Aggregate Model-Based Performance Analysis of an Emergency Department

Ivo Adan, Eindhoven University of Technology, Eindhoven, Netherlands

Erjen Lefeber, Eindhoven University of Technology, Eindhoven, Netherlands

Jordi Timmermans, Eindhoven University of Technology, Eindhoven, Netherlands

Annet van de Waarsenburg, Catharina Hospital Eindhoven, Eindhoven, Netherlands

Mireille Wolleswinkel-Schriek, Catharina Hospital Eindhoven, Eindhoven, Netherlands

ABSTRACT

In this paper the authors present an aggregate simulation model for a real-life emergency department, which is based on the concept of effective process times and which uses a token system to model patients claiming multiple resources simultaneously. Although it has been developed for a specific hospital, the model is flexible, and capable to describe different settings. The modeling steps, model specification and model validation are explained in detail. By using a process-based simulation language, the resulting model is transparent, intuitive and easy to use in quantitatively and efficiently evaluating the effect of proposed changes in the operational processes of the emergency department on patient service levels and resource utilizations.

Keywords: *Aggregate Modeling, Effective Process Time, Health Care, Simulation, Simultaneous Resource Possession*

1. INTRODUCTION

Due to rising costs, the health care sector is forced to work more efficiently and to better utilize their resources. Therefore, LEAN principles have been introduced in health care. Also at the Emergency Department (ED) of the Catharina Hospital in Eindhoven (CZE) the LEAN concept has been introduced to improve operational processes (Wolleswinkel, 2012). The aim is to streamline operational processes, i.e., to eliminate unnecessary operations to achieve better performance using existing resources.

To support decision making in process improvement programs, simulation has proved to be an effective tool (Brailsford, 2007; Duguay & Chetouane, 2007; Sinreich & Marmor, 2005). A literature review on the use of simula-

DOI: 10.4018/ijphim.2014070101

tion and modeling in the health care domain can be found in (Jun, Jacobson, & Swisher, 1999; Brailsford, Harper, Patel, & Pitt, 2009; Brailsford & Vissers, 2011), showing evidence that simulation and modeling are growing in popularity. This approach is also followed for the CZE: we develop a simulation model for the ED, based on actual data from the electronic hospital information system (EZIS), and exploit the concept of *effective process time* (EPT), cf. (Hopp & Spearman, 2008). The basic idea is that the various details of patient treatment times are not modeled in detail, but their contribution is *aggregated* into an EPT distribution, the parameters of which are directly estimated from the available data. This concept has been originally developed in semi-conductor manufacturing (Jacobs, Etman, Campen, & Rooda, 2003; Etman, Veeger, Lefeber, Adan, & Rooda, 2011). Its applicability in health care modeling, in particular for an MRI department, has recently been explored in (Jansen, Etman, Rooda, & Adan, 2012), and it is further investigated in the current paper. Characteristic features of the ED are that (i) patients *simultaneously* require multiple resources (e.g., treatment room, nurse, and physician) and (ii) nurses and physicians can spread their attention over *multiple* patients. We propose a novel *token system* to model the above mentioned features of simultaneous resource possession and multi servicing, a concept which has also been instrumental in analyzing multi-server queueing systems with concurrency constraints (Berezner, Kriel, & Krzesinski, 1995; Krzesinski, 2010). This token system, in combination with EPTs, describes the ED at an aggregate level, which is suitable and sufficiently flexible to support the improvement program of CZE, and it distinguishes the current model from other, typically more detailed models proposed in the literature, cf. (Duguay & Chetouane, 2007) and the references therein.

The resulting model, specified in the process-based simulation language Chi 3.0 (Hofkamp & Rooda, 2012), is transparent, flexible and intuitive, and hence, in the spirit of the principles set out in (Sinreich & Marmor, 2005).

The aim of this study is to investigate the capacity level needed to deliver the health care services within the target times set by the hospital management. The capacity consists of (ED-) physicians, (ED-) medical interns, (ED-) nurses and treatment rooms. Using the developed model, we are able to address questions such as:

- What capacity is at least required on a typical Monday to meet the target maximal waiting times?
- How much does the waiting time decrease if the number of nurses increases?
- How does the waiting time change if patients arriving by ambulance are treated with the same priority as other patients?

The contribution of this paper is in the modeling approach: instead of a detailed model we use an aggregate model, which is suitable and sufficiently flexible to support the improvement program of CZE. Various details of patient treatment times are not modeled in detail, but aggregated into an EPT distribution whose parameters are directly estimated from data: a concept that originally has been developed in semi-conductor manufacturing and in this paper is successfully transferred to health care modeling. Furthermore we propose a novel token system to model patients requiring multiple resources (treatment room, nurse and physician) simultaneously and physicians and nurses participating in more than one activity simultaneously. A third contribution in our modeling approach is the usage of recursive partitioning to derive a reliably fitted treatment time distribution from data.

Using this aggregate modeling approach we are able to obtain answers to the above mentioned questions, showing which improvement actions are most effective. In particular, these capacity questions and their effect on waiting times are also the only conclusions that can be drawn from the model. Clearly, for more detailed questions, a more detailed model is required.

In the next section we first explain the modeling steps and challenges. We explain how we model patient arrivals, how we fit treatment time distributions to data, and how we model resources handling multiple patients simultaneously. Then, in Section 3, the model is validated, and its decision supporting power is demonstrated in Sections 4-5.

2. MODELING

As mentioned in the introduction, we developed a simulation model of the ED of the CZE. Before we introduce this model, we first describe in more detail, and through the eyes of an arriving patient, what happens at the ED of the CZE (and what is also typical for EDs at other Dutch hospitals). The patient flow is described using the map in Figure 1, and the patient flowchart, shown in Figure 2.

Prior to the actual arrival, for most referred patients, it is already known that they will arrive in the near future. The paramedic or general practitioner contacts the senior nurse in order to keep him/her up-to-date. The senior nurse then adds a new entry to EZIS. New patients arrive by own transportation or by ambulance. Both patient flows register at the reception. At that time, patients are also logged in on EZIS. This means that the patient is physically present at the ED.

Figure 1. Map of the emergency department of the CZE

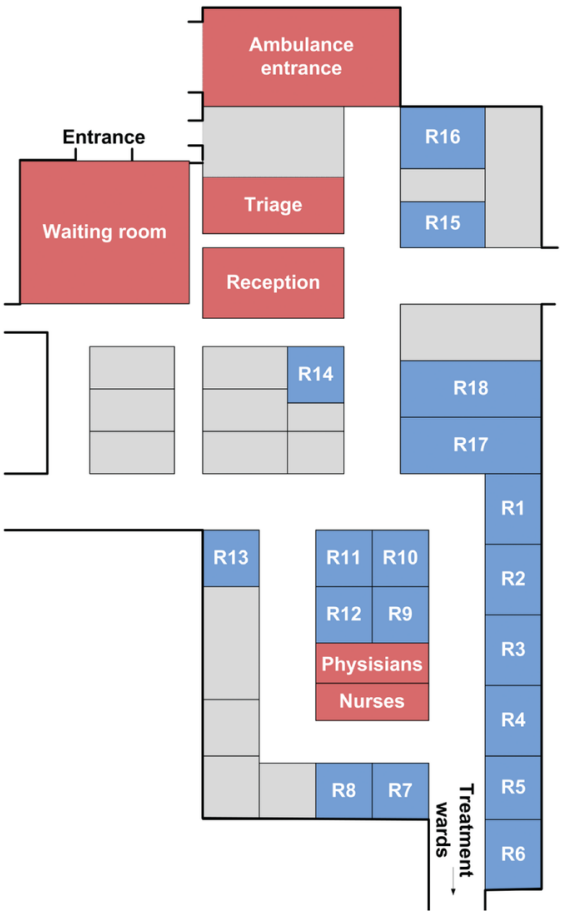
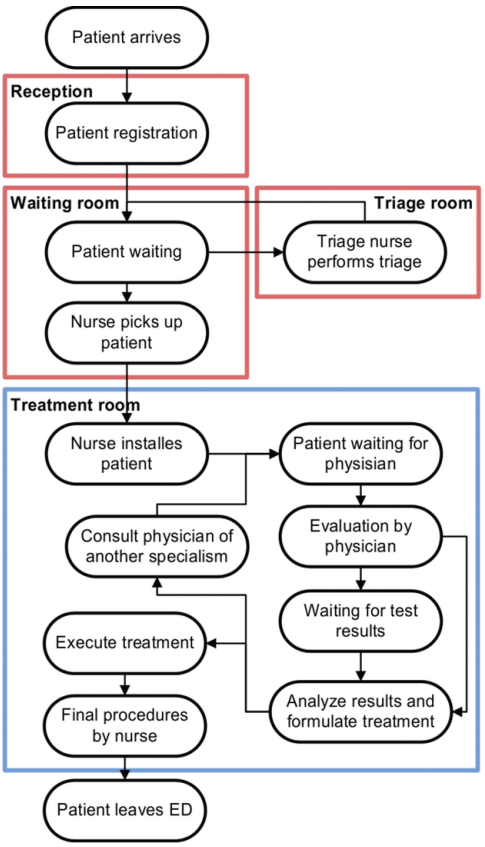


Figure 2. Schematic view of the patient flow in an ED



After registration, the patient will wait in the waiting room. Generally, patients go in order of arrival to the triage room to undergo triage according to the *Dutch Triage Standard*. This system uses four emergency levels: acute (red), urgent (yellow), standard (green) and not urgent (blue). When finished, the patient returns to the waiting room. If a nurse is free and a treatment room is available, the nurse picks up the longest waiting patient with the highest priority to accompany him/her to the treatment room.

