchers are the second call center agent class. They take care of the outbound calls, i.e., the dispatch process and the communication with the ambulance team and hospital in various follow-up contact moments of the service. An example of a follow-up call is a request from the ambulance team for additional health condition information while driving to the incident. The dispatcher has logistic skills and can be supported by decision support software (DSS) to determine the most appropriate ambulance to send to the incident.

The just mentioned type of EMS call centers make a distinction between call takers and dispatchers. This is called function differentiation. Other call centers have one class of call center agents, called generalists, that do both tasks. One can consider a generalist as a call taker and dispatcher embedded in one person.

Figure 1 shows a high level overview where the position of the call taking domain and dispatching domain is clarified. The call taking domain contains a triage procedure, in which is decided if a request is honored,

Table 1 Models and their call center agent classes

	Model name	Call takers	Dispatchers	Generalists
1	Function differentiation	✓	✓	
2	Solely generalists			✓
3	Mixed model	✓	✓	✓

i.e., an ambulance is required. Sometimes, when the need for an ambulance has been determined, additional instructions are given by the call taker while at the same time the dispatcher is assigning a call to an ambulance. A clear example of additional instructions are the reanimation instructions to an applicant that has no medical know-how. The dispatch process and follow-up contacts are modeled through a block sequence, which is described in detail in Section 2.3.

The three EMS call center models use different call center agent classes; see Table 1.

Prioritizing Calls Requests can be made by several disjunct applicant classes, such as civilians, general practitioners, or police officers. The incoming request for each applicant class has its own priority, i.e., an ambulance service provider can choose to give civilians a higher priority than hospitals to prevent call center agents from taking a hospital line if there are civilians waiting in a life-threatening situation.

The applicant's class may affect the service-time distributions in the call taking domain. Requests applied for by a general practitioner or a police officer generally do not require extensive triage and therefore may have a shorter service time than requests from civilians. Requests from civilians always need to be triaged in order to determine the need and urgency of the service.

Requests for EMS are often categorized into urgency levels: high urgency in the case of a potential life-threatening situation, to low urgency where a patient is stable and an immediate life-saving response is not required. Low urgency requests are for instance planned transports from and to a hospital.

Recall that a request is called honored if the call taker—or generalist, depending on the model—decides to send an ambulance. Only honored requests get an urgency assigned. The urgency is used for prioritizing in the dispatch domain: dispatchers only address low urgency tasks if all potential life-threatening high urgency tasks are completed.

Note the difference between priority and urgency: the priority for an incoming request is used to decide what telephone line to answer when there are multiple waiting calls in the telephone system, while an urgency, which is assigned during the triage procedure, gives information as regards the seriousness of the injuries, and consequently how EMS should respond in the dispatch domain.

1.2 Literature Review

We consider various fields of operation research in EMS, and we treat them separately.

1.2.1 Non-Call Center Related Simulation in EMS

Simulation models are powerful techniques to understand tactical and operational problems in EMS. These models can be used to explore possible solutions to tackle certain problems in resources, facility location, and deployment of ambulances.

Most models, like the first EMS simulation model by Savas in 1969 [23], were designed as a tool for the road domain of the ambulance services, i.e., distribution or availability of ambulances. These models can be used to explore facility locations, deployment of ambulances, and resources needed for optimal services. They do not include the call center, or include it on a very basic level to handle decision rules or call center overhead corrections in response time calculations.

Henderson and Mason discuss simulation and data visualization in EMS by means of an example of a simulation model of the ambulance service in New Zealand [10]. This simulation model is used as decision

and management support tool for the logistic part of the ambulance services. Arenghieri et al. use simulation to analyze and optimize the ambulance service in Milano [3] . A recent review article on EMS simulation models is provided by Aboueljinane [1].

The TIFAR framework [27] has recently been developed for simulation of ambulance services and can be used for studying various dispatch regimes. It has been extended to include the call center domain for this paper.

1.2.2 Operation Research in EMS

Optimal facility location is an important subject in the operational field of EMS. Proper base locations help to realize short response times. Facility location problems are studied since the early seventies of the previous century and this research field has evolved in different directions. Structured overviews of the research done in facility location in EMS can be found in literature [6, 15]. An important subject of facility location in EMS is the uncertainty in demand for ambulance care and the need for low response times. This uncertainty is addressed in a review of facility location modeling [25].

Another important subject for EMS in order to realize a short response time is to dispatch the right ambulance to an incident. This dispatch problem is closely related to facility location, as an optimal facility location is a prerequisite for an optimal dispatch. Dynamic relocation and dynamic optimization of facility locations is often included in the dynamic management and dispatching of ambulances. Dynamic relocation is studied in different ways, combinatorial [7, 19, 20] and as of lately more and more by dynamic programming [2, 16, 17, 24].

1.2.3 Call Center Simulation

Regular Call Center Simulation By regular call centers we mean call centers that handle incoming calls for a service of a company, with questions on a particular subject. The literature on the more general call centers is broad and vast, in contrast to EMS call centers. Topics of research are arrival processes, optimal staffing of call centers, estimation of demand and expected future demand for services, and routing of specific types of demands. Koole and Mandelbaum give a good introduction into this subject [13], while other papers review research on customer call centers [9, 12]. The surveys show that call centers are mainly analyzed within the framework of queueing theory. In order to keep the queueing models tractable, most papers deal with a single call type with a homogeneous pool of agents. The extension to multiple call types and a heterogeneous pool of servers leads to complex and intractable models [4, 22]. Because of this complexity, simulation is an appropriate tool to analyze more complex call centers, such as EMS call centers.

EMS Call Center Simulation There are limited simulation models that focus on the call center domain of the EMS process. Kozan and Mesken have modeled the EMS call center within a simulation context [14]. They developed a simulation model to analyze the effects of varying call volumes, personnel resources, and work-load distributions on the performance of the call center. Ross studies the Toronto EMS call center and develops a simulation model to examine the effect of changes of the dispatch processes on the work-load of the call center staff [21]. For this research, communication flows at the call center are identified and different dispatch systems are evaluated. Other preliminary work on EMS call center simulation can be found [8, 18]. A preliminary version of the present paper is presented [26].

1.3 Contribution of our Models

This paper is the first on EMS call centers that:

1. contains multiple call center agent classes, as can be found in a modern EMS call center,
2. contains follow-up contact moments,
3. simulates EMS call center processes in high detail,
4. uses real EMS call center data sets as simulation input to gain insights, and
5. provides a comparison between multiple EMS call center models.

Difference Between Regular and EMS Call Centers There are two main differences between an EMS call center and a regular call center. The first is the urgency at which calls need to be taken and processed. Requirements on the quality of service of EMS call centers are much stricter than regular call centers due to the urgency. A noticeable subject is the respectable uncertainty of demand in combination with the fact that a request for EMS may not wait too long, because it might involve a life-threatening situation. As a result, we take in the numerical analysis the strict requirement of at most six seconds waiting time as a key performance parameter, whereas in a regular call center a typical response time requirement is in the order of minutes rather than seconds.

The second difference between an EMS call center and a regular call center is the fact that the EMS call center has more communicational tasks in the coordination of the EMS services. EMS call center agents communicate with hospitals and ambulance teams and if necessary also with police, firefighters, and other EMS or homeland security call centers. These tasks are also done under time pressure. Not all of these different communication processes can be captured in a mathematical model with nicely shaped distributions. The number and characteristics of these communication processes are uncertain and difficult to catch in an analytical model. Simulation models do not have the difficulty that an analytic solution to a mathematical problem needs to be determined, because they calculate the model numerically. As analytic modeling is not possible for EMS call centers, simulation techniques seem the best ways to analyze and evaluate EMS call center systems.

Contribution to Practice The implemented simulation models can be used by decision makers, managers of EMS call centers, and other policy makers in managing the operational subjects of the EMS call center. All processes are assumed to be stochastic, the simulations make use of the uncertainties in call volumes and lengths of the communications. The models give insight into the variance of the work-load in different situations, depending on the scenarios. It is possible to study economies of scale in the case of increasing call volumes. The outcomes of individual simulation runs include the work-load of the system and the waiting times for different applicant classes. These performance indicators can be examined to explore optimal staffing strategies, staffing levels, and service level requirements.

