RESEARCH PAPER FOR THE SPECIAL ISSUE: HEALTHCARE BEHAVIOURAL OR

Machine Learning for Healthcare Behavioural OR: Addressing Waiting Time Perceptions in Emergency Care

Daniel Gartner[a] and Rema Padman[b]

[a]School of Mathematics, Cardiff University, Cardiff, United Kingdom
[b]The H. John Heinz III College, Carnegie Mellon University, Pittsburgh, USA

ARTICLE HISTORY
Submitted: February 15, 2018; 1^{st} revision: October 13, 2018; 2^{nd} revision: December 26, 2018; 3^{rd} revision: January 8, 2019; accepted for publication: January 14, 2019

ABSTRACT
Recent research has highlighted the need to improve patient satisfaction by reducing perceived waiting times in hospitals. This study examines factors that are associated with waiting time estimation behaviour and how to control flow of patients who overestimate waiting times. Using data from more than 250 patients, we test the applicability of machine learning methods to understand under-, correct and overestimation behaviour of waiting times in two emergency department areas. Our attribute ranking and selection methods reveal that actual waiting time, clinical attributes, and the service environment are among the top ranked and selected attributes. The classification methods reveal that the precision to classify a patient to the true outcome of overestimating waiting times reaches almost 70% in the first waiting area. If a patient waits in a treatment room which is the second waiting area under study, this precision level reaches almost 78%. We developed a discrete-event simulation model which we linked with the machine learning models of each waiting area. Our scenario analysis revealed that changing staffing patterns can lead to a substantial drop-off in the number of patients overestimating waiting times. Our results can be employed to control waiting time perceptions and, potentially, increase patient satisfaction.

KEYWORDS
Waiting Time Perceptions; machine Learning; attribute Selection; classification; discrete-event simulation

1. Introduction

Research in behavioural operations management has highlighted the need to focus on healthcare (Brailsford and Schmidt (2003); Fügener, Schiffels, and Kolisch (2017)) and the modelling of individuals' perceptions (White (2016)). When patients access systems of emergency care they are typically accompanied by waiting times – a result from variations in arrivals and service times. Individuals, however, perceive waiting times differently. Moreover, when patients are in need of emergency care services, they are concerned about being served immediately (Welch (2009)).

To understand which factors are associated with under-, correct and overestimation of waiting times in an emergency environment and to classify individuals' under-,

CONTACT D. Gartner and R. Padman. Email: gartnerd@cardiff.ac.uk and rpadman@cmu.edu

correct and overestimation behaviour is the aim of this study which is different from the traditional objective of minimizing waiting time. Previous research has shown that perceived waiting times are dependent on five dimensions (Welch (2009)):

(1) Empathy and attitude,
(2) information dispensation,
(3) technical competence (both technical skills and available technology),
(4) pain management and
(5) acceptable waiting times.

While our research is concerned with studying information dispensation by nurses and acceptable waiting times, the main focus is to evaluate factors that are associated with the under-, correct and overestimation of waiting times and how to manage flow in the ED based on these factors.

We investigate factors that influence the behaviour of under-, correct and overestimation of waiting times at a service encounter in a hospital by using machine learning methods. We focus on achieving accurate classification of under-, correct and overestimating waiting times in two subsequent waiting areas. We analyze patient data from emergency department visits consisting of more than 250 records from a hospital in southern Germany. Our results show that machine learning approaches can identify attributes related to under-, correct and overestimation behaviour of patients' waiting times. The methods achieve up to 78% classification accuracy and when combined with discrete-event simulation modelling, they can inform staffing decisions to reduce the number of patients overestimating waiting times.

The remainder of the paper is structured as follows. Section 2 provides a survey of relevant literature. Section 3 introduces the methods that are evaluated in this study. The analysis of the performance of these methods is given in Section 4, followed by concluding remarks in Section 5.

2. Literature Review

Recent research in behavioural healthcare operations management have addressed problems such as biases in surgeons' determination of uncertain surgery time leading to inefficient usage of operating rooms (Fügener et al. (2017)), and the behavioural impact of how queues are designed on human servers resulting in potentially longer waiting lines (Shunko, Niederhoff, and Rosokha (2018)). Literature reviews on patient satisfaction in the emergency department are Boudreaux and O'Hea (2004) and Welch (2009). With respect to waiting times, a survey on socio-demographic attributes associated with waiting is provided by Landi, Ivaldi, and Testi (2018). Outside healthcare, an example to predict waiting times for the arrival of transportation services is Sadat Zadeh, Anwar, and Basirat (2012).

In what follows, we review publications in which main factors that impact perceived waiting times were discovered. We focus on recent journal articles published after 1995.

To obtain a greater insight into the nature of PWT and its relationship to patient satisfaction, Table 1(a) introduces research on waiting times broken down by the following research fields and publication types:

- General services,
- internet services and
- ED services.