Paramedics, transporting patients by ambulance, have to wait at the reception before the patient can be dropped off at a treatment room. The target maximal waiting time is 15 minutes for these patients. This patient flow is not drawn in Figure 2. In the current situation, patients arriving by ambulance are served with priority over patients in the waiting room.

When the patient has arrived in the treatment room, the treatment process starts and the nurse ensures that the patient is installed properly in the emergency room. In some cases, the nurse already starts up a few small examinations, such as taking a blood sample. Next, a physician visits the patient for a first evaluation of the complaints, in most cases done by the medical intern. After consultation with a medical specialist, it is decided which extra examinations are needed, e.g. an X-ray. When the tests are finished and the results are reviewed, the physician determines what treatment is needed to cure the patient. If the physician is uncertain about the complaints and how to treat, then a medical specialist of

another specialty is paged to examine and treat the patient. During the treatment, the responsible nurse keeps monitoring and nursing the patient when needed.

When the treatment is finished, several options are possible. A patient can go home and the nurse can schedule a follow-up appointment at the general practitioner or at the policlinic. Another option is that the patient has to stay for hospitalization, or is transported to another hospital. In those cases, the patient can only leave if a nurse from the ward or a paramedic has arrived to pick up the patient. During the delay that occurs, the treatment room stays occupied and is therefore not available for a new patient.

The above way of working at the ED of the CZE forms the basis of our model. However, we are limited by the available data in EZIS. In the remainder of this section we outline how the available data is incorporated in our model. In particular, limited or no data is available on the activities taking place while the patient is in the treatment room (e.g., number and duration of visits of the responsible nurse and physician), though the *entrance* and *exit time* of the patient in the treatment room are accurately recorded. Therefore, we lump the treatment room process in the blue box of Figure 2) into a single EPT distribution. The parameters of this distribution, however, depend on several patient characteristics. This calls for further lumping, which is done by application of data mining techniques, as described in Section 2.2.

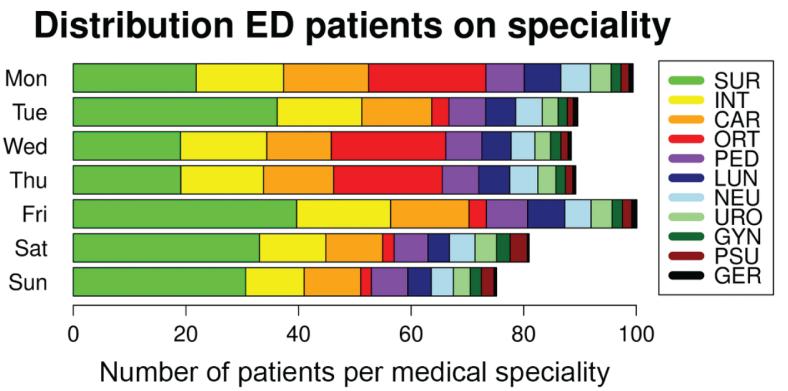
2.1. Patients Arrivals and Diversity

In Figure 3 we show the average number of patient arrivals from Monday to Sunday, as well as the distribution of patients over the most important specialties. In 2011, over 20 different medical specialties were consulted. In our simulation model, only the 11 most visited specialties are included. The ones that are left out have, on average, less than one patient visit per day. The included specialties are surgery, internal medicine, cardiology, orthopedics, pediatrics, lung diseases, neurology, urology, gynecology, plastic surgery and geriatrics.

From Figure 3 it can be observed that the number of patient arrivals on Monday and Friday are significantly higher than on the other days (about 100 patients per day versus almost 90 patients per day). Also, a significant difference in these numbers for both surgery and orthopedics can be seen. The latter is due to agreements between the surgeons and orthopedists: on Monday, Wednesday, and Thursday more patients are seen by the orthopedist, whereas on the other days, these patients are seen by the surgeon.

As mentioned before, a triage system is adopted as waiting room control system, us-

Figure 3. Number of patients per specialism on the different weekdays



ing four emergency levels: acute (red), urgent (yellow), standard (green) and not urgent (blue). The percentages of respectively red, yellow, green and blue patients vary over the different medical specialties, as can be seen in Figure 4.

Figure 5 shows that the arrival rate of patients strongly varies over the day, say from 1 patient per hour during the night, up to 10 patients per hour during peak office hours. Also a significant difference in arrival rate between the different weekdays exists. Monday and Friday are the busiest days in terms of arrival rate. The lowest arrival rates occur during the weekend days. The rates on Saturday and Sunday as well as the rates on Tuesday to Thursday are corresponding.

From the EZIS data it follows that within an interval of 1 hour, it is reasonable to as-

Figure 4. Triage color distribution per specialty

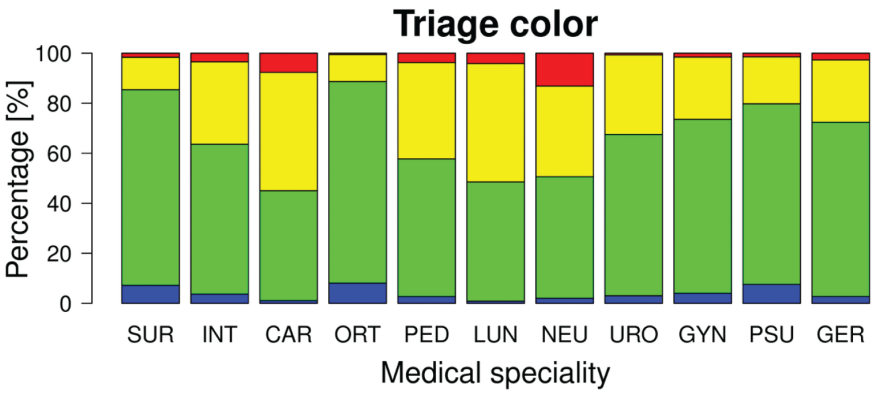
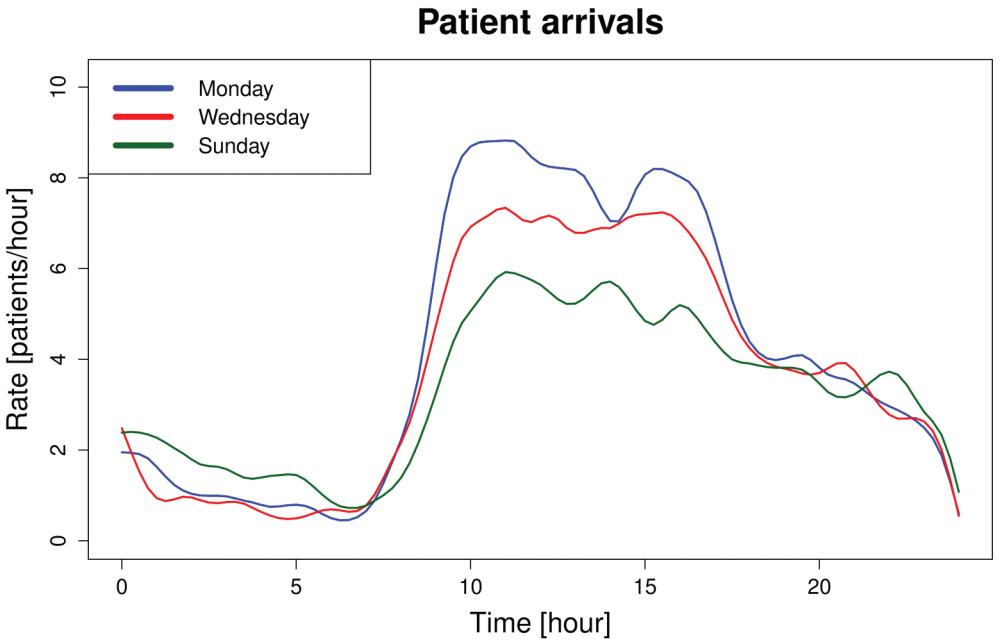


Figure 5. Patient arrival rates on Monday, Wednesday and Sunday



sume that arrivals are Poisson distributed. We therefore model the patient arrivals as an *inhomogeneous Poisson process* (as also proposed in (Alexopoulos, 2008)), with piecewise constant arrival rates during one hour, where we distinguish between ambulance arrivals and arrivals by own transportation. For each day of the week, the hourly arrival rates are estimated from EZIS data. To each arriving patient we assign its required specialty and triage color, according to probabilities which can be read from Figure 3 and Figure 4.