This article compares the two most common staffing policies that are used in practice to see their behavior, to find which performs better under given circumstances. A third model, the so-called mixed model, is proposed and analyzed.

The importance of an efficient EMS call center together with the small number of existing models raises the need for the development of a DES model for the EMS call center domain.

Contribution to Other Call Centers The models are also applicable to other call centers that both take calls and coordinate a service, in a context where high performance counts. Many emergency call centers, amongst others police, fire fighters, general practice centers, and even the emergency room desk, are similar to EMS call centers in the sense that they have a comparable structure of decision and communication processes while they have to deal with short time spans and multiple priorities. The line of analysis we perform in this paper can also be applied in these contexts. Also out of the scope of life-saving applications the models can be useful. In the case of a taxi service there are multiple types of incoming lines, since regular customers may have a higher priority. A taxi call centers also handles the coordination of the vehicles to incoming calls. Too long waiting times can lead to a potential customer abandoning a call and instead take his chances at the competitor, which rises the need to take a call within seconds.

2 Model Description

This section describes the three models in detail. In Section 2.1 we formulate our assumptions on the ambulance practice, followed by the model with function differentiation in Section 2.2. All models contain a block sequence, the inner working of which is explained in Section 2.3. Section 2.4 describes the second model holding the classic regime with solely generalists who do both call taking and dispatching. We conclude with the mixed model in Section 2.5.

2.1 Assumptions on ambulance practice

We assume that an ambulance region has exactly one regional EMS call center that is responsible for both the call taking and the dispatch of ambulances. As stated before, we assume that there are three classes of EMS call center agents, each with their own set of skills and costs: *call takers, dispatchers*, and *generalists*.

A request is called honored if the call taker concludes from the triage process that an ambulance must be sent to the incident. Each request that is honored gets an urgency assigned. We assume that this urgency not subject to change during the remainder of the call handling. In practice urgency mutations do occur. We assume that an ambulance service provider uses the totally ordered set of urgency classes $\mathscr{U}$.

We assume that to every honored request exactly one ambulance is dispatched. After dispatch, it can generate follow-up calls. In practice, there are situations where multiple vehicles are dispatched to a request, each generating follow-up calls.

Performance Indicators The quality standard and volume of care differ per country, and they depend on legislation, tradition, culture, and prosperity level. In some countries ambulances are staffed with paramedics, while other countries employ specially trained nurses; Hoogeveen provides an European overview [11]. Hence, various countries use different key performance indicators (KPIs) for EMS care. Most countries use a constraint for the fraction of late arrivals. KPIs of EMS call centers are not always well defined for lower urgency calls.

In this paper we use the waiting time before a call is taken, the average work-load for a call center agent, and the total call center time as KPIs. Recall that, by definition, the call center time is the duration from the moment when a incoming request enters the system until the moment the ambulance has been dispatched by a dispatcher. Follow-up contacts and the additional instructions are not contained in the call center time, but they are, however, part of the work-load of an agent.

Our goal is to gain insight into the required number call takers $n_{calltaker} \in \mathbb{N}$ and dispatchers $n_{dispatcher} \in \mathbb{N}$ in the model with function differentiation, or generalists $n_{generalist} \in \mathbb{N}$ in the model with solely generalists, to reach certain performance indicator thresholds. In the mixed model we use all these variables. Let n denote the total number of agents in the system.

Inspired by the Dutch quality of service requirements, these being applicable to most other Western countries, we limit ourselves to the following performance indicators for all models in the remainder of this paper, under the assumption that there will be no abandonments:

1. The fraction α_1 of the calls that are picked up by the call taker (or generalist) within at most r_1 time units.
2. The fraction $\alpha_2(u)$, $u \in \mathscr{U}$ of the honored calls that have a call center time at most $r_2(u)$ time units.
3. The average work-load on a call center agent, if applicable split by call center agent class.

2.2 EMS Call Center Model with Function Differentiation

Our first model describes the communication processes at an EMS call center with function differentiation in high detail; a schematic for this model is displayed in Figure 2. Inspired by the literature, the EMS call center is modeled as a queuing model; it encapsulates all communication moments and delays in queuing systems, in which call center agents are the servers and communication moments are the tasks. Using this approach,

each domain can be modeled as a queueing system with priority queues and a server pool, conditional routing statements, and assignment blocks.[2]

Priority Queues at the Two Queuing Systems The priority queues in our models all have a similar behavior. All tasks are non-impatient, meaning that, once they have entered the system, they wait until they are served. All priority queues have an infinite capacity and no overflow regulations. Tasks are handled on a preemptive priority first-in first-served basis by a fixed number of servers. This means that when a task enters the highest priority queue and finds all servers busy, a task of the lowest priority in service is put on hold and the respective server starts serving the task from the highest priority queue. This task resumes service from the point at which it was put on hold when a server becomes available and stays in service with a duration of the remaining service time, but only if there are no tasks to serve with a higher priority. As a result, higher priority tasks are not influenced by the presence of lower priority tasks.

Call Taking Domain Requests from applicant classes whose requests have a similar behavior are bundled into one incoming stream. We allow an incoming stream to be filled by only one applicant class.

Inbound requests are modeled as new tasks for the queuing system of the call taking domain. The request arrival process is assumed to be dependent on time and the applicant class.

Dependent on the incoming stream of a tasks, the call is routed to one of the priority queues at the call taker. Tasks originating from the same incoming stream are led to the same priority queue of the call taker queuing system.

The call taker queuing system has a fixed number of priority queues $M_1 \in \mathbb{N}$, denoted by $Q_1^C, Q_2^C, \ldots, Q_{M_1}^C$. The priorities are strictly decreasing, i.e., Q_1^C is the priority queue with calls of the highest priority. These

[2] Technically, the dispatcher domain contains multiple queuing systems, since each block has one, as we show later on.

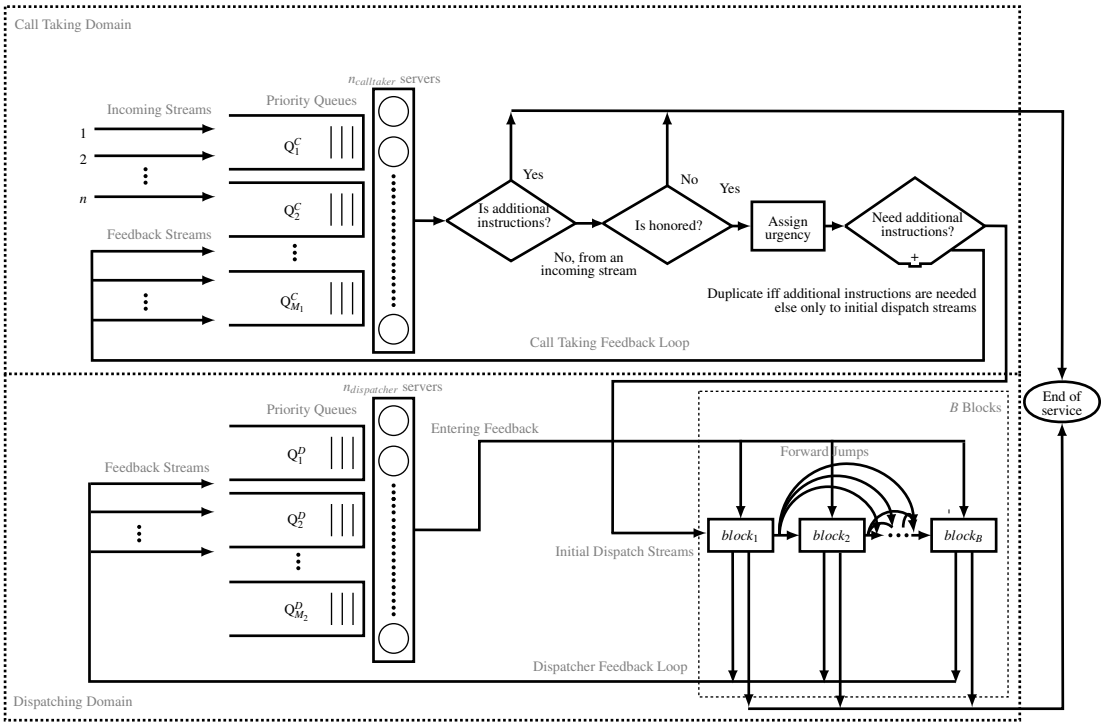

Fig. 2 Queuing model representing the call taker and dispatching domains of the EMS call center.

priority queues are also filled by a feedback loop containing the extra communication moments in the case the call taker gives additional instructions.

