AGGREGATE MODEL BASED PERFORMANCE ANALYSIS OF AN EMERGENCY DEPARTMENT

I.J.B.F. Adan[(a)], E. Lefeber[(a)], J.J.D. Timmermans[(a)], A. v.d. Waarsenburg[(b)], M. Wolleswinkel-Schriek[(b)]

[(a)]Manufacturing Networks Group, Department of Mechanical Engineering, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.
[(b)]Catharina Hospital Eindhoven, Michelangelolaan 2, 5623 EJ Eindhoven, The Netherlands.

[(a)][I.J.B.F.Adan,A.A.J.Lefeber]@tue.nl

ABSTRACT

In this paper we 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 evaluating proposed changes in the operational processes of the emergency department.

Keywords: Aggregate modeling, effective process time, health care, simulation.

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 and Chetouane 2007, Sinreich and Marmor 2005). A literature review on the use of simulation and modeling in the health care domain can be found in (Jun et al. 1999, Brailsford et al. 2009, Brailsford and 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 and 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 developed in semi-conductor manufacturing (Etman et al. 2011). Its applicability in health care modeling, in particular for an MRI department, has recently been explored in (Jansen et al. 2012), and it is further investigated in the current paper. Typical features of the ED are that (i) patients *simultaneously* require multiple resources (e.g., treatment room, nurse, 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. This token system, in combination with EPTs, describes the ED at an aggregate level, 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 and Chetouane 2007) and the references therein.

The resulting model, specified in the process-based simulation language Chi 3.0 (Hofkamp and Rooda 2012), is transparent, flexible and intuitive, and hence, in the spirit of the principles set out in (Sinreich and Marmor 2005). It can be used 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 this model, it is possible to address questions such as, for example:

- 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?

In the next section we will first explain the modeling steps and challenges. 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 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 1a, and the patient flowchart, shown in Figure 1b. Prior

Figure 1a: Map of the Emergency Department of the CZE.

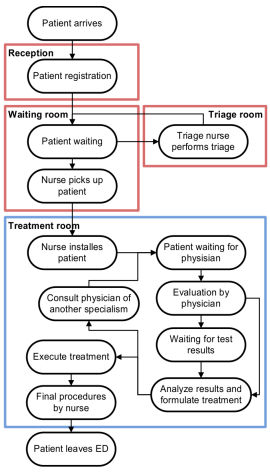

Figure 1b: Schematic view of the patient flow in an ED.

to the actual arrival, for most referred patients, it is already known that they will arrive in the near future. The paramedic or general partitioner 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.

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 non 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 1b. 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 speciality 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 partitioner 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 will lump the treatment room process in the blue box of Figure 1b) into a single EPT distribution. The parameters of this distribution, however, depend on several patient characteristics. This calls for further lumping, which will be done by application of data mining techniques, as described in Section 2.2.

2.1 Patients arrivals and diversity

In Figure 2 we show the average number of patient arrivals from Monday to Sunday, as well as the dis-

tribution of patients over the most important specialities. In 2011, over 20 different medical specialities were consulted. In our simulation model, only the 11 most visited specialities are included. The ones that are left out have, on average, less than one patient visit per day. The included specialities are surgery, internal medicine, cardiology, orthopedics, pediatrics, lung diseases, neurology, urology, gynecology, plastic surgery and geriatrics. ¿From Figure 2 it can be observed that the number of patient arrivals on Monday and Friday are significantly higher than on the other days. Also, a significant difference in these numbers for both surgery and orthopedics can be seen. This 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.

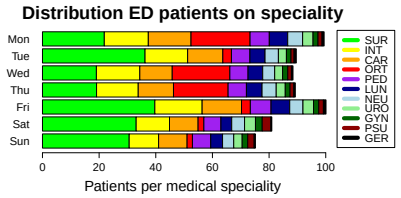

Figure 2: Patients per specialism on the different weekdays.

Though not visible in Figure 2, 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. We therefore model the patient arrivals as an *inhomogeneous Poisson process* (Alexopoulos 2008), with piecewise constant arrival rates during one hour, where we distinguish between ambulance arrivals and arrivals by own transportation. The hourly arrival rates come from EZIS data. To each arriving patient we assign its required speciality, according to probabilities which can be read from Figure 2.