In 2011, 12% of the patients arrived by ambulance. These patients are served first. The other patients arrive by own transportation. After treatment, 31% of the patients is hospitalized.

2.2. Treatment Times

Statistical analysis shows that the total treatment times of patients (while being in the treatment room) depend on several factors:

- Whether the patient requires a second consult or not;
- Type of attending physician (ED-physician or other medical specialist);
- Medical specialty;
- Triage color;
- Age; and
- The number of patients currently treated by the physician.

The dependence on the number of consults during treatment is illustrated in Figure 6, depicting the empirical distributions of total treatment time for patients requiring one and two consults, respectively. As expected, the total treatment time for patients with one consult is (stochastically) smaller than the total treatment time for patients with two consults, cf. (Ross, 1995).

Besides the second consult, also the medical specialty, the triage color, the age, the type of attending physician (ED-physician or other medical specialist) and the number of patients that the physician treats at that moment (PIP) are factors that have a significant influence. Table 1 shows for all factors and categories the

Figure 6. Empirical distribution of total treatment time for patients requiring one and two consults, respectively

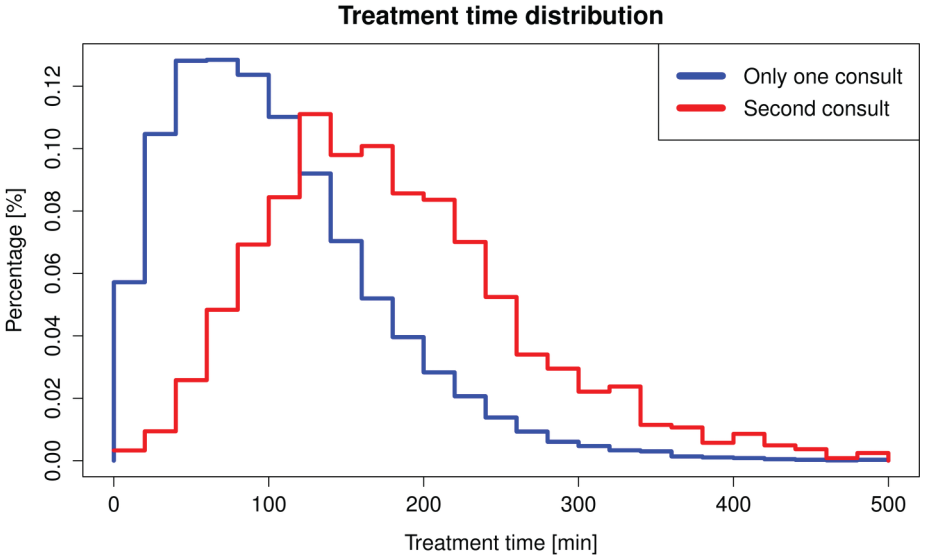


Table 1. The mean and variance of the historical treatment times if all patient treatment times are only split on one category

Classification		Mean	Variance	Classification		Mean	Variance
Second consult	No	107	5874	Triage	Red	126	5035
	Yes	184	7927		Yellow	141	6271
Type Physician	ED	98	5180		Green	106	5679
	Other	133	5842		Blue	67	4278
Specialism	SURt	153	4606	Age	0-3	93	4340
	SURn	90	5447		4-16	82	4282
	INT	160	6946		17-35	96	5819
	CAR	115	7245		36-55	112	6478
	ORT	83	4014		56-75	123	6303
	PED	105	4830		76+	142	6891
	LUN	157	4767	PIP	0	116	7816
	NEU	127	5576		1	117	5834
	URO	116	4561		2	118	6238
	GYN	114	7522		3	112	5785
	PSU	75	4310		4	103	5407
	GER	141	4409		5+	100	4919

mean and variance of the treatment, if all patient treatment times are only split on one factor.

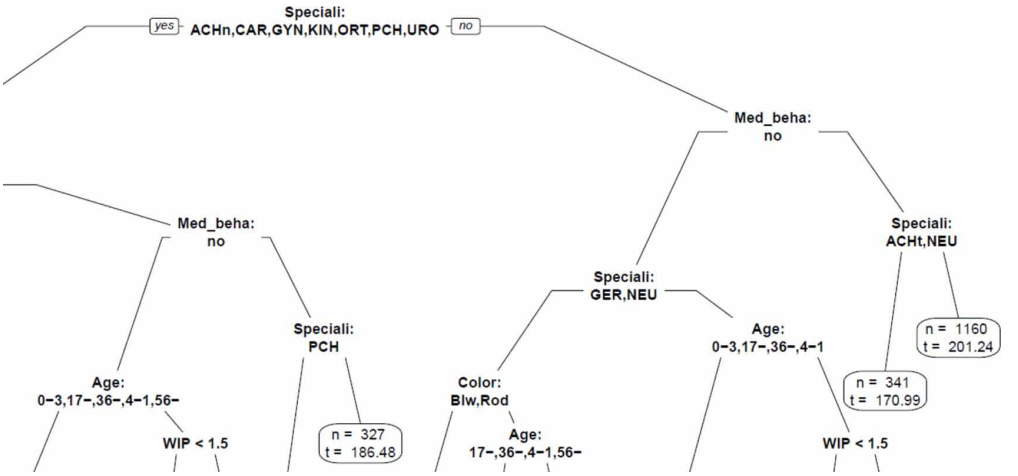
This combination of factors that have an effect on the treatment time leads to almost 7000 treatment groups. Since the ED has been visited by 34.000 patients in 2011, we get, on average, 5 treatment time realizations per group. Hence, it is unreliable to sample treatment times from empirical data or fitted (EPT) distributions for these groups. To cope with this problem, the data mining technique of *recursive partitioning* has been used. The package 'rpart' (Therneau & B. Atkinson, 2012) of the statistical software program R (Gentleman & Ihaka, 2012) has been used to generate a decision tree, which specifies the partitioning. The first part of this tree is shown in Figure 7.

In the root, all groups are lumped together. Then, in each decision step, the current group is split into two groups by the factor with the highest influence on the mean treatment time. The 'n' in a leaf denotes the number of patient treatment times that fit into that group. The "t" stands for the mean of the treatment times into that group. The entire tree can be found in (Timmermans, 2012b, Appendix A); it consists of only 37 leaves, each containing patient treatment times from many groups. The treatment time of each group is now assumed to be Gamma distributed, fitted to the mean and variance of *all* treatment times in the corresponding leaf. As a result, many of the almost 7000 groups use the same, but now *reliably fitted*, treatment time distribution.

2.3. Resource Capacity

Nurses, but also physicians, are capable of handling multiple patients simultaneously. To capture this 'multi-processing' feature, we adopt a novel and flexible *token system* to model the capacity of the physicians, nurses and triage nurses. To start the treatment, a patient claims a combination of tokens representing the resources that are simultaneously needed. Four nurse tokens are used to represent one

Figure 7. Part of the treatment time decision tree



nurse, because (s)he can treat a maximum of four patients at the same time. So each patient needs one nurse token. Moreover, the triage nurse, senior nurse and physician are modeled as respectively one, two and two or three tokens.

So assume that for the simulation model we have 2 triage nurses, 5 nurses, 1 ED-physician who can see three patients simultaneously, and 1 specialist for each specialism who can see two patients simultaneously. Then we have 2 triage nurse tokens, 20 nurse tokens, 3 ED-physician tokens, and 2 specialist tokens for each specialism. Whenever a patient is waiting and a treatment room is available, the patient can enter the treatment room, provided both 1 nurse token, and one ED-physician token are available. In that case, the patient can enter the treatment room and takes those tokens. When the treatment of the patient has been finished, the tokens are released and become available for the other patients.