When a server is available the task immediately receives service. The service-time distribution depends on the applicant class and the urgency. In practice the urgency is assigned during the service, and the call taking duration depends on that urgency. Therefore we actually calculate the urgency at the start of the service, though in the schematic we have not found a way to clearly denote it and we choose to display it at service completion.

The first time a task arrives at the *'Is additional instructions?'* statement, the answer is no. In the next routing statement, either with probability qh the call is not honored and the request ends service. Otherwise, with probability ph an ambulance is to be sent to the incident. In the latter case the request gets an urgency $u \in \mathscr{U}$ assigned and we say that a request is honored. The probability $ph = 1 - qh$ depends on the incoming stream.

With probability $pe(i,u)$ the call taker gives the applicant additional instructions; see the *'Need additional instructions?'* conditional fork in Figure 2.

When the applicant receives additional instructions in the model, the task splits into two separate tasks. One goes to the priority queues at the call takers using a feedback stream, while the other directly goes to the dispatcher. Using this conditional fork both a call taker and dispatcher can work simultaneously. If the priority queue at the dispatcher is of the highest priority Q_1^C, it acts like an uninterrupted call at the call taker. An additional instruction task ends service when the agent puts down the telephone.

If there are no additional instructions required, the task directly enters the dispatcher domain. This happens through an initial dispatch stream and $Block_1$ to one of the priority queues of the dispatcher. The routing to the priority queue at the queuing system in the dispatching domain depends on the urgency of the request.

Dispatching Domain Task handling at the dispatcher server pool is done by a $M/G/n_{dispatcher}/\infty/PNPN$ queuing policy, just like at the call taker server pool. The $M_2 \in \mathbb{N}$ priority queues $Q_1^D, Q_2^D, \ldots, Q_{M_2}^D$ are filled by honored requests of the call taker and communication moments with ambulance and hospital, represented by blocks. The service-time distribution depends on the call's urgency, incoming stream, and, if applicable, the block it originates from.

The model contains a sequence of blocks representing the inter arrival times, contact probabilities, routing, and the service time of follow-up contact moments with ambulance teams of the dispatching domain.

A block starts with a period of waiting, possible contact with the dispatcher in a feedback contact, and continuation to a succeeding block or the end of service. Let us give a more detailed description of the block sequence.

2.3 Block Description

All models have a *block sequence* that consists of a fixed number of $B \in \mathbb{N}_{\geq 1}$ blocks; a schematic of a block is displayed in Figure 3.

The number of priority queues open to receive feedback from the blocks is denoted by M. Note that $M = M_2$ for the model with function differentiation, and a similar behavior can be found for the other two models.

A special case is $Block_1$. This block is always present and always leads to a feedback; this represents the dispatch of an ambulance where the priority and service time of the dispatch equal the priority and service time by the assigned server in the dispatch server pool.

Without loss of generality we describe the structure of $Block_b$, $b \in \{1,2,\ldots,B\}$, and we illustrate its behavior for a request with urgency $u \in \mathscr{U}$. The incoming tasks originate from previous blocks (for $b > 1$) or the newly honored calls that enter the dispatch domain (for $b = 1$). To mimic the behavior of a time interval in which no communication occurs, tasks are handled by an infinite-server pool with general distributed service time. This service time can be interpreted as a delay by a driving time or contact moment with a patient at the incident location or hospital.

At this point in time the decision has to be made whether the ambulance team and call center agent have a contact moment using priority queue $m \in \{Q_1^D, Q_1^D, \ldots, Q_{M_2}^D\}$. With probability $qf_b^m(u)$ a contact occurs

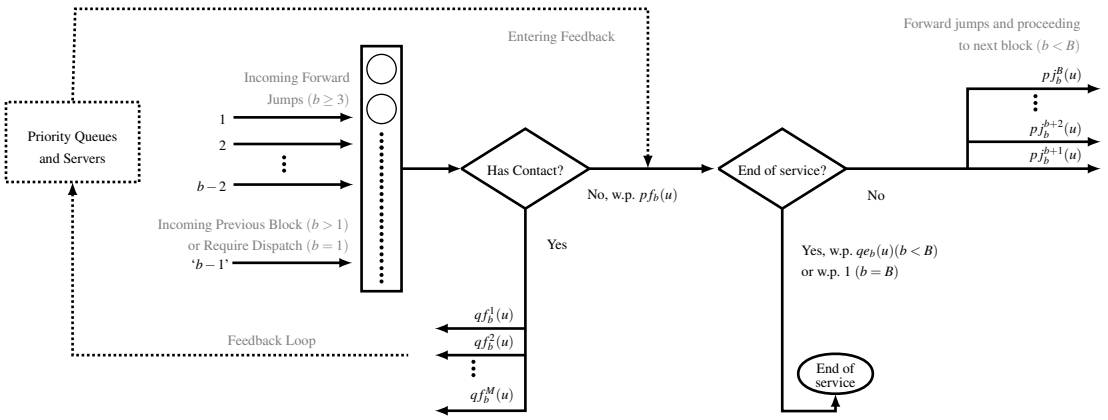

Fig. 3 Block description.

through the feedback loop mechanism to priority queue Q_m, and with probability $pf_b(u)$ no contact occurs in this block and the call moves forward to the next decision moment. The service time is generally distributed, and the parameters may depend on b and u. Note that $pf_b(u) + \sum_{1 \leq m \leq M} qf_b^m(u) = 1, \ \forall\, b,u$.

Right after service completion, the feedback task enters $Block_b$ again. Both the tasks from the feedback loop and tasks that had no feedback continue to the next decision moment. Now we determine whether there is an end of service, with probability $qe_b(u) = 1 - pe_b(u)$, or alternatively if the task moves to a succeeding block. If $b = B$ we take $qe_b(u) = 1$; this leads to an end of service. The last decision moment of $Block_b, b \neq B$ tells us at which block the task continues, the so-called forward jump. With probability $pj_b^{b'}(u)$ we redirect to $Block_{b'}$. For $b' \leq b$ we have $pj_b^{b'}(u) = 0$, and $\sum_{b'>b} pj_b^{b'}(u) = 1 \ \forall\ b \neq B,u$. Notice that "End of Service" can also be interpreted as a special case of a forward jump to a dummy block $B+1$.

2.4 EMS Call Center Model with Solely Generalists

In this section we describe the EMS call center model with solely generalists. Figure 4 provides a schematic overview of this model.

There is a major difference between this model and the model with function differentiation of Section 2.2. In this model there is only one finite server pool with agents, thus the feedback from blocks and additional instructions is mixed with the incoming demand. In this section we build upon the assumptions of Section 2.1 and the blocks from Section 2.3.

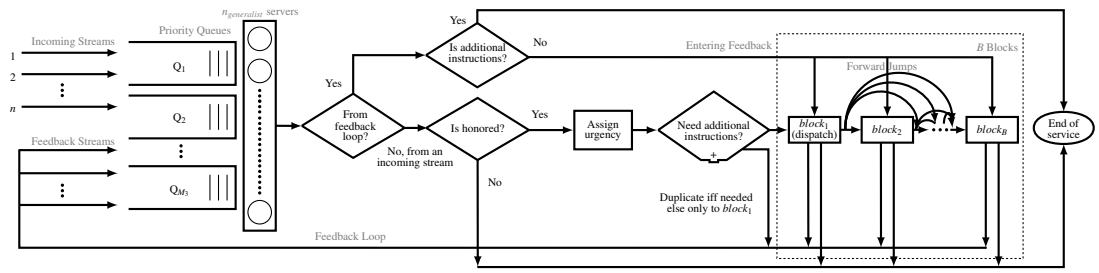

Fig. 4 Model with solely generalists.