General services	Antonides, Verhoef, and van Aalst (2002); Luo, J. Liberatore, L. Nydick, B. Chung, and Sloane (2004)
Internet services	Hong, Hess, and Hardin (2013)
ED services	Boudreaux and O'Hea (2004); Hedges (2002); Nanda et al. (2012); Shaikh, Witting, Winters, Brodeur, and Jerrard (2013); Soremekun, Takayesu, and Bohan (2011); Thompson, Yarnold, Adams, and Spacone (1996); Thompson, Yarnold, Williams, and Adams (1996); Welch (2009)

(a) Classification of Articles by Field and Publication Type

PWT is related to and impacts satisfaction strongly	Boudreaux and O'Hea (2004); Soremekun et al. (2011); Thompson, Yarnold, Williams, and Adams (1996); Welch (2009)
Satisfaction depends more on PWT than AWT	Hedges (2002); Thompson, Yarnold, Williams, and Adams (1996)
Patients inaccurately estimate AWT and PWT	Thompson, Yarnold, Adams, and Spacone (1996)
Distraction shortens PWT	Antonides et al. (2002); Nanda et al. (2012); Shaikh et al. (2013); Thompson, Yarnold, Williams, and Adams (1996); Welch (2009)
Distraction increases PWT	Hong et al. (2013)
Process reengineering impacts PWT	Luo et al. (2004)

(b) Overview of Perception-related Outcomes

Table 1. Classification of Publications in the Field of Waiting Time Perceptions

Table 1(a) reveals that there are two relevant studies related to general services. Antonides et al. (2002) investigate the effects of the waiting environment with distraction by music, TV and queue information on PWT. They disclose that overestimation of waiting time is one crucial reason for an unsatisfying service perceived. Moreover, they conclude that providing information about the remaining waiting time reduces overestimation and consequently increases customer satisfaction. Additionally, Luo et al. (2004) discover the positive influence of changing the processes on AWT and PWT.

The internet services research area is the focus of Hong et al. (2013) who evaluate online waiting time perceptions. Their study reveals that time-related visual content shown to customers who wait for a download makes users perceive the download slower.

The acceptability of a time-tracker display is studied by Shaikh et al. (2013) who conclude that patients prefer an ED in which the estimated wait time is displayed. Another empirical study was conducted by Nanda et al. (2012) whose objective was to analyze the effect of visual art on patients' and visitors' behaviour in the ED. They found a significant reduction in restlessness, noise level, and people staring at other people in the room. They conclude that visual art has positive effects on the ED waiting experience.

Table 1(b) presents the results directly related to the customer's perception by clustering the findings into five groups. The majority of the selected articles is associated

with PWT in the EDs of hospitals. For instance, Thompson, Yarnold, Williams, and Adams (1996) interviewed the patients of a community hospital's ED to determine the effects of actual waiting time, perception of waiting time, information delivery, and expressive quality on patient satisfaction. The authors identify PWT as one of the main impacting factors on patient satisfaction. Soremekun et al. (2011) validate this result and recommend an intense focus and improvement on the service perception. With regard to patient satisfaction, Welch (2009) identifies the strongest correlation of acceptable wait times and empathy. In accordance with Boudreaux and O'Hea (2004), all the results show the strong impact of PWT on patients' satisfaction.

Thompson, Yarnold, Williams, and Adams (1996) and Hedges (2002) find that the effect of satisfaction depends more on PWT than on AWT. Also, an evaluation of the accuracy of patients' time estimation by Thompson, Yarnold, Adams, and Spacone (1996) shows that on the one hand, the examined participants tend to overestimate the period from triage until the first examination. On the other hand, they observe an underestimation of the total amount of time spent in the ED.

In addition, there are distraction-related factors inducing changes of contentment. Experimental studies prove that information about the remaining waiting time communicated by staff reduce the PWT (see e.g. Antonides et al. (2002)). Furthermore, music, television or time tracker systems are widely recognized tools to improve the well-being of patients in a waiting atmosphere (see e.g. Shaikh et al. (2013)).

In summary, PWT depends on various factors and is strongly related to patient satisfaction. Therefore, the reduction of PWT is a promising variable to improve patient satisfaction in the ED.

The approaches proposed in our paper can be categorized into and differentiated from the literature on the management of perceived waiting times in emergency departments as follows: First, with respect to the study design to find relations between actual and perceived waiting times, the approach of Thompson, Yarnold, Adams, and Spacone (1996) is similar. However, instead of asking patients 2-4 weeks after their service experience in the Emergency Department (which was the study setting in Thompson, Yarnold, Adams, and Spacone (1996)), we directly get feedback from patients during the service process.

The second major difference in comparison to previous work is that we employ different attribute ranking and selection techniques to evaluate and select a concise set of attributes that potentially explain the under-, correct and overestimation behaviour of patient's waiting times. Our work therefore differs from previous work because we refrain from using regression models which underly the assumption on having independent variables. Also, we model the problem as a three-class classification problem because we want to characterize patients who are likely to under-, correctly or overestimate their waiting times.

The third major difference in comparison to previous work is that we evaluate different classification techniques on a variety of metrics such as classification accuracy and overestimation precision.