Note that the token system describes the capacity claim by patients at an *aggregate level*: nurse and physician capacity are claimed during the treatment, but the number and duration of visits during the treatment are not modeled. In other words, nurses and physicians can spread their capacity (tokens) over multiple patients present in the treatment rooms, but we do not exactly model how and when.

Each patient demands one triage nurse token during the triage process and one nurse and one physician token during treatment. An exception is made for acute (red) patients. They demand more intensive care for the first 15 to 30 minutes of their treatment. So, all red patients claim more tokens for the first part of the treatment.

Not only the arrival rate is different for each hour of each day of the week, the model also assumes a different working roster for each weekday. The roster specifies the available capacity at each point in time during the day. For example, the staffing levels are lower at night. Also the transfer from night to morning shift is taken into account by decreasing the available capacity between 7:30h and 9:00h.

2.4. Discrete-Event Simulation Model

For specifying our discrete-event simulation model we used the programming language Chi 3.0, see (Hofkamp & Rooda, 2012), which is an instance of a parallel processes formalism. A system is abstracted into a model, with cooperating processes, connected to each other via channels. The channels are used for exchanging material and information. The model of the ED consists of a number of concurrent processes

connected by channels, denoting the flow of patients or information.

The process to simulate a *single day* of the ED at CZE, is depicted in Figure 8. It consists of the generators *G*, the arrival process delay *D*, the waiting room *W*, the triage process *T*, the treatment rooms *R* and the exit *E*.

The generators G^k represent the arrival processes and create not only the individual patients with a certain arrival rate, but also their attributes (such as, e.g., specialty, triage color, age). The first generator creates patients arriving by ambulance and the second one generates patients arriving by own transportation. A new patient is sent via *D* to the waiting room *W*. The delay process *D* represents the time needed to register a new arrival.

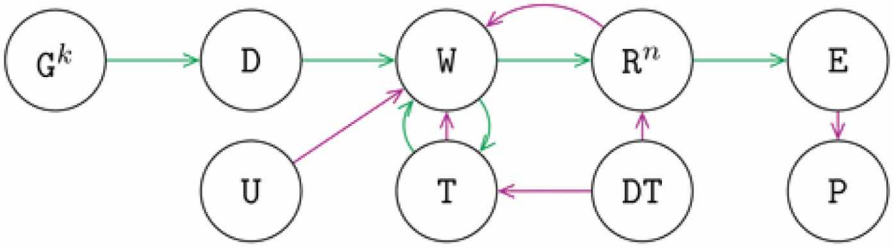
The waiting room process keeps track of all waiting patients and of the availability of the recourses. It uses this information to determine when and which patient will go to the triage process or to the treatment room first. The triage process *T* receives patients from the waiting room and sends the patients back after a certain amount of time, required for performing the triage. A triage can start if both the triage room and triage nurse are available. When the triage is completed, the waiting room process is informed that the triage nurse and triage room are available again. Process *DT* is used to sample the triage time. If the treatment can start, the patient is sent to one of the treatment rooms R^n. The update process *U* is used to report staffing changes to process *W*. The waiting room *W* also receives information from *T* and R^n about their availability.

The treatment rooms are modeled individually and each room can be occupied by one patient at a time. Treatment room *R* receives a patient and requires nursing and physician capacity. The treatment itself is modeled as a time delay for the patient, an occupation of nursing and physician tokens and an occupation of the treatment room. The total number of available tokens is decreased by the number of tokens required by the patient for the entire treatment time. As mentioned in Section 2.3, red patients need more capacity for the first part of their treatment. If this first part is finished, the required capacity is reduced to normal level, i.e., to one nurse token and one physician token. For patients that need a second consult, the specialty of required physician capacity is changed when the treatment is halfway. Process *DT* is also used to sample the treatment time. After (and possibly during) this delay, capacity is released again and process *R* informs the waiting room.

When the treatment is finished, the patient goes to the exit *E*. This represents the departure of a patient. The patient either goes home or is hospitalized. The exit process sends information to the print process *P* which takes care of the simulation output. Next, the print process also signals when the system is empty and a new simulation day can start.

Note that the available resource capacities (such as, number of triage and treatment rooms, (triage) nurses, physicians, and so on) are model

Figure 8. The process to simulate a single day. Patients are transferred using the green channels, other information uses the purple channels.



parameters, which can be easily adapted to the situation at hand. For more details on the model and code, see (Timmermans, 2012b, Chapter 3 and Appendix B).

3. MODEL VALIDATION

The model has been validated by (i) team discussion, and (ii) comparing the model output with the historical data.

The simulation structure and the results have been discussed in several team discussions. Hospital managers, the head of the ED, ED-physicians and senior nurses have been involved in these discussions. During these meetings, a software package was used for the analysis of historical and simulation data. The software package consists of three tools, see Figure 9. The first tool is an Excel program to transform historical data into input files for the other two tools. The second tool is the simulation model implemented in the language Chi 3.0 (Hofkamp & Rooda, 2012). The third one has been developed in R (Gentleman & Ihaka, 2012) to visualize and analyze historical and simulation output data. For more information, see the software package manual (Timmermans, 2012a).

As a result of these discussions, most assumptions were confirmed, but the meetings also led to new insights, such as, e.g., the need of a separate stream for patients arriving by ambulance. Also, these patients get a higher priority while waiting. Another remark during the discussions was that not all treatment rooms are used equally. Some rooms are more suitable for gynecology or otolaryngology patients and other rooms are more often used for small traumas. The latter, however, has not been taken into account in the simulation model.

As part of the validation, the historical data and simulation output are compared. As mentioned in Section 2.1, Monday and Friday are the busiest days. Therefore, the results of the simulated Mondays are used in this section to show the match between historical data and simulation results. All colored bands shown in the figures are the 95% confidence intervals.

In Figure 10 and Figure 11, the historical and simulated average occupations are given for patients that are present in the waiting room, present in the treatment rooms and for the total number of patients at the ED. The results for the first hours of the day are different because the simulated ED begins each day empty. Starting from 9:00h, the waiting room fills to