A new request generates an incoming task from an incoming stream into one of the $M_3 \in \mathbb{N}$ priority queues for the server pool with $n_{generalist} \in \mathbb{N}$ generalists. The priority queue assignment is based upon the incoming stream and temporary hidden urgency of the request, like in the model with function differentiation. At service completion of the call taking process, the call has probability $ph(i)$ to be honored by the call center agent. An urgency $u \in \mathscr{U}$ gets assigned to each honored request, based on a categorical distribution. Its parameters depend only on the incoming stream. Not honored requests exit the system. Directly after the urgency assignment, there is a conditional fork that generates an extra task in the case there is an additional instructions contact. Whether this is the case is determined by a Bernoulli distribution that depends on the incoming stream.

Further distribution parameters depend on the urgency, the incoming stream and the block last visited. The extra task is led to one of the priority queues with the highest priority. This additional instruction task exits the system at service completion. The other task leaves the conditional fork to $Block_1$, which acts like the dispatch.

Like the call taker queue in the model with function differentiation, this is a priority based preemptive FCFS queue with an infinite capacity. A block sequence mimics the behavior of the dispatch, driving to an incident, taking care of the patient at the incident, et cetera. The distribution can be determined using a best fit approach. After the last block the call is ended, but it is also possible that another block already ended the call.

2.5 Mixed Model

The last model, the so-called *mixed model*, contains call takers, dispatchers and generalists. It is very similar to the model with function differentiation; see Section 2.2. The major difference between the two models is that there is an extra server pool with $n_{generalist}$ generalists. A generalist starts to provide service to a task when there are no call takers or dispatchers available to take it, i.e., it handles the overflow from both the call takers and the dispatchers. Generalists respect priority of jobs like any other call center agent type. If an agent of another type becomes available, the generalist finishes the task that he services instead of giving the remaining service to the available centralist.

3 Data and Parameter Estimations

In this section we discuss the input data and parameter estimations. We use two data sources of the EMS call center in the city of Utrecht in the Netherlands over a period of three months. The first is the telephone information (TI) data set, that originates from a server that monitors all in- and outbound telephone calls. The second contains call record details of the ambulance services in this period, which include status updates from the ambulance teams. Every inbound request to the ambulance service provider has its own row in the ambulance services call record details data set. Parameters that could not be determined from the data sets are estimated through expert opinion.

We group the input parameters as Bernouilli distributions and service-time distributions. The parameters that are discussed in Section 3.1 to Section 3.5 are the same for both models.

First we describe the data sets that we used as an input source more intensively. Section 3.2 explains our aggregation to reduce multiple applicant classes into incoming streams. Sections 3.3 and 3.4 describe the Bernoulli decision parameters for the call taking and dispatching domains, respectively. Section 3.5 gives the service-time distributions for all call taking and dispatching processes. Finally, Section 3.6 discusses the routing policies toward the priority queues for both models separately.

3.1 Processing the Data Sets

The TI data set contains the timestamps of the telephone communication, i.e., the moment the EMS call center agent lifted the handset, when it was hung up, and if it was an in- or outbound call. This data set consists of 109,000 inbound and outbound telephone calls, of which 79% are inbound. The TI data set does not include any information on the content of the calls, such as the applicant class (e.g. civilian or police), and whether the

Table 2 Distribution of new requests over the incoming streams.

Incoming Stream	Applicant Classes	Percentage
Civilian	1-1-2, 9-1-1	24.4
Hospital - Low	Hospitals*	17.4
Others	Unspecified, Others	15.3
EMS call center	EMS Call Center Agents	13.9
GP - Low	GPs*, GP Centers*	11.7
GP - High	GPs*, GP Centers*	10.2
HC Institutions	Psychi'c, Midwifes, Homecare	3.3
Police	Police, Fire Fighters	2.6
Hospital - High	Hospitals*	1.3

Differences between starred entries are addressed in Section 3.5.

request was honored. The records do not indicate if the call was a new incoming call or a follow-up call by an ambulance team.

Additional information was linked to the TI records by matching it to the database of the call record details. In this time period there were 41,000 requests of ambulance care. The matching is done probabilistically, in the sense that we coupled the start of the ambulance service request to the most likely corresponding telephone contact in the TI data set. In the matching process, usually one inbound call was matched to one request and, thereby, to one service. In some cases an inbound call was matched to multiple services; we omitted these multiple matches in the estimation of the parameters. Fraction 92% of the call record details were matched to an TI record. From these matched data records we identified applicant classes, urgencies, priorities, and service times. Unmatched TI records were classified as follow-up calls.

3.2 Aggregation to Incoming Streams

Every applicant class is mapped onto one incoming stream. For example, the incoming stream *Health Care Institutions* contains every request from the applicant classes *Psychiatric, Midwifes,* and *Home Care*. This grouping is based on applicant classes with similar substantive grounds, service-time distributions and priorities. Table 2 lists the call volume per incoming stream as percentages of the total call volume; these percentages are obtained from the call center records. It also shows which applicant classes are contained in each stream.

We assume the arrival process for applicant class k to be a Poisson process with rate f_k. Then the arrival process for an incoming stream can be modeled as Poisson arrivals with rate

$$\lambda_i = \sum_{\substack{k \text{ is included} \\ \text{in stream } i}} f_k, \qquad i \in \{1,\ldots,n\}$$

for a fixed number of incoming streams $n \in \mathbb{N}$, where applicant k is included in incoming stream i.

The call flow matches the Dutch practice. An underlying assumption is that civilians have inadequate knowledge of what to do in emergency situations and GPs can determine very well when they require ambulance services on short notice, yielding the result that the two streams are answered directly. A call from these classes even may force ongoing calls of a lower priority 'on hold'. All other incoming callers have some basic knowledge on how to keep the patient stable. Therefore, they have a somewhat lower urgency, and they are routed to the medium priority queue. The low priority calls are made from hospitals, and often contain a request for patient transfer to another hospital or house address.

3.3 Bernoulli Parameter Estimations for the Call Taking Domain

The call taking domain holds Bernoulli distributions for a request being honored, call urgency assignment, and receiving additional instructions.

Table 3 The urgency distribution (in %) for every incoming stream.

Incoming Stream	Urgency		
	High	Medium	Low
Civilian	60.1	37.8	2.1
Hospital - Low	0	0	100
Others	20.1	64.6	15.3
EMS call center	5.1	84.1	10.8
GP - Low	0	23.7	76.3
GP - High	72.6	27.4	0
Health Care Institutions	7.1	17.7	75.2
Police	48.7	49.7	1.6
Hospital - High	38.9	61.1	0

Only civilian requests qualify for not being honored; this happens with probability 22%, a value obtained from the call records data set. Every honored request gets an urgency assigned. We use three urgencies: $(u_1, u_2, u_3) = (\text{High, Medium, Low})$. Table 3 shows the urgency distribution for each incoming stream.

With probability 10% a request from a civilian gets additional instructions, which is independent of the urgency. This value is obtained from expert guesses.

3.4 Bernouilli Parameter Estimations for the Dispatching Domain

The Bernouilli distributions at the dispatching domain are structured in blocks: for each block these parameter consist out of the follow-up contact, end of service, and forward jump probabilities. Let us describe these parameters for the $B = 6$ blocks that we use in our simulation runs.

The probabilities for a follow-up contact are listed in Table 4; they are all based on expert opinion, and they are addressed for each block individually in high detail. The probability distribution for the infinite server agents for each block is obtained from the service detail records; these are the durations for the travel times, treatment duration, transfer duration at the hospital, etc. We fitted a log-normal, normal, and exponential distribution and used the mean squared error to determine a best fit. These distributions depend only on the urgency and status, and in particular not on the incoming stream. Another choice one could have made is to use an empirical distribution. The best fit distributions and their parameter values from our data set are listed in Table 5.

Forward jumps only occur from $Block_4$: *During Patient Treatment* to $Block_6$: *Transfer at the hospital*; that is when a patient does not require transport to a hospital. The probability that a forward jump $pj_4^6(u)$ occurs is 41.87% for high urgency, 40.34% for medium urgency, and 15.29% for low urgency requests. We only use a redirect to *End of Service* in the last block.

We look at each of the blocks individually to set the remaining routing parameters. Parameters that are not mentioned explicitly are zero valued.

$Block_1$: *Dispatch* This block includes the dispatch process of choosing the right vehicle for every honored request, and the contact moment from the EMS call center to the ambulance team to give initial instructions. Also time for optimizing the coverage of the fleet using a dynamic ambulance management (DAM)

Table 4 Probability $\sum_{i=1}^{M} qf_b^i(u)$ (in %) that a follow-up contact occurs.