Finally, we develop a discrete-event simulation model of the ED under study and link the under-, correct and overestimation of waiting time classification with the model. By running a scenario analysis we can find out how the number of patients who overestimate waiting times can be reduced.

3. Methods

3.1. Machine Learning Methods

We provide a formal description of the classification problem of under-, correct and underestimating waiting times. Let $\mathcal{I}$ denote a set of individuals (patients who enter the emergency department) and $\mathcal{W} := \{1, 2\}$ denote the set of waiting areas, see Figure 2 in Section 4. Let $\mathcal{E}_w$ denote the set of outcomes for waiting area $w \in \mathcal{W}$ where

$$e_i = \begin{cases} 0 & \text{if patient's estimate waiting time} < \text{actual waiting time} \\ 1 & \text{if patient's estimate waiting time} = \text{actual waiting time} \\ 2 & \text{otherwise} \end{cases}$$

For our classification problem, we have three outcomes for each waiting area $\mathcal{E}_w := \{0, 1, 2\}$ which means that a patient is either someone who under-, correctly or overestimates waiting time. For each patient $i \in \mathcal{I}$, we observe a set of attributes $\mathcal{A}_w$ collected until the time point when he leaves waiting area $w \in \mathcal{W}$. The patient's true outcome, $e_i \in \mathcal{E}_w$, is computed based on an indicator function of the actual waiting time and the patient's estimate. Let $\mathcal{V}_a$ denote the set of possible values for attribute $a \in \mathcal{A}_w$ and let $v_{i,a} \in \mathcal{V}_a$ denote the value of attribute a for patient i. We wish to classify e_i when patient i is a patient who under-, correctly or overestimates waiting time given the patient's values $v_{i,a}$ for each attribute $a \in \mathcal{A}_w$ and waiting area $w \in \mathcal{W}$. In this supervised learning problem, we assume the availability of labeled training data from many other patients $j \in \mathcal{I} \setminus i$ whose attribute values $v_{j,a}$ and outcomes e_j (under-, correct or overestimates) are known. This training data is used to learn a classification model.

3.1.1. Attribute Ranking and Selection Techniques

Attribute ranking techniques provide a list of attributes sorted by decreasing attribute quality. In contrast, attribute selection techniques select a classification-relevant subset of attributes. In this paper, we will evaluate both attribute ranking and attribute selection techniques.

3.1.1.1. Relief-F Attribute Ranking. Attribute ranking methods can be divided into two broad categories: Statistical and entropy-based (Novaković (2016)). The Relief-F algorithm (see Kononenko et al. Robnik-Šikonja and Kononenko (2003)) not only provides a quick estimate of relevant attributes (Gartner, Kolisch, Neill, and Padman (2015)), it also performed well in a setting where Naive Bayes (NB) was used as a classification approach (Novaković (2016)). We will use NB in our experimental analysis as well which is why evaluate Relief-F in combination with NB and other classifiers. Another advantage is that Relief-F can handle multiple values in the class attribute which we have because we take into account under-, correct or overestimation of waiting time. The result is a quality measure of each attribute Q_a which can provide an attribute ranking.

In order to describe the algorithm we first define the "k-nearest hits" and "k-nearest misses" for a sampled instance $i \in \mathcal{I}$. Let the set of k-nearest hits $\mathcal{H}_i(k) \subset \mathcal{I} \setminus i$ of an instance $i \in \mathcal{I}$ contain at most k instances $j \in \mathcal{I}, j \neq i$ which have the same

class value (e.g. overestimated waiting time) as instance i. More precisely, we choose those instances with $e_j = e_i$ which have the lowest $\mathit{diff}_{i,j}$-values as defined by Eqs. (1) and (2).

$$\mathit{diff}_{i,j} = \sum_{a \in \mathcal{A}} \mathit{diff}_{i,j,a} \tag{1}$$

$$\mathit{diff}_{i,j,a} = \begin{cases} 0, & \text{if} \quad v_{i,a} = v_{j,a} \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

Furthermore, for each class value $e \neq e_i$, let the set of k-nearest misses $\mathcal{M}_{e,i}(k) \subset \mathcal{I} \setminus i$ of instance i contain at most k instances $j \in \mathcal{I}, j \neq i$. More precisely, we choose those instances with $e_j = e$ which have the lowest $\mathit{diff}_{i,j}$-values as defined by Eqs. (1) and (2). Both the k-nearest hits and the k-nearest misses for each class value $e \in \mathcal{E}$ are used by Equation (3) which computes the quality measure Q_a for attribute $a \in \mathcal{A}$.

$$Q_a = \frac{1}{k \cdot |\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(- \sum_{h \in \mathcal{H}_i(k)} \mathit{diff}_{i,h,a} + \sum_{e \in \mathcal{E} \setminus d_i} \frac{p(e)}{1 - p(e_i)} \sum_{m \in \mathcal{M}_{e,i}(k)} \mathit{diff}_{i,m,a} \right) \tag{3}$$

For each instance $i \in \mathcal{I}$ the k-nearest hits and k-nearest misses for each sampled instance $i \in \mathcal{I}$ are selected and used in Equation (3). Then, the attributes with highest values of the quality measure are considered most relevant for classification. A detailed multi-class example is provided in Gartner (2015).