Figure 9. Software package containing Excel, Chi 3.0 and R tools

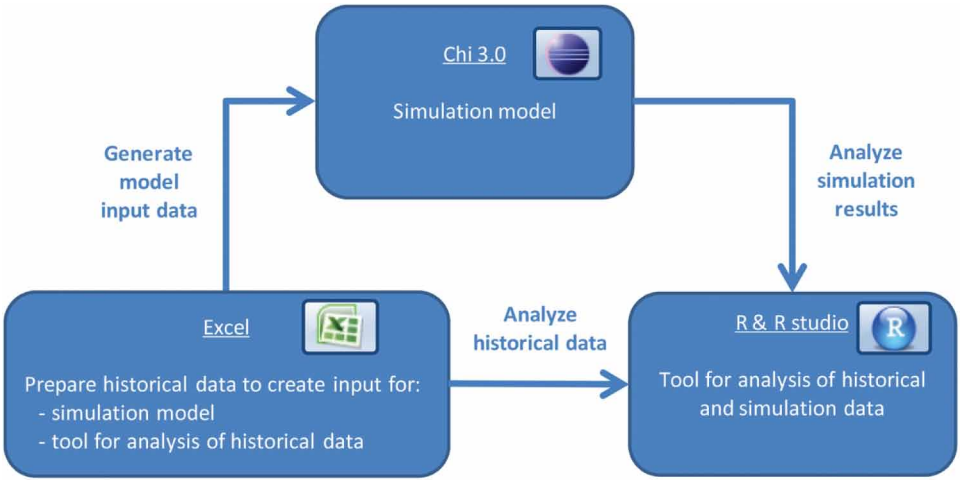


Figure 10. Historical average occupation of patients on Monday

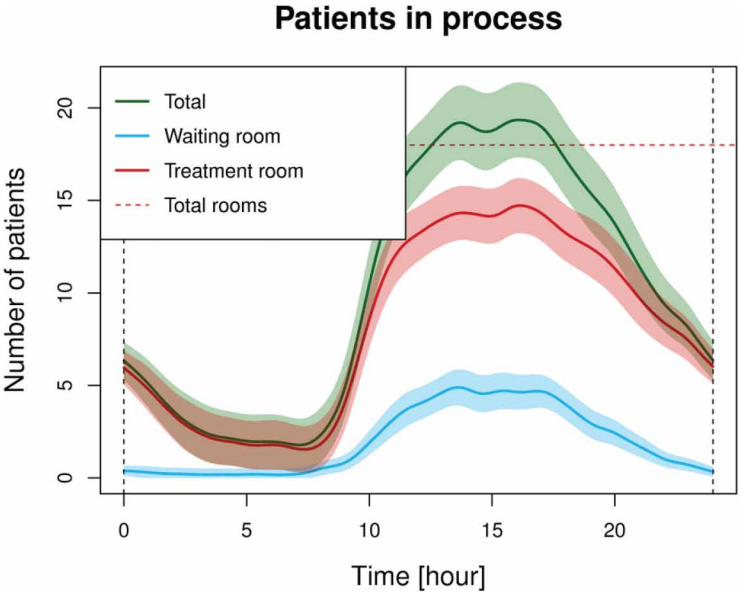
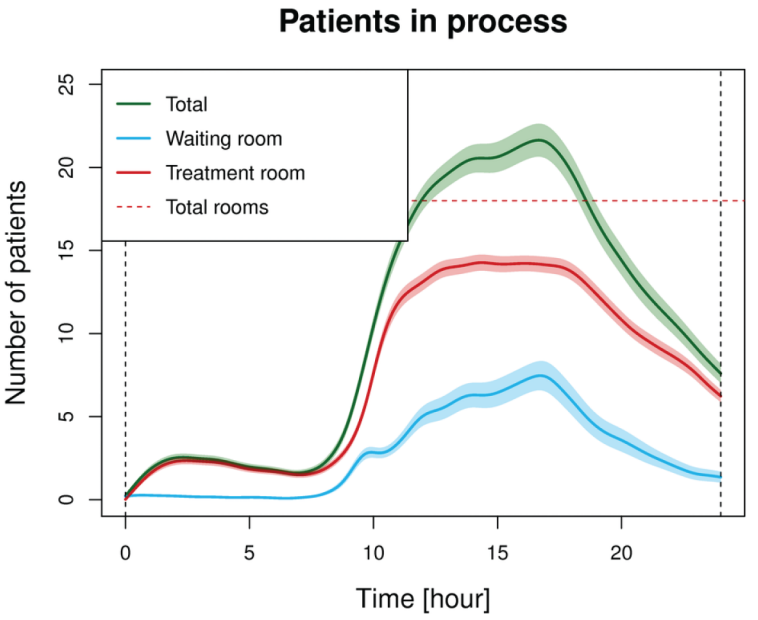


Figure 11. Simulated average occupation of patients on Monday



approximately five patients in the afternoon, on average. Both figures show that the waiting room is empty at the end of the day and that there are on average around 7 patients present in the treatment rooms during the day.

Next, the waiting times are discussed. As can be seen in Figure 12 and Figure 13, the distributions of the waiting times for yellow patients give a relatively close match. Similar results can be obtained by comparing other patient categories. In Figure 14, the average *cycle time factor* (Hopp & Spearman, 2008) is plotted during the day. This factor is the total time a patient is present at the ED divided by the treatment time, and thus provides an indicator of logistic efficiency. If, for patients starting their treatment at time t, this factor is close to one, their waiting time is short compared to their treatment time. Figure 14 shows a good match between historical and simulated data. During peak hours the cycle time factor raises to 1.5-2.0, which is, in terms of manufacturing, a good performance.

4. ANALYSIS OF 2011

In addition to the simulation model, a tool for analysis of simulation output has been developed in R, see also Figure 9. The output is processed and *adaptive* plots are created using the R package *playwith*. That tool has been used to conduct the analysis in this section. By means of this tool, improvement opportunities can be evaluated, such as, for example, opportunities to reduce waiting times, to reduce the number of patients waiting or to improve the utilization of treatment rooms or nursing capacity. Here we restrict ourselves to investigating opportunities to reduce the percentage of yellow and green patients, present on Monday between 10:00h and 20:00h, that exceed the target maximal waiting time of respectively 60 and 120 minutes. This time window is chosen, because it is one of the busiest moments at the ED.

Before investigating possible improvements, first the original situation is simulated. The simulation results for Monday are shown

Figure 12. Historical waiting time for yellow patients arriving on Monday

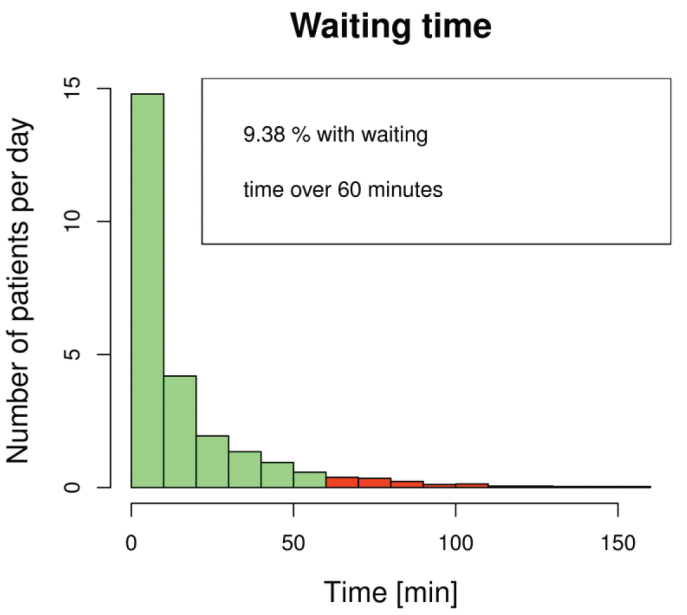


Figure 13. Simulated waiting time for yellow patients arriving on Monday

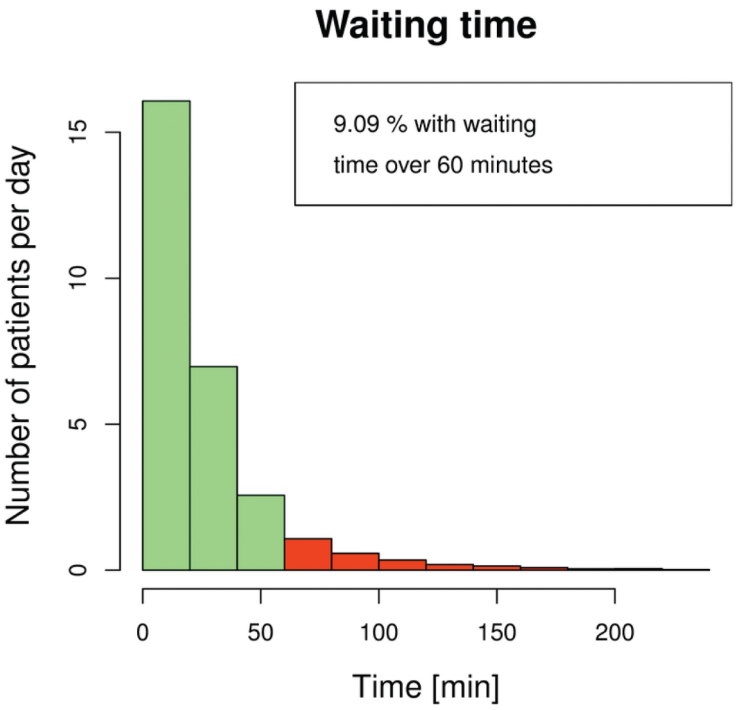
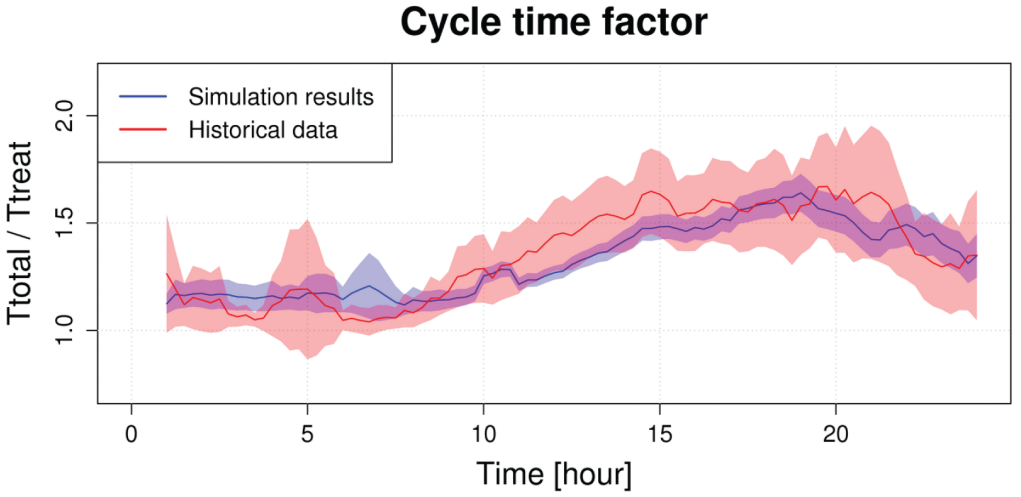


Figure 14. Cycle time factor for historical and simulated patients on Monday. The colored band represents the 95% confidence interval. The patient's cycle time factors are located according to the time their treatment starts.



in Figure 15. The colored bands in the third graph are the 95% confidence intervals. One can see that over 8% of the yellow patients exceed the target maximal waiting time. The vertical black dotted lines in the plot on the right mark the time interval of 10:00h to 20:00h. Possible opportunities for improvement are:

- More treatment rooms;
- No priority for ambulance patients;
- More nursing capacity;
- More physician capacity;
- Treatment time reduction; or
- A combination of these opportunities.