Block	Status	Urgency (in %)		
		High	Medium	Low
$Block_1$	Dispatch	100	100	100
$Block_2$	Leaving the Base Location	15	15	15
$Block_3$	Driving to Patient	30	10	10
$Block_4$	During Patient Treatment	10	0	0
$Block_5$	Driving to Hospital	0	0	0
$Block_6$	Patient Transfer at Hospital	100	100	100

Table 5 Service time distribution for the infinite-server pool of every block. Units are in seconds, or min:sec.

Block	Status	Urgency	Distribution	μ	σ	Mean	SD
$Block_1$	Dispatch	H/M/L	Deterministic	0	*N.A.*	00:00	00:00
$Block_2$	Leaving the Base Location	High	Log-normal	3.926	0.624	01:02	00:42
	Chute Time	Medium	Log-normal	4.244	0.717	01:30	01:14
		Low	Exponential	145.280	*N.A.*	02:25	02:25
$Block_3$	Driving to Patient	High	Log-normal	5.836	0.482	06:25	03:16
		Medium	Normal	622.342	300.994	10:22	05:01
		Low	Normal	734.297	444.314	12:14	07:24
$Block_4$	During Patient Treatment	High	Log-normal	7.102	0.471	22:55	11:16
		Medium	Log-normal	6.799	0.683	18:53	14:33
		Low	Log-normal	6.783	0.480	16:30	08:53
$Block_5$	Driving to Hospital	High	Log-normal	6.518	0.549	13:07	07:46
		Medium	Normal	847.788	411.833	14:08	06:52
		Low	Log-normal	6.886	0.536	18:50	10:52
$Block_6$	Patient Transfer at Hospital	High	Log-normal	6.917	0.491	18:59	09:55
		Medium	Normal	975.745	439.378	16:16	07:19
		Low	Normal	955.944	470.430	15:56	07:50

methodology is included in this block. Giving the initial instructions often is done digitally by sending a notification to the team's pagers. Note that the block's infinite-server pool has a zero service time for all calls, because there is no time delay between call taking and dispatching other than the queues of the dispatcher pool. There is a feedback loop which directs to the dispatch-queue whose name is similar to the calls urgency $qf_1^{D:High}(High) = qf_1^{D:Medium}(Medium) = qf_1^{D:Low}(Low) = 1$; see Section 3.6. On the assumption that enough ambulances are available, a request cannot end at this stage: $pe_1(u) = 1$. When an ambulance is already on the road, there is no departure from base, but the EMS team has motivation for similar questions to the EMS call center and $Block_2$ is not skipped. Notice that in this case its block name is not fully accurate, although the behavior can be included in the service-time distribution for the infinite-server pool in $Block_2$. There is always a redirect to $Block_2$: $pj_1^2(u) = 1\ \forall u \in \mathscr{U}$.

$Block_2$: *Leaving the Base Location* When an ambulance team reads the incident description on the on-board monitor, there may be pressing questions as regards medical uncertainties or special equipment that must be taken on board. Another reason for contact is that the incident location might be unclear to the EMS team. Under the assumption that a request is not canceled at this stage, the request is passed to $Block_3$: $pe_2(u) = 1, pj_2^3(u) = 1\ \forall u \in \mathscr{U}$.

$Block_3$: *Driving to Patient* When the ambulance arrives at the incident location, there can be a contact moment with the EMS call center for various reasons: the EMS team can find the patient or there is a request for assistance by another team. In most cases the EMS call center specifies the location in more detail. Because the EMS team is assumed to search for the patient or start treating, there is a forward jump to $Block_4$: $pj_3^4(u) = 1\ \forall u \in \mathscr{U}$.

$Block_4$: *During Patient Treatment* Depending on the findings at the incident locations, there are multiple possible outcomes. When a patient is treated and needs transportation to a hospital, the crew can contact the EMS call center to ask them to notify the hospital's emergency department. When a patient is not found, not yet ready for transportation, or can be treated at the incident location, the crew becomes available again. In that case they can give a situation update to the EMS call center and go to a base location to wait for a new request being assigned to them. When a patient needs transport to a hospital or other destination we redirect to $Block_5$. If a patient is treated at the incident location, forward jumps occur with probabilities $pj_4^6(u)$ as discussed earlier.

$Block_5$: *Driving to Hospital* It is unlikely that a contact occurs on arrival at a hospital. Ambulance providers whose EMS call center agents provide extra motivation to the EMS team to become available again on short notice may include those contact moments in this block. We include this stage as a delay, although it could be merged with $Block_6$ in our case. The only non-zero parameters are $pe_5(u) = 1$ and $pf_5(u) = 1\ \forall u \in \mathscr{U}$.

Block$_6$: *Transfer at the Hospital.* At the hospital the patient is transferred to another health care provider, and the EMS team take a few minutes to refresh themselves. Ambulances in some EMS regions refill medical materials at the hospital. When the ambulance has delivered the patient, the EMS call center is notified that they are available for dispatch again. Since this is the last block in line, it results by definition in an *End of Service*: $qe_5 = 1$.

3.5 Service Time Distributions

Call Taking Domain Motivated by literature [5, 8, 28], we assume that the service time at the call taker is log-normal for each incoming stream and urgency couple. For both models we take the same service-time distributions. The parameter values for every incoming stream and priority couple are listed in Table 6.

Hospitals and GPs each have two separate lines to reach the EMS call center and are able to prioritize their request using these lines. For hospitals we assumed that the high urgency line is used if and only if it leads to a high urgency call. For GPs we assumed that high and low urgencies are from the high priority and low priory lines, respectively. For the medium urgency we assumed that the calls were evenly distributed over the high and medium urgency lines.

The service-time distribution for the additional instructions equals the chute time of high urgency calls, because we have no data to support a better assumption. The chute time is the time it takes the EMS team to leave the base location. We took the urgency dependent chute time distribution as a proxy for the dispatch time distribution, based on the motivation that a high urgency call contains more critical information.

Table 6 Log-normal service-time distribution for the EMS call center agents. Units are in seconds, or min:sec.

Incoming Stream	Priority	μ	σ	Mean	SD
	High	4.604	0.695	2:07	1:40
Civilian	Medium	4.579	0.697	2:04	1:38
	Low	4.689	0.616	2:11	1:29
	High	4.432	0.761	1:52	1:39
Police	Medium	4.417	0.689	1:45	1:22
	Low	4.744	0.496	2:10	1:09
	High	4.559	0.746	2:06	1:49
Others	Medium	4.645	0.667	2:10	1:37
	Low	4.620	0.606	2:02	1:21
Health	High	4.518	0.538	1:46	1:01
Care Inst.	Medium	4.821	0.502	2:21	1:15
	Low	4.615	0.599	2:01	1:19
	High	4.471	0.582	1:44	1:06
Hospital - High	Medium	4.695	0.500	2:04	1:06
	Low	*N.A.*	*N.A.*	*N.A.*	*N.A.*
	High	4.538	0.579	1:50	1:10
GP - High	Medium	4.744	0.390	2:04	0:50
	Low	*N.A.*	*N.A.*	*N.A.*	*N.A.*
	High	*N.A.*	*N.A.*	*N.A.*	*N.A.*
GP - Low	Medium	4.744	0.390	2:04	0:50
	Low	4.715	0.540	2:09	1:15
	High	*N.A.*	*N.A.*	*N.A.*	*N.A.*
Hospital - Low	Medium	*N.A.*	*N.A.*	*N.A.*	*N.A.*
	Low	4.644	0.605	2:05	1:23
	High	4.471	0.737	1:55	1:37
EMS Call Center	Medium	4.626	0.687	2:09	1:40
	Low	4.579	0.657	2:01	1:29
Exit Information	All	3.926	0.624	1:02	0:43
	High	3.926	0.624	1:02	0:43
Dispatch	Medium	4.244	0.717	1:30	1:14
	Low	3.988	0.570	1:03	0:39
Follow-up call	Medium	3.389	0.846	0:42	0:43

Table 7 Priority Queues for EMS call center agents, model with function differentiation.