2.2 Treatment times

Statistical analysis shows that the total treatment times of patients (while being in the treatment room) depend on several factors: medical speciality, triage color, age, type of attending physician (ED-physician or other medical specialist), the number of patients currently treated by the physician, and whether the patient requires a second consult or not. This combination of factors leads to almost 7000 treatment groups. Since the ED has been visited by 34.000 patients in 2011, that gives, 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 et al. 2012) of the statistical software program R (Gentleman and Ihaka 2012) has been used to generate a decision tree, that specifies the partitioning. The first part of this tree is shown in Figure 3. In the root, all groups are lumped together.

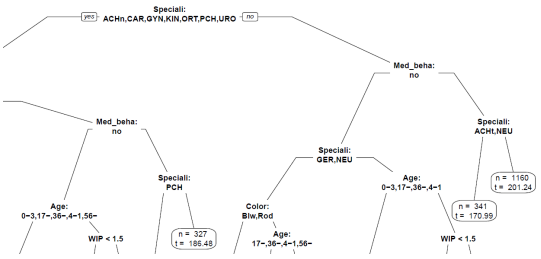

Figure 3: Part of the treatment time decision tree.

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 *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 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, the simulation model uses, for example, 20 nurse tokens to represent 5 nurses. 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 and 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 4. It consists of the

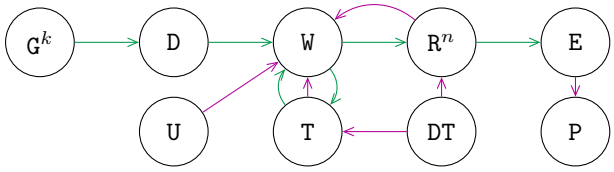

Figure 4: The process to simulate a single day. Patients are transferred using the green channels, other information uses the purple channels.

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., speciality, 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. 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 speciality 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 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 tool developed in R (Gentleman and Ihaka 2012) was used for the analysis of historical and simulation data. This tool has been developed to increase the ease of visualizing and analyzing the 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 gynaecology 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 5a and Figure 5b, 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

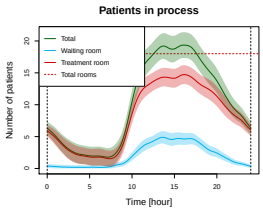

Figure 5a: Historical average occupation of patients on Monday.

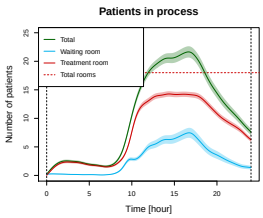

Figure 5b: Simulated average occupation of patients on Monday.

fills to 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 6a and Figure 6b, 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 7, the average *cycle time factor* (Hopp and 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 7 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.

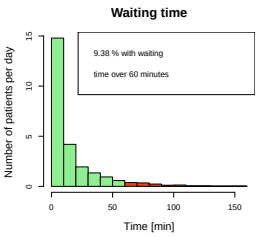

Figure 6a: Historical waiting time for yellow patients arriving on Monday.

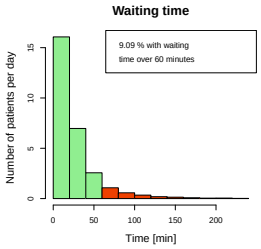

Figure 6b: Simulated waiting time for yellow patients arriving on Monday.

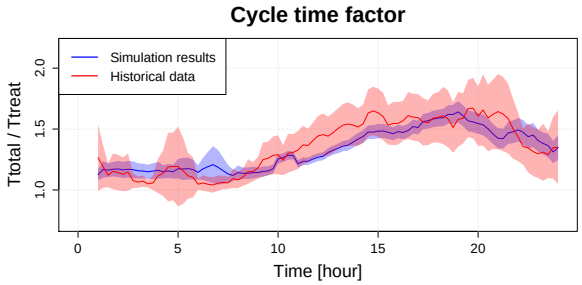

Figure 7: Cycle time factor for historical and simulated patients on Monday. The colored band represents the 95% confidence interval. The patient's cycle time factors are located according to the time their treatment starts.

4 ANALYSIS OF 2011

In addition to the simulation model, a tool for analysis of simulation output has been developed in R. 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 in Figure 8. One can see that over 18% 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, e.g., more treatment rooms, no priority for ambulance patients, more nursing capacity, more physician capacity and treatment time reduction, or a combination of these opportunities. Be-

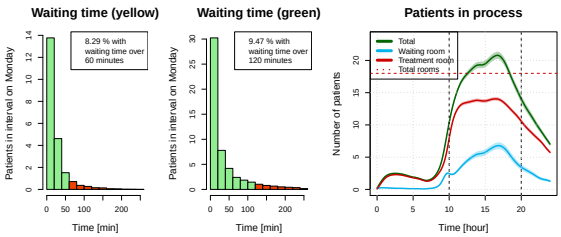

Figure 8: Unadapted simulation results on Monday.

low we only investigate the effect of treatment time reduction.