3.1.1.2. Markov Blanket Attribute Selection. With respect to attribute selection, we study Markov blanket (MB) which can be employed to model attribute-dependencies, see Saeys, Inza, and Larranaga (2007). Since MB is not capable of detecting redundant attributes, we evaluate Markov blanket attribute selection which uses conditional independence relations between the class and all other attributes. A Markov blanket is a specific Bayesian network that encodes this conditional independence in a graph. In our study, we evaluate a Markov blanket (MB) search as devised by Ramsey (2006).

3.1.1.3. Correlation-based Feature Selection. Another method to detect attribute-dependencies is correlation-based feature selection (CFS), see Saeys et al. (2007). It searches feature subsets according to the degree of redundancy among the features. The goal is to eliminate irrelevant features (Khalid, Khalil, and Nasreen (2014)) and the evaluation process aims to find subsets of features that are individually highly correlated with the class but have low inter-correlation. Intercorrelation betweeen two nominal attributes is computed via the symmetrical uncertainty between the attributes using conditional information entropies. In our case, the attribute subset $\mathcal{A}_w^*$ for waiting area $w \in \mathcal{W}$ is selected which maximizes the normalized sum of conditional symmetrical uncertainties between each attribute and the class (Hall and Holmes (2003)).

3.1.2. Classification Techniques

A machine learning classifier learns information from a dataset of labeled training examples. For each instance or individual, the true class is known to the classification method. Now, having learned the classifier's structure from the training examples, the classifier is applied to a separate dataset of unlabeled test examples. The task is to classify the true outcome of each individual patient which is unknown to the classification method.

Many machine learning classifiers and attribute selection methods have been published in the literature. Only few benchmarking studies have been performed on the evaluation of combined attribute selection, ranking and classification methods. One of these studies is Hall and Holmes (2003) who benchmarked Naive Bayes (NB), classification trees (also called decision trees) (DT), Bayesian networks (BN) and decision rules. Furthermore, they combined the classifiers with CFS and Relief-F. This benchmarking study motivated us for choosing the previously introduced attribute ranking and selection methods as well as the classifiers which are introduced next.

3.1.2.1. Naive Bayes. For our first classifier, we first learn the prior probability $p(e)$ of each class value $e \in \mathcal{E}_w$ in waiting area $w \in \mathcal{W}$ from the training data. This can be done by maximum likelihood estimation, see Gartner et al. (2015). Similarly, the conditional probability $p(v_{i,a}|e)$ represents the relative frequency of training instances that belong to class $e \in \mathcal{E}_w$ and for which $v_{i,a} = 1$. Finally, the classifier assigns the patient to class value e_i^* for which the likelihood function is maximized.

3.1.2.2. Bayesian Networks. One limitation of Naive Bayes is that each patient's attribute is only dependent on the class attribute. In a Bayesian network, however, conditional independence relations can be encoded by a graphical model in which attributes are encoded as vertices and dependencies between attributes are encoded as edges between vertices. When learning the conditional probabilities (e.g. using maximum likelihood) for each vertex, we must condition on the parents Π_a of the given attribute a in the network. This includes the class attribute and, similar to Naive Bayes, we assign the patient to class e_i^* that maximizes the posterior probability.

3.1.2.3. Classification Trees. Decision tree learners automatically learn a classification tree from labeled training data. There are various methods to learn the structure of a classification tree from data: We use a decision tree learner which has been investigated by Hall and Holmes (2003) in combination with attribute selection. We employ this algorithm because we can control the over-fitting of the classification tree as well as the tree size during the learning process using parameter optimization.

3.1.2.4. Decision Rules. Finally, a simple decision rule learner can be stated as follows (Hall and Holmes (2003)): In the training set, we count how often a value of an attribute occurred with respect to each class attribute value. For each value, we create a mapping to the most frequent class. Now, for each instance in the testing set, we assign the instance to the class which is described by the decision rule and the observed attribute value.

3.2. Patient Flow Modelling using Discrete-Event Simulation

The machine learning methods introduced in the previous subsection help us to identify relevant and non-redundant attributes that are associated with the under-, correct and overestimation of patients' waiting time. This can be important for the senior management of the ED if, for example, the ambiance turns out to be a relevant attribute which the manager may immediately want to improve. Additional relevant and non-redundant attribute may be the actual waiting time (as our experimental study will reveal). As a consequence a metric to improve for a ED manager may be the reduction of actual waiting time and, as a consequece, the number of patients who overestimate waiting times. To this end, we followed Karnon et al. (2012)'s guide to develop a discrete-event simulation model. The model is used to evaluate the perceived waiting times as a function of different shift staffing decisions. The modelling approach and arrival patterns have similarities to Crawford, Parikh, Kong, and Thakar (2013). Differences are, however, that we evaluate waiting times and perceived waiting times for staff rather than beds. Also our level of detail is higher since we track each patient's access time for each of the resources (staff and room). Figure 1 represents the patient flow observed in the hospital's ED.