Call Taking Domain		
Priority Queue	Name	Task originates From
Q_1^C	C: Ultra High	Exit information
Q_2^C	C: High	Incoming streams: Civilian, GP - High
Q_3^C	C: Medium	Incoming streams: Police, Others, Health Care Institutions, Hospital - High, GP - Low, EMS call center.
Q_4^C	C: Low	Incoming stream: Hospital - Low.
Dispatching Domain		
Priority Queue	Name	Task originates From
Q_1^D	D: High	High Urgency, from $Block_1 : Dispatch$.
Q_2^D	D: Medium	Medium Urgency, from $Block_1 : Dispatch$. All feedback from $Block_2$ up to and including $Block_6$.
Q_3^D	D: Low	Low Urgency, from $Block_1 : Dispatch$.

Dispatching Domain The service-time distribution for the dispatch time is assumed to equal the service-time distributions of the chute time for high and medium urgency calls. Because the dispatch of low urgency calls require less time and are assumed to be log-normally distributed, we have omitted outliers that are over three standard deviations above the mean, and made a best fit log-normal distribution.

The service-time distribution of the dispatcher's follow-up contacts is obtained from unmatched TI records. Because we were unable to distinguish between the status in the process, priority or urgency, we used the same best fit log-normal distribution for every follow-up contact with a mean of 42 seconds.

3.6 Routing Policies toward the Priority Queues

Regardless of the model, all agent pools handle tasks on a priority based first-in first-served policy. This means that tasks with a higher priority are handled first, and for tasks with the same priority the call takers handle them at a first-come first-served base. Notice that the infinite-server pools in the blocks have no queues and priorities because there are enough agents to start any incoming service directly.

Model With Function Differentiation The priority queues in the model with function differentiation is listed in Table 7. In the call taking domain the civilian calls and high urgency general practitioners calls are of high priority. Giving their exit instructions a slightly higher priority leads to an uninterrupted call: these exit instruction tasks are directly picked up by a call taker after finishing the triage process and the number of exit instructions cannot exceed the number of call takers, thus all exit instruction have a zero waiting time.

In the dispatching domain, the dispatch of high urgency calls have the highest priority, and the dispatches of low urgency, often ordered patient transports, are done when there are no remaining tasks left. Relocations are considered part of the dispatch procedure.

Model With Solely Generalists Table 8 gives the priority queues of the model with solely generalists: the priority queues of the model with function differentiation are zipped together. The main idea is that dispatching an ambulance to a request with a certain urgency is more important than taking a call with compatible priority.

Table 8 Priority Queues for the model with solely generalists.

Priority Queue	Name	Task originates from
Q_1	C: Ultra High	Exit information
Q_2	D: High	High Urgency, from $Block_1 : Dispatch$.
Q_3	C: High	Incoming streams: Civilian, GP - High
Q_4	D: Medium	Medium Urgency, from $Block_1 : Dispatch$. All feedback from $Block_2$ to $Block_6$.
Q_5	C: Medium	Incoming streams: Police, Others, Health Care Institutions, Hospital - High, GP - Low, EMS call center.
Q_6	D: Low	Low Urgency, from $Block_1 : Dispatch$.
Q_7	C: Low	Incoming stream: Hospital - Low.

In fact, communication with a team that may need additional assistance may be more important than taking a new request from the police.

4 Simulations and Results

To assess performance and to gain insight in the optimal staffing decisions in EMS call centers, we have performed extensive simulation experiments based on the models with and without function differentiation described above. Recall that the KPIs are:

1. the waiting time before the call is taken, for high, medium and low priority classes,
2. the average call center time, i.e., the time duration from call entering the call taker queue until a call is dispatched, and
3. the average work-load for a call center agent, for call takes, dispatchers and generalists.

Table 9 Simulation results for the optimal policies for function differentiation (FD) and solely generalists (SG), for various rates.

λ	Best Model	$\underline{\mathbf{n}}$	n	Costs (in k€)	Fraction Within 6 Sec (in %)			Call Center Time (in min:sec)			Busy Fraction CT Gen	Busy Fraction Disp (in %)
					High	Medium	Low	High	Medium	Low		
100	FD	(1, 1, 0)	2	125	96.3	87.4	85.9	3:09	3:22	3:45	14.5	11.1
		(0, 0, 2)	2	180	99.8	98.1	96.9	2:59	3:04	3:15		12.8
200	SG	(0, 0, 2)	2	180	99.4	92.9	89.8	3:02	3:09	3:35		25.5
		(2, 1, 0)	3	195	99.7	97.2	96.3	3:01	3:10	3:42	14.5	22.1
300	SG	(0, 0, 2)	2	180	98.7	85.6	80.2	3:06	3:17	4:13		38.3
		(2, 1, 0)	3	195	99.3	94.2	92.5	3:03	3:16	4:12	21.7	33.2
400	FD	(2, 1, 0)	3	195	98.9	90.2	87.8	3:07	3:24	4:54	28.9	44.3
		(0, 0, 3)	3	270	99.8	94.4	91.5	3:00	3:06	3:29		34.0
500	FD	(2, 2, 0)	4	250	98.2	85.3	82.0	3:07	3:19	3:47	36.2	27.7
		(0, 0, 3)	3	270	99.7	90.4	84.9	3:02	3:09	3:50		42.6
600	FD	(2, 2, 0)	4	250	97.6	79.7	75.6	3:12	3:27	4:09	43.4	33.2
		(0, 0, 3)	3	270	99.4	85.3	77.0	3:04	3:14	4:22		51.1
700	FD	(2, 2, 0)	4	250	96.8	73.7	68.0	3:17	3:37	4:37	50.6	38.7
		(0, 0, 4)	4	360	99.90	93.6	88.8	3:00	3:06	3:36		44.7
800	FD	(3, 2, 0)	5	320	99.5	90.9	88.7	3:02	3:12	3:58	38.6	44.2
		(0, 0, 4)	4	360	99.8	90.4	83.2	3:01	3:08	3:53		51.0
900	FD	(3, 2, 0)	5	320	99.4	88.0	84.8	3:03	3:15	4:16	43.4	49.8
		(0, 0, 4)	4	360	99.8	86.4	77.0	3:02	3:11	4:19		57.4
1,000	FD	(3, 2, 0)	5	320	99.2	84.8	80.3	3:05	3:18	4:38	48.2	55.2
		(0, 0, 4)	4	360	99.7	81.3	68.7	3:04	3:15	4:57		63.8
1,200	FD	(4, 3, 0)	7	445	99.90	92.7	89.7	3:01	3:08	3:37	43.4	44.2
		(0, 0, 5)	5	450	99.96	88.0	77.6	3:01	3:09	4:12		61.2
1,400	FD	(4, 3, 0)	7	445	99.8	88.7	84.4	3:02	3:11	3:56	50.6	51.6
		(0, 0, 6)	6	540	100	92.4	83.5	3:00	3:06	3:49		59.5
1,600	FD	(4, 3, 0)	7	445	99.6	83.6	77.0	3:05	3:16	4:26	57.8	58.9
		(0, 0, 6)	6	540	99.98	87.1	72.5	3:01	3:08	4:29		68.0
1,800	FD	(5, 3, 0)	8	515	99.90	91.7	87.3	3:01	3:09	4:35	52.0	66.3
		(0, 0, 7)	7	630	99.98	91.7	79.1	3:00	3:06	4:00		65.6
2,000	SG	(0, 0, 7)	7	630	99.98	88.0	69.4	3:01	3:08	4:44		72.8
		(6, 4, 0)	10	640	99.97	95.8	92.8	3:00	3:05	3:38	48.2	55.2
2,200	FD	(6, 4, 0)	10	640	99.96	94.0	89.2	3:00	3:07	3:53	53.0	60.7
		(0, 0, 8)	8	720	99.99	92.1	76.3	3:00	3:06	4:10		70.1
2,400	FD	(6, 4, 0)	10	640	99.95	91.6	84.9	3:01	3:08	4:14	57.7	66.2
		(0, 0, 8)	8	720	99.99	88.3	65.7	3:01	3:08	4:57		76.4
2,600	FD	(6, 4, 0)	10	640	99.95	88.7	80.1	3:02	3:11	4:45	62.5	71.7
		(0, 0, 9)	9	810	99.98	92.1	73.2	3:00	3:06	4:18		73.6
2,800	FD	(7, 5, 0)	12	765	99.97	93.8	87.6	3:00	3:06	3:46	57.7	61.8
		(0, 0, 10)	10	900	99.99	94.6	79.3	2:59	3:05	3:56		71.3
3,000	FD	(7, 5, 0)	12	765	99.96	91.6	84.2	3:01	3:08	4:00	61.8	66.1
		(0, 0, 10)	10	900	100	92.3	71.7	3:00	3:06	4:26		76.3

In general, the cost of hiring different agent types depend on the required level of eduction. The total yearly cost for call takes, dispatchers, and generalists is assumed to be €70k, €55k, and €90k, respectively; these numbers are representative for the Netherlands. For convenience, we denote the *staffing policy* by the triple

$$\underline{n} := (n_{calltaker}, n_{dispatcher}, n_{generalist}), \tag{1}$$

where $n_{calltaker}$, $n_{dispatcher}$, and $n_{generalist}$ are the numbers of call takers, dispatchers, and generalists, respectively.