The effect of these opportunities is analyzed in the remainder of this section.

4.1. Extra Treatment Room

To investigate the effect of more treatment room capacity, a simulation is performed with one extra room, 19 in total. The results are shown in Figure 16.

The increase of one treatment room leads to a decrease of 0.5% and 20.3% for respectively yellow and green patients that exceed maximal target waiting time. Corresponding with this result, one can see that the waiting room gets less filled.

Figure 15. Unadapted simulation results on Monday

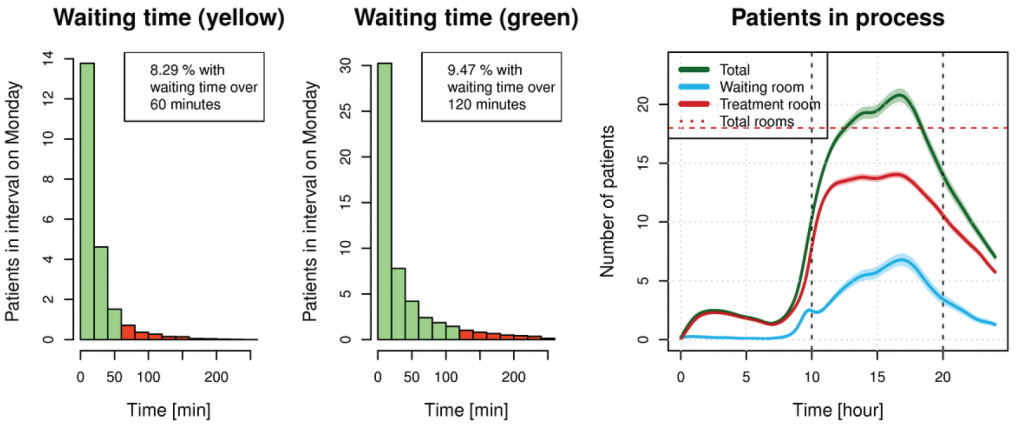
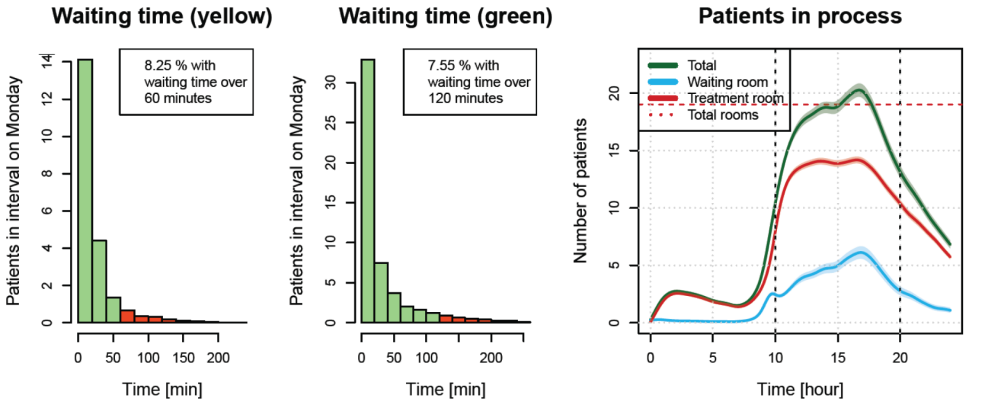


Figure 16. Simulation output using 19 treatment rooms



4.2. No Priority for Ambulance Patients

In the current situation, ambulance arrivals are given priority. The paramedics have to continue with their next request and therefore they would like to have minimal delays at the ED. What if an extra waiting room is created for patients arrived by ambulance? Using this extra waiting room, the senior nurse can follow the same procedure for ambulance arrivals as for patients arriving by own transportation. The results are shown in Figure 17.

A small increase in waiting times can be seen for yellow and green patients. This result can be explained by the fact the former ambulance patients are not directly transferred to a treatment room but have to wait for available capacity. The effect is not very large due to the fact that only 12% of the total patients arrived by ambulance.

4.3. Increase Nursing Capacity

Next, the influence of nursing capacity is examined. A simulation is conducted in which the nursing capacity is increased by four tokens from 10:00 h until 20:00 h. As mentioned in Section 2.3, four tokens represent one nurse. Figure 18 shows the results of the simulation including the roster and used nursing capacity on the right-hand side.

Figure 17. Simulation output without distinguishing patients by type of arrival

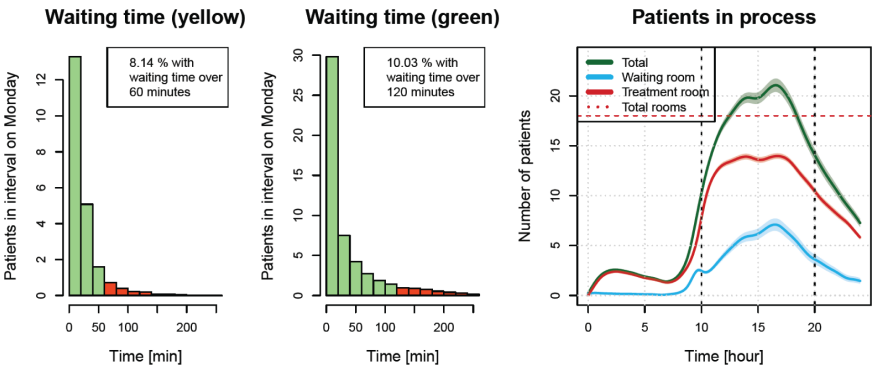
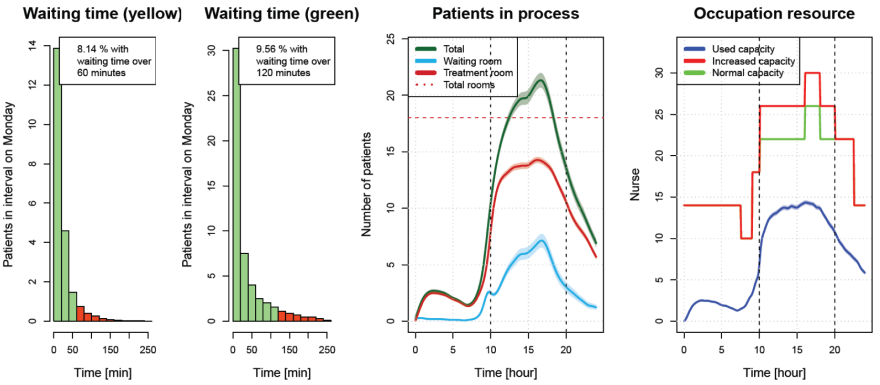


Figure 18. Simulation output with extra nursing capacity



Despite the increase of nursing capacity, the number of patients that exceed the target time remains the same.

4.4. Increase Physician Capacity

The following improvement opportunity is the increase of physician capacity. First, the capacity of the ED-physician is increased by three tokens. These tokens represent one extra physician. Figure 19 shows that, similar to the results of increasing nurse capacity, an increase of ED-physician capacity does not result in a decrease of patients that exceed the target time.

Next, instead of increasing the ED-physician capacity, the maximum capacity of all other physician types is raised with one token between 10:00h and 20:00h. The results are shown in Figure 20.

In addition, the occupation of the cardiologist is shown to illustrate the increase of capacity. The percentage of patients that exceed the target time substantially drops with 63.1% and 66.8% for respectively yellow and green patients.