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

One potential way to 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 9. 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.

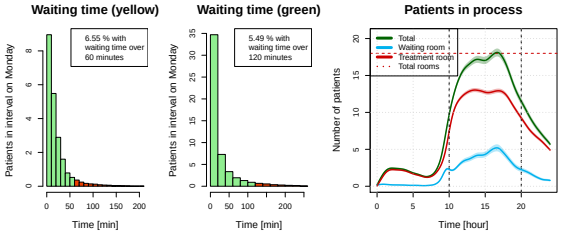

Figure 9: Simulation output for 10 minutes treatment time reduction per patient.

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?
- 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 speciality.

Here we consider the second scenario only: An increase of the arrival rate.

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 10. 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%.

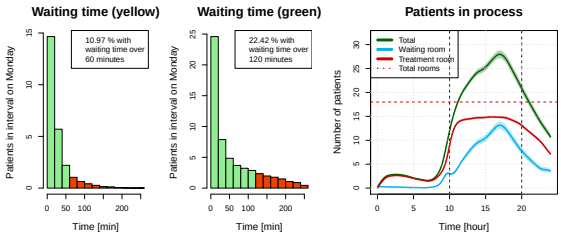

Figure 10: Simulation output for the CZE ED on Monday with a 15% growth of patient arrivals.

6 CONCLUSIONS

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 in the process improvement program at CZE, and possibly also at other hospitals.

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., 2008. Modeling patient arrivals in community clinics. *Omega: The International Journal of Management Science*, 36, 33–43.

Brailsford, S. and Vissers, J., 2011. OR in healthcare: A european perspective. *European Journal of Operational Research*, 212, 223–234.

Brailsford, S.C., 2007. Tutorial: Advances and challenges in healthcare simulation modeling. Proceedings of the 2007 Winter Simulation Conference, pp. 1436–1448. Washington D.C.

Brailsford, S.C., Harper, P.R., Patel, B. and Pitt, M., 2009. An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*, 3, 130–140.

Duguay, C. and Chetouane, F., 2007. Modeling and improving emergency department systems using discrete event simulation. *Simulation*, 83, 311–320.

Etman, L.P.F., Veeger, C.P.L., Lefeber, E., Adan, I.J.B.F. and Rooda, J.E., 2011. Aggregate modeling of semiconductor equipment using effective process times. Proceedings of the 2011 Winter Simulation Conference, pp. 1795–1807. Phoenix (Arizona, USA).

Gentleman, R. and Ihaka, R., 2012. R project. `http://www.r-project.org`. [accessed 25 October 2012].

Hofkamp, A.T. and Rooda, J.E., 2012. Chi 3.0.0 documentation. `http://chi.se.wtb.tue.nl/`. [accessed 28 January 2013].

Hopp, W.J. and Spearman, M.L., 2008. *Factory Physics: Foundations of Manufacturing Management*. 3rd ed. New York: IRWIN/McGraw-Hill.

Jansen, F.J.A., Etman, L.F.P., Rooda, J.E. and Adan, I.J.B.F., 2012. Aggregate simulation modeling of an MRI department using effective process times. Proceedings of the 2012 Winter Simulation Conference, pp. 1795–1807. Berlin (Germany).

Jun, J.B., Jacobson, S.H. and Swisher, J.R., 1999. Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society*, 50, 109–123.

Sinreich, D. and Marmor, Y., 2005. Emergency department operations: The basis for developing a simulation tool. *IIE Transactions*, 37, 233–245.

Therneau, T., Atkinson, B. and Ripley, B., 2012. Package 'rpart' — recursive partitioning. `http://cran.r-project.org/web/packages/rpart/rpart.pdf`. [accessed 2 November 2012].

Timmermans, J.J.D., 2012a. *Documentation of the simulation model tools of the emergency department at Catharina Hospital Eindhoven.* Report MN-420703, Eindhoven University of Technology, Manufacturing Networks Group, Department of Mechanical Engineering, Eindhoven, The Netherlands.

Timmermans, J.J.D., 2012b. *Efficiency of the Emergency Department of the Catharina Hospital Eindhoven.* Thesis (MSc). Eindhoven University of Technology, Manufacturing Networks Group, Department of Mechanical Engineering, Eindhoven, The Netherlands. `http://www.tue.nl/uploads/media/2012_MN-420704_JJD_Timmermans.pdf`.

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