Solely Generalists vs. Function Differentiation. Table 9 shows the results of extensive simulations for the models with solely generalists and function differentiation. For each arrival rate between 100 and 3,000, and for both models, we determined what the lowest cost and corresponding policy is for which the KPI requirements are met. For brevity of the table we only show results in steps of 200 after $\lambda = 1,000$, although we consider every λ that is a multiple of 100 in our analysis. The requirements used are that at least 95% of the calls must be taken within 6 seconds, and low urgency calls must have a call center time of at most 5 minutes. We do not pose a boundary on the busy fractions. We call a policy better than another one when it has a lower cost. Various interesting insights can be gained from this data.

First, there is not one model that outperforms the other for all rates. For the lowest arrival rate we see that we need two agents, though we can do it at a lower cost with function differentiation. For $\lambda = 300$ two generalists can reach the required performance. From that point, the lowest cost model with function differentiation is slightly cheaper than the model with solely generalists. The only exception can be found at $\lambda = 2,000$, where the two models are nearly equal, since solely generalists are €10k cheaper.

Second, the latencies for high priority calls are exceeding 99.5% in nearly all of the cases, especially for higher arrival rate values λ. For lower and medium priority calls we can see a clear, though not completely unanimous, difference in favor for the model with function differentiation.

Third, busy fractions of call center agents do not exceed 76.4%. In general one can say that generalists have a higher work-load than call center agents in the model with function differentiation. The latter model has at most two call center agents more, which helps to explain this difference.

Fourth, we notice that the best policy with function differentiation in the cases considered has at least the same number of EMS call center agents as the model with solely generalists. Thus the number of work stations will not decrease when switching from a solely generalists policy to function differentiation.

We observe that good prioritizing of tasks by the EMScall center agents, in combination with a good KPI requirement for the low urgency and low priority calls, leads to good results for medium and high urgency and priority calls. This is due to the fact that KPIs of higher urgency calls are not affected by the lower priorities.

Mixed Model Table 10 shows the results of extensive simulations of mixed policies, combined with the other two models. More precisely, it shows the KPIs for those combinations $\underline{n}$, defined in (1), for which no additional agent (of any type) can be hired within the budget constraint, and for which also the system is stable. We call a model stable when every queue length stays within bounds. In the simulations, the total annual budget for hiring agents is 500k€. The table shows an KPI for generalists, additionally to the three already mentioned:

4. The work distribution, i.e., the average percentage of time that the generalists spend taking calls versus dispatching.

To gain insight into the implications of staffing decisions on the KPIs, we have performed extensive simulation studies. The results are briefly outlined below. In the examples discussed below the call arrival rate was taken to be $\lambda = 2,000$, i.e. a rate of 83.33 requests per hour, and the service-time distributions were taken from Table 6. The results lead to a number of interesting observations.

First, comparing the results for the cases (0, 4, 3) and (4, 2, 1), it is quite remarkable that we can observe that the fraction of calls that meets the 6-second target is higher for *all* priority classes in the case (4, 2, 1). This seems rather counter-intuitive, since it would be natural to say that swapping agents between different classes would favor at most one or two classes at the expense of another class. Note also that the total cost for the case (4, 2, 1) is 480, which is less than the cost of 490 for the case (0, 4, 3). Looking at the call center times, we also

Table 10 Simulation results for the mixed policies.

n	n	Fraction Within 6 Sec (in %)			Call Center Time (in min:sec)			Busy Fraction (in %)			Work Distribution Gen. (in %)	
		High	Medium	Low	High	Medium	Low	CT	Disp	Gen	Call Taking	Dispatching
(0, 4, 3)	7	95.8	29.9	2.5	05:02	07:01	33:51	*N.A.*	53.5	98.7	98	2
(0, 2, 4)	6	99.3	57.9	27.0	03:24	03:50	10:27	*N.A.*	77.3	88.8	81	19
(1, 4, 2)	7	95.8	30.3	3.3	05:03	07:08	22:34	98.9	53.6	98.2	97	3
(1, 2, 3)	6	99.3	58.9	28.8	03:24	03:51	10:21	91.2	78.4	87.3	76	24
(1, 1, 4)	6	99.9	70.6	39.1	03:10	03:26	09:37	88.5	84.1	84.3	59	41
(2, 4, 1)	7	95.8	30.6	4.8	05:05	07:06	06:02	98.2	54.0	97.3	95	5
(2, 3, 2)	7	99.3	66.8	46.5	03:18	03:40	05:56	83.7	64.1	75.2	81	19
(2, 1, 3)	6	99.9	72.4	42.5	03:10	03:26	10:56	84.7	86.7	84.6	47	53
(2, 0, 4)	6	99.9	77.2	46.6	03:06	03:20	11:26	83.7	*N.A.*	85.7	36	64
(3, 5, 0)	8	95.8	31.2	8.0	05:07	06:57	42:30	96.3	44.2	*N.A.*	*N.A.*	*N.A.*
(3, 3, 1)	7	99.3	68.0	49.9	03:18	03:39	05:57	79.1	66.4	73.4	70	30
(3, 2, 2)	7	99.9	82.2	66.0	03:05	03:16	05:30	73.9	75.0	69.1	49	51
(3, 0, 3)	6	99.9	78.7	51.0	03:07	03:21	24:08	77.7	*N.A.*	92.3	20	80
(4, 4, 0)	8	99.3	70.7	57.3	03:15	03:35	05:02	72.2	55.2	*N.A.*	*N.A.*	*N.A.*
(4, 2, 1)	7	99.9	83.9	71.1	03:04	03:16	07:33	67.0	82.3	77.2	27	73

see that the case (4, 2, 1) outperforms (0, 4, 3) for all urgency classes. To understand why this is the case, we observe from the work distribution results that in the case (0, 4, 3) the generalists are heavily loaded (in fact, 98.7% utilization), of which 98% is spent on call taking and only 2% on dispatching, whereas in the (4, 2, 1) case the work-load is much more balanced.

Second, the results show that in none of the cases considered only generalists were chosen, whereas intuitively it would make sense to do so, because of the facts that a generalist is flexible and hiring a generalist is less expensive than hiring both a call taker and a dispatcher. Consider for example the case (0, 2, 4) and compare this with the case (0, 0, 5). In the former case, in which six persons are hired, the system is stable, whereas in the latter case the system is unstable.

Third, there is an important difference between general call centers and EMS call centers with respect to the utilization of specialists versus generalists. This is so because in general call centers generalists are expensive, and hence, call center planners will tend to maximize their busy fraction, because idle time of generalists is costly. On the contrary, in EMS call centers there is an incentive for planners to keep the utilization of generalists low, because the generalists' idle times can be used to support other EMS services. For example, compare the case (3, 3, 1) to the case (3, 2, 2). Their performances are similar, while in the case (3, 2, 2) the generalists are less heavily loaded.

Technicalities This study contains 2,676 unique simulations runs: 2,200 for a call center with function differentiations, 450 for a call center with solely generalists, and 26 for the mixed policy study. For each simulation run we simulated 50,000 incoming calls. If there were over a thousand calls in the system, that is the total of all queues, we said that the simulation is unstable and excluded it from further analysis.

The simulation engine is written in a C++ and MariaDB using the TIFAR-framework that we developed at CWI in Amsterdam. The correct working of the software was validated by intensively tracing individual tasks and call center agents. We validated the input distributions and parameters against the output database.