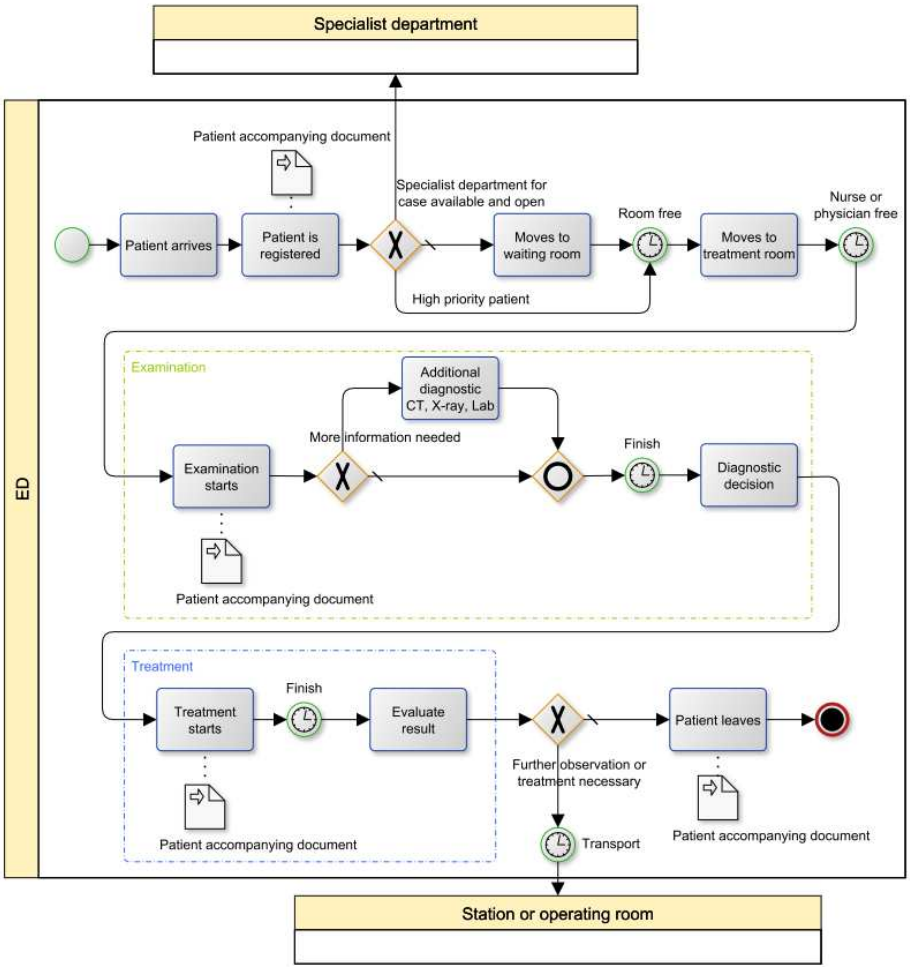

Figure 1. Patient flow observed in the hospital's ED

The figure reveals that every patient first undergoes a registration in which the triage level is determined. Afterwards, patients are assigned an examination room in which they are seen by a nurse, physician or both. Then, if necessary, patients undergo radiology diagnostics followed by a non-mandatory treatment. Finally, patients leave the ED.

4. Experimental Investigation

In the following, we provide an experimental investigation of the presented methods. We first give an overview of the data employed in our study, followed by a presentation of the attribute selection results and an evaluation of the classification techniques. We then show how we setup the DES and explain how we carried out our scenario analysis. The section closes with a breakdown of the results based on different staffing patterns.

4.1. Data and Information Documented for Classifying the Estimation Behaviour

Figure 2 shows the simplified patient flow at the ED of our collaborating hospital and how we collected the data for our study.

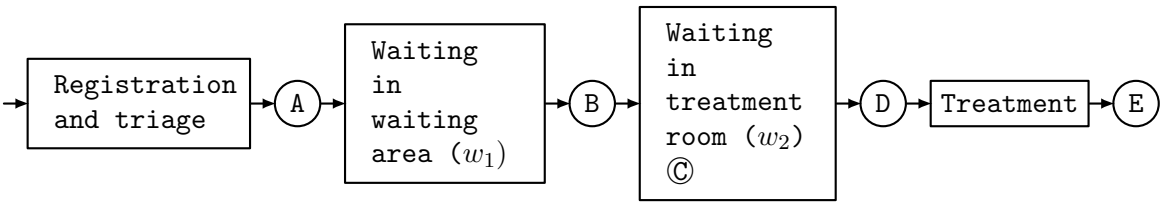

Figure 2. Simplified patient flow at the ED of the collaborating hospital where the data collection points are marked with (A)–(E)

In the ED of our collaborating hospital, patients undergo a registration process where a nurse collects demographic information about the patient (data collection point A). In addition, the registration nurse categorizes the patients using a triage system. Afterwards, the patient waits in the waiting area (w_1) until he/she is called by name in order to enter the treatment room (w_2). The time stamp when the patient is called as well as the arrival time of the patient at the treatment room are documented (data collection point B). In this room, the patient waits for the treatment by a physician, nurse or both. The patient fills out our questionnaire (data collection point C). Once the physician or nurse enter the treatment room, the time stamp is documented (data collection point D). After the treatment, the patient fills out a second questionnaire in order to collect data for the waiting time in the treatment room (data collection point E). An overview of all attributes is shown in Table A1 while a summary statistics for our data set is provided in Table 2(a).