4.5. Treatment Time Reduction

As mentioned in the introduction, the LEAN concept has been introduced in the CZE to improve operational processes (Wolleswinkel, 2012). The aim is to eliminate unnecessary operations to achieve better performance us-

Figure 19. Simulation output with extra ED-physician capacity

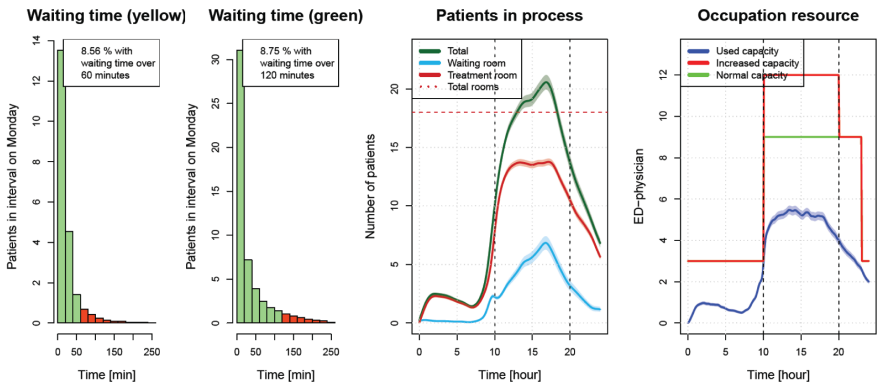
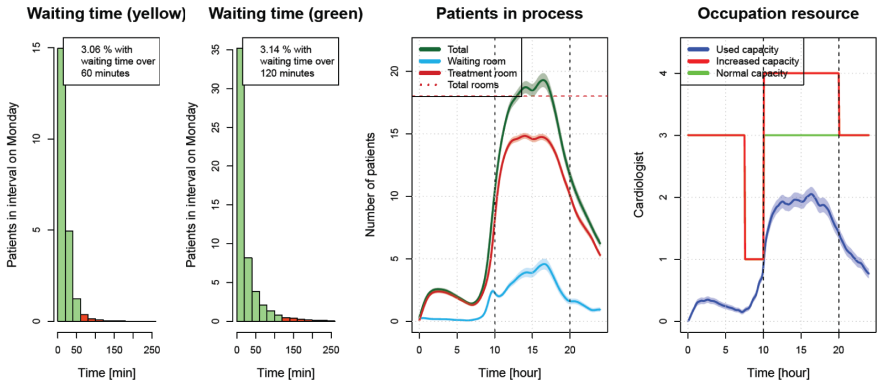


Figure 20. Simulation with one extra token per physician type, except for the ED-physician



ing existing resources. This approach can, for example, result in a reduction of the treatment time by 10 minutes per patient.

One way to potentially achieve this reduction is to shorten the time for hospitalization. This time starts from the moment that the treatment actually finishes until the patient is picked up by the nurse of the ward. About 30% of the patients, mostly elderly, are hospitalized.

A simulation is performed in which the treatment time per patient is reduced with 10 minutes. The results are shown in Figure 21. The number of waiting patients as well as the number of occupied treatment rooms is decreased. The percentage of patients that exceed the target maximal waiting time is reduced to 6.55% and 5.49% for respectively yellow and green patients, i.e., a reduction of 21.0% and 42.0%.

If the treatment time is increased by 10 minutes per patient, an opposite result can be observed. This increase results in 9.12% and 14.76% of the yellow and green patients exceeding the target.

4.6. Combination of Improvements

In the previous sections, several improvement opportunities are elaborated. Subsequently, the most influential opportunities are (partially) combined and the results are discussed in this section. The simulation that is performed includes a 10 minutes decrease in treatment time for 50% of the patients, one extra treatment room and one extra token for the cardiologist, medical intern, neurologist and lung specialist from 10.00h until 20.00h. The simulation results are shown in Figure 22.

With the introduced adaption, the percentage of patients that exceed the target waiting time is decreased by 47.0% for yellow patients and by 68.3% for green patients.

In this section, the simulation model is used to explore several improvement opportunities. Treatment time reduction and increased availability of physicians from other departments tend to be powerful improvement opportunities. It can be concluded that a combination of these opportunities also leads to a large improvement.

5. SCENARIO ANALYSIS

In the previous section, the simulation model has been used to investigate improvements based on the 2011 situation. Alternatively, by modifying the input files for the simulation, different trial scenarios can be studied, such as:

- In general, older patients have longer treatment times. What effect has an increase of ED visits by elderly patients, due to the aging population?

Figure 21. Simulation output for 10 minutes treatment time reduction per patient

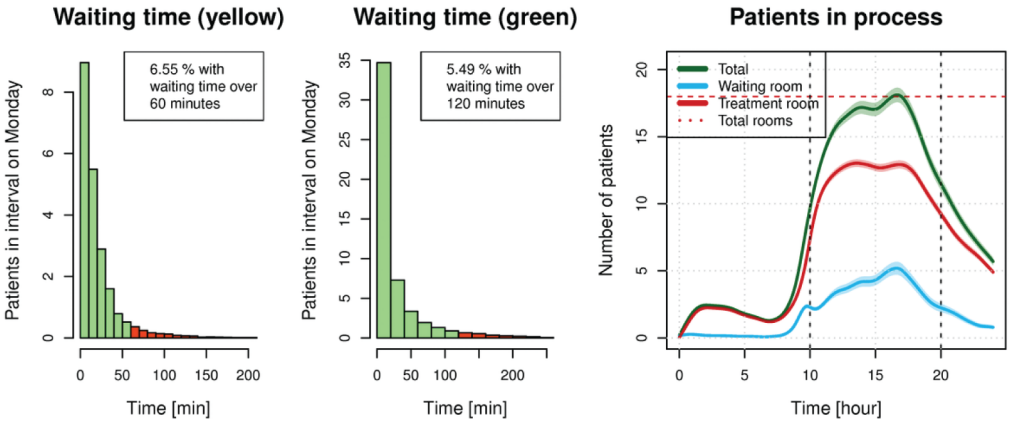


Figure 22. Simulation output for treatment time reduction, physician capacity increase and an extra treatment room

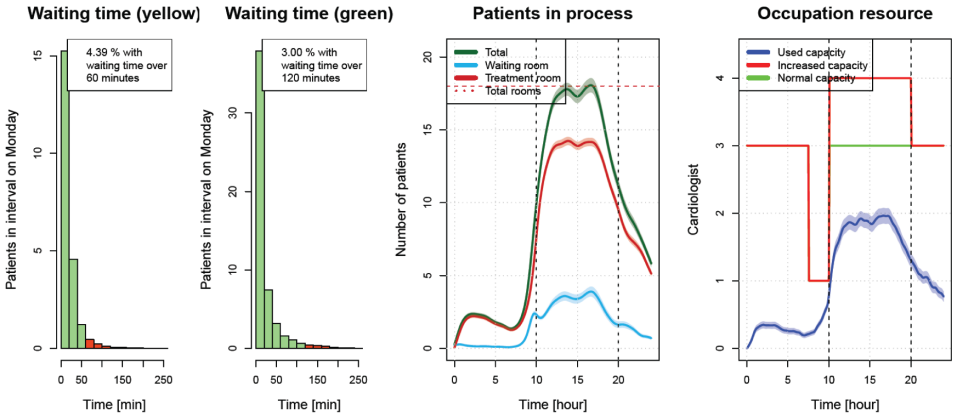


Figure 23. Simulation output for the CZE ED on Monday with a 15% growth of patient arrivals

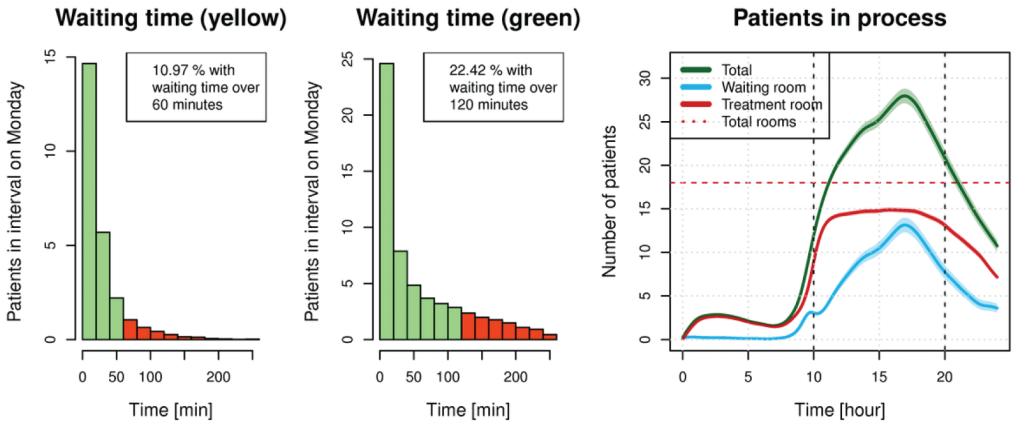
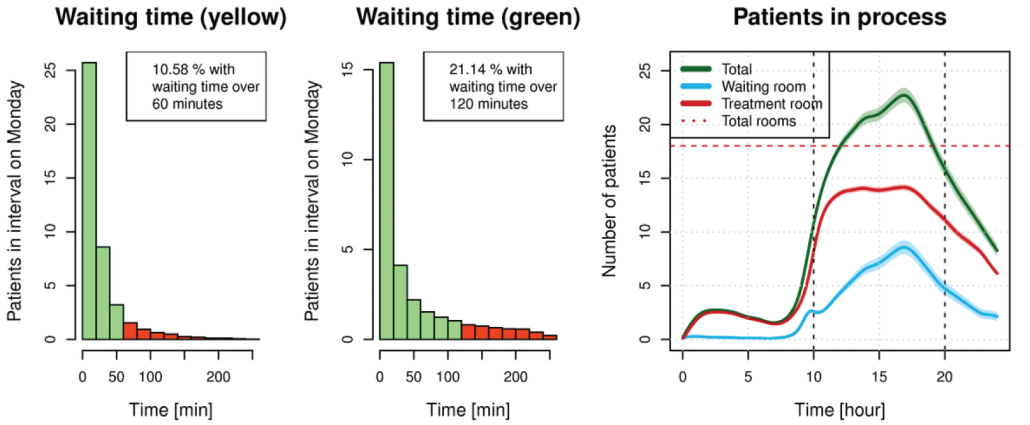