The simulations have run on a Calleo Application Server 2260 that is dedicated to this project. This machine contains two Intel Xeon Processor E5-2640v2 (8 cores, 20MB cache, 2.0GHz) and hyper-threading enabled, resulting in maximal 32 cores. The RAM memory equals 256GB at 1,600MHz. Running all 2,676 required simulations took 37,871 wall seconds (10h 31m), and during this time period on average 14.12 CPU's where in use simultaneously. TIFAR makes use of multi-threading and was limited to 16 cores maximal, as the database and operation system also needed processing power.

5 Conclusions and Discussion

In this paper we presented a performance and capacity simulation model for EMS call centers. The model includes two classes of EMS call center agents: call takers and dispatchers. A key feature of the model is that it includes follow-up calls from EMS teams and hospitals. The model also discriminates between multiple types of applicants which differ in priorities. The model enables EMS planners to better understand the impact of these features on the response time performance of EMS call centers.

We assume that the service-time distributions of both call takers and dispatchers are independent of the work-load. In reality, the time used for servicing a call may depend on the work-load and service time decreases when the work-load increases. The inclusion of work-load dependent service-time distributions may have a significant impact on the response time performance of EMS call centers.

The input originates from actual call center databases. There are some estimations in our results; they cannot directly be used for management decisions without further research. Probability distributions and parameters may differ for various ambulance regions. Our model has been made for general EMS call centers, but it can also be used for other applications, such as firefighter call centers.

Our model assumes that switching occurs instantaneously. However, in practice switching between tasks takes time, which is a motivation for the six-second response time threshold. Inclusion of non-negligible switching times is an interesting subject.

There are some secondary advantages and disadvantages in function differentiation. In an EMS call center with function differentiation a general assumption is that dispatchers work faster than generalists in doing the same tasks. This is because the full focus is on the logistic domain. An advantage of generalists is that they are easy to deploy in the case of sickness absence.

The models presented in this paper are also applicable to other call centers that take calls, dispatch and perform coordination, in situations where short response times are required. These include, but are not limited, to police, firefighters, taxi service, and roadside assistance. However, before proper use in another context the model's all the model and input paramaters should be adapted, amongst others, the arrival distributions, the number of urgencies, priorities, queus, and blocks.

6 Future Research

This paper entails interesting topics for further research.

One may consider an EMS call center as a multi-skilled call center. This means that there are people who can only take the easy calls, the junior call takers, from police and firefighters, and senior call takers who can take the life-threatening calls. We can even introduce a third call taker class, equipped to handle both.

Let us present some thoughts on extensions of the models. So-called 'ramping' occurs when the emergency departments of the hospitals are over capacity, which implies that ambulances have to wait in line to transfer a patient. Regions suffering from occasional ramping may want to include extra moments for contact with an EMS call center to cope with this load. We have not included ramping since this is not seen in the regions where we obtained validation data from.

We have not included the effect from EMS call center agents who do both dispatching and low priority call taking, which is a combination that also exists in practice.

It is interesting for future research to evaluate the use of these models in other, non-EMS, emergency call centers, and conclude if the claims we make also hold in these contexts.

As a final remark, note that the effect on the choice of triage protocols, which is also in focus in various parts of the world, is not considered in this study. This choice is part of the service-time distributions, and therefore it is considered as an input for our models.

Acknowledgements We thank Regionale Ambulance Voorziening Utrecht (RAVU) for sharing the EMS call center data. We thank numerous EMS drivers and nurses participating for their expert guesses. This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organization for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs.

References

1. Aboueljinane, L., Sahin, E., Jemaï, Z., Marty, J.C.: A simulation study to improve the performance of an emergency medical service: Application to the french val-de-marne department. Simulation Modelling Practice and Theory 47, 46–59 (2014)
2. Andersson, T., Petersson, S., Värbrand, P.: Dynamic ambulance relocation for a higher preparedness. In: Proceedings of 35th Annual Meeting of Decision Sciences Institute, Boston, MA (2004)
3. Aringhieri, R., Carello, G., Morale, D.: Ambulance location through optimization and simulation: the case of milano urban area (2007), unknown if it is published
4. Bhulai, S.: Dynamic routing policies for multi-skill call centers. Probability in the Engineering and Informational Sciences 23(1), 75–99 (2009)
5. Bolotin, V.A.: Telephone circuit holding time distribution. Proceedings of the 14th International Teletraffic Congress 1(1), 125–134 (1994)
6. Brotcorne, L., Laporte, G., Semet, F.: Ambulance location and relocation models. European Journal of Operational Research 147(3), 451–463 (2003)
7. Drezner, Z.: Dynamic facility location: the progressive p-median problem. Location Science 3(1), 1–7 (1995)
8. Dwars, R.P.: Capacity planning of emergency call centers. MSc thesis, VU University Amsterdam (2013)
9. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management 5(2), 79–141 (2003)
10. Henderson, S., Mason, A.: Ambulance service planning: simulation and data visualisation. Operations Research and Health Care pp. 77–102 (2005)
11. Hoogeveen, M.: Ambulance care in europe (January 2010)
12. Koole, G., Pot, A.: An overview of routing and staffing algorithms in multi-skill customer contact centers. Submitted for publication (2006)
13. Koole, G., Mandelbaum, A.: Queueing models of call centers: An introduction. Annals of Operations Research 113(1-4), 41–59 (2002)
14. Kozan, E., Mesken, N.: A simulation model for emergency centres. In: Zerger, A., Argent, R. (eds.) Proceedings of the International Congress on Modelling and Simulation. Advances and Applications for Management and Decision Making. pp. 2602–2608. Modelling & Simulation Society of Australia & New Zealand Inc., Australia, Victoria, Melbourne (2005)
15. Li, X., Zhao, Z., Zhu, X., Wyatt, T.: Covering models and optimization techniques for emergency response facility location and planning: a review. Mathematical Methods of Operations Research 74(3), 281–310 (2011)
16. Maxwell, M., Henderson, S., Topaloglu, H.: Ambulance redeployment: an approximate dynamic programming approach. In: Winter Simulation Conference (WSC), Proceedings of the 2009. pp. 1850–1860. IEEE (2009)
17. Maxwell, M., Restrepo, M., Henderson, S., Topaloglu, H.: Approximate dynamic programming for ambulance redeployment. INFORMS Journal on Computing 22(2), 266–281 (2010)
18. Puts, J.: Emergency Call Center: Finding a balance between costs and quality of service when dealing which emergency calls. MSc research paper, VU University Amsterdam (2011)
19. Rajagopalan, H., Saydam, C., Xiao, J.: A multiperiod set covering location model for dynamic redeployment of ambulances. Computers & Operations Research 35(3), 814–826 (2008)
20. Restrepo, M.: Computational methods for static allocation and real-time redeployment of ambulances. Ph.D. thesis, Cornell University (2008)
21. Ross, E.: Simulation Analysis of Toronto Emergency Medical Service's Communications Centre. BSc thesis, University of Toronto (2007)
22. Roubos, D., Bhulai, S.: Approximate dynamic programming techniques for skill-based routing in call centers. Probability in the Engineering and Informational Sciences 26, 581–591 (2012)
23. Savas, E.S.: Simulation and cost-effectiveness analysis of new york's emergency ambulance service. Management Science 15(12), 608–627 (1969)

24. Schmid, V.: Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. European Journal of Operational Research 219(3), 611–621 (2012)
25. Snyder, L.: Facility location under uncertainty: A review. IIE Transactions 38(7), 547–564 (2006)
26. Van Buuren, M., Kommer, G.J., Van der Mei, R., Bhulai, S.: A simulation model for emergency medical services call centers. In: Proceedings of the Winter Simulation Conference. WSC '15, Winter Simulation Conference (2015)
27. Van Buuren, M., Van der Mei, R., Aardal, K., Post, H.: Evaluating dynamic dispatch strategies for emergency medical services: Tifar simulation tool. In: Proceedings of the Winter Simulation Conference. pp. 46:1–46:11. WSC '12, Winter Simulation Conference (2012)
28. Zuzáková, B.: Optimal emergency medical service system design. MSc thesis, VU University Amsterdam (2012)