Table 2. Summary statistics (a) and #original and selected attributes (b)

w_i	Parameter	n
1	# patients responded	228
2	# patients responded	176
1	$\#\ \lvert\{e_i \in \mathcal{E}_1 : e_i = 1\}\rvert$	91
1	$\#\ \lvert\{e_i \in \mathcal{E}_1 : e_i = 2\}\rvert$	80
1	$\#\ \lvert\{e_i \in \mathcal{E}_1 : e_i = 3\}\rvert$	57
2	$\#\ \lvert\{e_i \in \mathcal{E}_2 : e_i = 1\}\rvert$	35
2	$\#\ \lvert\{e_i \in \mathcal{E}_2 : e_i = 2\}\rvert$	31
2	$\#\ \lvert\{e_i \in \mathcal{E}_2 : e_i = 3\}\rvert$	110

(a)

w_i	original #attributes	#selected attributes using		
		Relief-F	MB	CFS
1	21	11	3	6
2	25	11	2	4

(b)

The table reveals that in waiting area 1 and 2, the number of patients who responded was 228 and 176, respectively. Of these patients, some only answered questions in area 1 while others only answered questions in area 2. This is why the total number of patients sums up to more than 250. A more detailed view is provided in Figures 3(a)-(b). They show the breakdown of how many patients answered the question of "How long do you estimate your waiting time in the waiting area?" while intervals range from "0-2", "3-5", "6-10", "11-15", "16-30", "31-45", "46-60", "61-90", "91-120", "121-150", ">150" minutes encoded as integers from 1 to 11. We call this perceived waiting time (PWT) and the same coding has been used for encoding the actual waiting time (AWT).

AWT ↓/ PWT →	1	2	3	4	5	6	7	8	9	10	11	Totals
1	15	4		1	1	1					2	24
2	7	8	6	1	4	1		1				28
3	7	8	6		5	3	1					30
4	1	3	5	3	4	3						19
5		4	10	8	21	5	1		1	1		51
6		1	1	5	8	8	3		1	1		28
7		1			1	3	5	2	3			15
8					3	2	4	7				16
9						1	1	4	5		1	12
10		1						1	1	1		4
11											1	1
Totals	30	30	28	18	47	27	15	15	11	3	4	228

(a)

AWT ↓/ PWT →	1	2	3	4	5	6	7	8	9	10	11	Totals
1	14	8	10	2	11	8	2	2	4		1	62
2	17	9	9	6	8	5	2	3	2			61
3	6	2	1	1	5	5	2	2		1	1	26
4		1	1	2	3	1	1	1				10
5	2	1	1	2	2	1	1		1			11
6					1	3					1	5
8					1							1
Totals	39	21	22	13	31	23	8	8	7	1	3	176

(b)

Figure 3. Actual (AWT) and perceived waiting time (PWT) as responded by patients in the waiting area w_1 (a) and the treatment room w_2 (b)

The figures show, for example, that values above the coloured diagonal in Figure 3(a) sum up to 57 which is exactly the number of individuals we labeled as overestimators (see Table 2(a)). Another observation in Figure 3(b) is that none of the patients waited more than 90 minutes in waiting area 2 which is encoded by interval number 8.

4.2. Attribute Ranking, Selection and Classification Results

4.2.1. Attribute Ranking Results

Table 3 provides the results of the attribute rankings where the nearest-neighbor parameter of the Relief-F algorithm is set to $k = 10$, as suggested by Robnik-Šikonja and Kononenko (2003).

Table 3. Results of the top ten attributes determined by Relief-F for waiting area w_1(a) and w_2(b)

Rank	Attribute name	Q_a
1	Waiting time in waiting area w_1	0.0369
2	How was your perception on waiting time in this area (w_1)?	0.0357
3	While waiting, did you look at your watch?	0.0095
4	Do you agree that patients who arrive after you are treated before you?	0.0088
5	Did you have company while waiting (Friend, relative)?	0.0086
6	Age	0.0080
7	The ambiance of the waiting area is pleasant.	0.0077
8	The staff informed me about my current waiting situation.	0.0066
9	Type of admission	0.0045
10	Type of arrival	0.0033
11	Under-, correct and overestimation of waiting time in waiting area w_1	

(a)