Figure 24. Simulation output for a shifted triage distribution to relatively more urgent patients



- What extra capacity is needed if a neighboring ED closes and the CZE ED has to partially take care of their patients?
- What if the average urgency of patients increases? For example, due to less self-referrals.
- What if more accurate triage results in less second consults and thus in a decrease of treatment times?
- What capacity of ED-physicians is needed if more patients are consulted by the ED-physician instead of the specialist of the attending medical specialty?

Here we consider the second and third scenario only: An increase of the arrival rate and an increase in average urgency.

5.1. Scenario: Increasing Arrival Rate

We consider the scenario: What happens if a neighboring ED has to (temporarily) close? This closure can be caused by a MRSA-outbreak or by financial cutbacks. In this case we assume that the introduced closure results in an increase of 15% in patient arrivals. The growth of patient arrivals by 15% results in a large increase of waiting times, as shown in Figure 23.

On average, there will be more than 12 patients waiting during peak hours. 22.42% of the green patients, present between 10:00h and 20:00h, exceed the target waiting time. For yellow patients, the percentage is 10.97%.

5.2. Scenario: Increasing Urgency

Since the general practitioner (GP) is first care and the ED is second care, patients are not supposed to visit the ED by self-referral. To reduce this group, GPs and EDs try to provide patients better information and the national government is planning to ask an extra charge for self-referrals. In this scenario, we assume that these actions lead to a shift in patient mix: more urgent (red and yellow) patients and less standard (green and blue) patients, as the latter are more likely to understand that their GP is also able to provide proper treatment. Figure 24 shows that the percentage of green patients exceeding the target time has become larger, which is as expected, since there are relatively more higher-priority patients.

6. CONCLUSION

In this paper we developed a simulation model, and we explained modeling challenges and solutions, implementation issues and usage. The resulting process description in the simulation language Chi is transparent, flexible and intuitive, and therefore more easily accepted by its potential users. Also, the use of the visualization and data analysis capabilities of software package R appeared to be crucial to the acceptance of the model. We can conclude that the developed simulation model, equipped with the user interface developed in R, bears the promise to play an important role as decision support tool in the process improvement program at CZE, and possibly also at other hospitals.

The main contribution of this paper is in the modeling approach. Various details of patient treatment times are not modeled in detail, but their contribution is aggregated into an EPT distribution of which the parameters are directly estimated from the available data. This concept has originally been developed and extensively tested in semi-conductor manufacturing and in this paper we successfully transfer this approach to health care modeling. We believe this aggregate modeling concept can more often be used in the modeling of health care systems. Second, we do not model in detail how patients simultaneously require multiple resources (treatment room, nurse, physician) and how nurses and physician spread their attention over multiple patients. Instead we use a token system to model these features in an aggregate way. A third novel aspect of our modeling approach is the usage of the data mining technique of recursive partitioning to reliably fit distributions of treatment times to data.

Currently there is a discussion in the Netherlands on reducing the number of EDs. Simulation models, like the one in this paper, can support this discussion by *quantitatively* evaluating the (logistic) effects of proposed closing or merging of EDs, or by comparing the efficiency of EDs.

REFERENCES

Alexopoulos, C., Goldsman, D., Fontanesi, J., Kopald, D., & Wilson, J. R. (2008). Modeling patient arrivals in community clinics. Omega. *International Journal of Management Sciences*, *36*(1), 33–43.

Berezner, S., Kriel, C., & Krzesinski, A. (1995). Quasi-reversible multiclass queues with order independent departure rates. *Queueing Systems*, *19*(4), 345–359. doi:10.1007/BF01151928

Brailsford, S. (2007). Tutorial: Advances and challenges in healthcare simulation modeling. In *Proceedings of the 2007 winter simulation conference* (pp. 1436–1448). Washington D.C.

Brailsford, S., Harper, P., Patel, B., & Pitt, M. (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*, *3*(3), 130–140. doi:10.1057/jos.2009.10

Brailsford, S., & Vissers, J. (2011). OR in healthcare: A European perspective. *European Journal of Operational Research*, *212*(2), 223–234. doi:10.1016/j.ejor.2010.10.026

Duguay, C., & Chetouane, F. (2007). Modeling and improving emergency department systems using discrete event simulation. *Simulation*, *83*(4), 311–320. doi:10.1177/0037549707083111

Etman, L., Veeger, C., Lefeber, E., Adan, I., & Rooda, J. (2011). Aggregate modeling of semiconductor equipment using effective process times. In *Proceedings of the 2011 winter simulation conference* (pp. 1795–1807). Phoenix (Arizona, USA).

Gentleman, R., & Ihaka, R. (2012). *R project.* Statistics Department of the University of Auckland. Retrieved October 25, 2012, from http://www.r-project.org

Hofkamp, A., & Rooda, J. (2012). *Chi 3.0.0 documentation.* Retrieved January 28, 2013, from http://chi.se.wtb.tue.nl/

Hopp, W., & Spearman, M. (2008). *Factory physics: Foundations of manufacturing management* (3rd ed.). New York: IRWIN/McGraw-Hill.

Jacobs, J., Etman, L., Campen, E., & Rooda, J. (2003). Characterization of operational time variability using effective process times. *IEEE Transactions on Semiconductor Manufacturing*, *16*(3), 511–520. doi:10.1109/TSM.2003.815215

Jansen, F., Etman, L., Rooda, J., & Adan, I. (2012). Aggregate simulation modeling of an MRI department using effective process times. In *Proceedings of the 2012 winter simulation conference* (pp. 1795–1807). Berlin (Germany).

Jun, J., Jacobson, S., & Swisher, J. (1999). Application of discrete-event simulation in health care clinics: A survey. *The Journal of the Operational Research Society*, *50*(2), 109–123. doi:10.1057/palgrave.jors.2600669

Krzesinski, A. (2010). Order independent queues. In R. J. Boucherie & N. M. van Dijk (Eds.), *Queueing Networks: A Fundamental Approach* (pp. 85–120). Springer.

Ross, S. (1995). *Stochastic processes*. Wiley.

Sinreich, D., & Marmor, Y. (2005). Emergency department operations: The basis for developing a simulation tool. *IIE Transactions*, *37*(3), 233–245. doi:10.1080/07408170590899625

Therneau, T., & Atkinson, B. B. R. (2012). *Package 'rpart'—recursive partitioning.* Retrieved November 2, 2012, from http://cran.r-project.org/web/packages/rpart/rpart.pdf

Timmermans, J. (2012a). *Documentation of the simulation model tools of the emergency department at Catharina hospital eindhoven* (MN Report No. MN-420703). Manufacturing Networks Group, Department of Mechanical Engineering, Eindhoven, the Netherlands: Eindhoven University of Technology.

Timmermans, J. (2012b). Efficiency of the emergency department of the Catharina hospital Eindhoven (MSc Thesis, Eindhoven University of Technology, Manufacturing Networks Group, Department of Mechanical Engineering, Eindhoven, The Netherlands). Permanent URL: http://www.tue.nl/uploads/media/2012 MN-420704_JJD_Timmermans.pdf

Wolleswinkel, M. (2012, April). In de start blokken — Tweegesprek over het meerjarenbeleidsplan. [In Dutch]. *Ons Catharien*, *2*(2), 6–7.