Rank	Attribute name	Q_a
1	Waiting time in waiting area w_2	0.0636
2	In this waiting area (w_2) I felt calm and unhurried.	0.0382
3	How was your perception on waiting time in this area (w_2)?	0.0354
4	Treatment time	0.0302
5	While waiting, did you look at your watch?	0.0288
6	How long do you estimate your waiting time in waiting area w_1?	0.0144
7	The staff informed me about my current waiting situation.	0.0061
8	Type of admission	0.0058
9	Under-, correct and overestimation of waiting time in waiting area w_1	0.0054
10	Waiting time in waiting area w_1	0.0043
11	Under-, correct and overestimation of waiting time in waiting area w_2	

(b)

Both attribute rankings, in waiting area w_1 and w_2, indicate that the attribute "waiting time" and the "waiting time perception" are among the top three ranked attributes. One explanation for this phenomenon is that waiting time perceptions have a direct influence on the under-, correct and overestimation of waiting time as our literature review revealed. Another important attribute as revealed by the attribute ranking is whether or not the staff informed the patient about the waiting situation. This is important because the direct implication of our study is that the emergency department can control under-, correct and overestimation of waiting times by informing the patient about the waiting situation. We expect this result because it confirms previous studies such as Antonides et al. (2002), Hong et al. (2013) and Thompson, Yarnold, Adams, and Spacone (1996).

4.2.2. Attribute Selection Results

The results from the CFS attribute selection are shown in Table 4.

Table 4. CFS results for waiting area w_1(a) and w_2(b)

Attribute name
Triage
Waiting time in waiting area w_1
In this waiting area (w_1) I felt calm and unhurried.
The staff informed me about my current waiting situation.
How was your perception on waiting time in waiting area (w_1)?

(a)

Attribute name
Waiting time in waiting area w_1
Waiting time in waiting area w_2
How was your perception on waiting time in this area (w_2)?

(b)

The results reveal that, in waiting area 1, five attributes were selected. The CFS results for waiting area w_2 reveal only three relevant and non-redundant attributes.

Among the attributes selected by the algorithm of Ramsey (2006), with the conditional independence tests at a significance level of 0.05 and search depth 1, the attributes "Waiting time in waiting area w_1" and "Did you occupy yourself with other things while waiting in the waiting area?" were selected for waiting area w_1. The selection of the waiting time attribute is consistent with Relief-F and CFS. For waiting area w_2, only the attribute "Waiting time in waiting area w_2" was selected. This is included in CFS, too and exactly the top attribute in the Relief-F attribute ranking for that waiting area.

Table 2(b) summarizes the attribute selection part where a comparison between the original number of attributes and the number of selected attributes is provided and broken down by waiting area. The original number of attributes used for waiting area w_1 comes from data collection point A, B, C and D . The original number of attributes used for waiting area w_2 has additional information from data collection E, see Figure 2 and Table A1 in the appendix.

4.2.3. Parameter Optimization for the Decision Tree Learner

We performed a parameter optimization for the decision tree approach and varied the minimum number of instances per leaf (MI) within the interval $[2, 15]$ for w_1 and w_2. The confidence factor (CF) is varied using the values 0.01 to 0.5 with 100 steps. The parameter combination which results in the maximum accuracy on the testing set is selected. Table 5 shows the results.

The results reveal that for waiting area w_2, the confidence factors are higher as compared to waiting area w_1. Similarly, the number of instances per leaf for waiting area w_2 is greater than or equal to the number of instances per leaf for waiting area w_1.

4.2.4. Classification Results for Waiting Area w_1

All classifiers are assessed using the same performance indicators. The overall performance is measured in terms of classification accuracy (proportion of correctly

Table 5. Optimized confidence factors (a) and minimum instances per leaf (b) for the decision tree learner broken down by waiting areas

	Attribute selection method			
w_i	w/o	Relief-F	MB	CFS
1	0.07	0.13	0.38	0.19
2	0.43	0.41	0.46	0.37

(a)

	Attribute selection method			
w_i	w/o	Relief-F	MB	CFS
1	2	2	2	2
2	4	10	2	3

(b)

classified patients) as well as precision (proportion of cases classified as belonging to the true "underestimation" and "overestimation" outcome that are correctly classified). The mathematical expression of how underestimation precision is calculated is: $\frac{TP(e_w=1)}{TP(e_w=1)\cdot FP(e_w=1)}$. Similarly, overestimation precision is calculated as $\frac{TP(e_w=2)}{TP(e_w=2)\cdot FP(e_w=2)}$. Finally, we report the area under the ROC curve of the "underestimation" and "overestimation" outcome. For example, to calculate the overestimation ROC area, we have to calculate the pairwise AUCs for outcome 0 and 1 as well as 0 and 2. For example, for outcome 0 and 1 this becomes $AUC_{0,1} = \frac{1}{m\cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{1}_{p_i>p_j}$. Here, i runs over all m data points with true label 1, and j runs over all n data points with true label 0. p_i and p_j denote the probability score assigned by the classifier to data point i and j, respectively. $\mathbb{1}_{p_i>p_j}$ is the indicator function: it outputs 1 iff $p_i > p_j$. Finally, we average across the two AUCs. All performance indicators are measured using 10-fold cross-validation and the training and test sets were separated randomly. We chose to use accuracy, precision and area under the ROC because they are standard metrics which have been used to benchmark supervised learning algorithms (Caruana and Niculescu-Mizil (2006)).

Table 6(a) shows the results using overall accuracy as metric and reveals that attribute selection can improve classification accuracy from 46.9% to 51.3%, comparing the highest accuracies of all classifiers. A detailed analysis of the decision rule learner revealed that the attribute "Waiting time in waiting area w_1" was always chosen for creating the decision rules. The decision tree branches which are linked with the waiting time output of the discrete-event simulation model are as follows:

$$e_i^* = \begin{cases} 2 \text{ (overestimate) if wait time in } w_1 \text{ is 0 to 2 minutes} \\ 2 \text{ (overestimate) if wait time in } w_1 \text{ is 3 to 5 minutes} \\ 1 \text{ (correct estimate) if wait time in } w_1 \text{ is 6 to 10 minutes} \\ 2 \text{ (overestimate) if wait time in } w_1 \text{ is 11 to 15 minutes} \\ 1 \text{ (correct estimate) if wait time in } w_1 \text{ is 16 to 30 minutes} \\ 0 \text{ (underestimate) if wait time in } w_1 \text{ is 31 to 45 minutes} \\ 1 \text{ (correct estimate) if wait time in } w_1 \text{ is 46 to 60 minutes} \\ 1 \text{ (correct estimate) if wait time in } w_1 \text{ is 61 to 90 minutes} \\ 1 \text{ (correct estimate) if wait time in } w_1 \text{ is 91 to 120 minutes} \\ 0 \text{ (underestimate) if wait time in } w_1 \text{ is 121 to 150 minutes} \\ 1 \text{ (correct estimate) if wait time in } w_1 \text{ is more than 150 minutes} \end{cases}$$

The results with underestimation precision as a metric, shown in Table 6(b) reveal

that without attribute selection, classifying patients as ones who underestimate their waiting time as true outcome ($e_i = 0$, see Section 3) is 57.4% using BN. The results with overestimation precision as a metric, shown in Table 6(c) reveal that with CFS attribute selection, classifying patients as ones who overestimate their waiting time as true outcome ($e_i = 2$, see Section 3) is 69.7% using DT.

The results using the underestimation ROC area as metric come up to 72.4% using BN and CFS attribute selection (see Table 6(d)). Interestingly, using CFS can boost the ROC area for all classification approaches not just compared to the results without attribute selection but also with respect to Relief-F or MB. The results using overestimation ROC area as metric, shown in Table 6(e) reveal that, again, with CFS attribute selection, the ROC area can be increased.

Table 6. Overall accuracy (a), underestimation precision (b), overestimation precision (c), underestimation ROC area (d) and overestimation ROC area (e) for waiting area w_1.

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	46.1	50.0	42.5	49.6
BN	46.9	49.6	42.5	49.1
DT	46.5	47.4	40.8	**51.3**
Rule	43.4	43.4	43.4	43.4

(a)

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	56.7	55.7	44.6	54.6
BN	**57.4**	54.7	44.6	55.8
DT	48.9	51.3	43.9	51.5
Rule	44.3	44.3	44.3	44.3

(b)

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	33.3	48.9	40.6	48.5
BN	35.1	53.7	39.4	48.5
DT	42.4	48.8	36.8	**69.7**
Rule	43.3	43.3	43.3	43.3

(c)

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	64.5	66.8	58.4	72.0
BN	65.4	67.5	58.9	**72.4**
DT	64.6	65.6	59.8	67.3
Rule	57.7	57.7	57.7	57.7

(d)

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	65.9	69.3	60.7	71.0
BN	65.8	69.8	61.1	**71.2**
DT	60.7	65.5	60.5	65.7
Rule	56.0	56.0	56.0	56.0

(e)

4.2.5. Classification Results for the Treatment Room w_2

Again, we will break down the results by our five evaluation metrics while all performance indicators are measured using 10-fold cross-validation and the training and test sets were separated randomly. Table 7(a) shows the results using overall accuracy as metric and reveal that attribute selection can improve classification accuracy from 66.5% to 69.9%, comparing the highest accuracies of all classifiers. Again, analyzing the decision rule learner in more detail revealed that now the attribute "Waiting time in waiting area w_2" was chosen. The branches of the decision tree which are relevant for the connection with the actual waiting time output of the discrete-event simulation model are as follows:

$$e_i^* = \begin{cases} 1 \text{ (correct estimate) if wait time in } w_2 \text{ is 0 to 2 minutes} \\ 0 \text{ (underestimate) if wait time in } w_2 \text{ is 3 to 5 minutes} \\ 2 \text{ (overestimate) if wait time in } w_2 \text{ is 6 to 10 minutes} \\ 2 \text{ (overestimate) if wait time in } w_2 \text{ is 11 to 15 minutes} \\ 0 \text{ (underestimate) if wait time in } w_2 \text{ is 16 to 30 minutes} \\ 1 \text{ (correct estimate) if wait time in } w_2 \text{ is 31 to 45 minutes} \\ 1 \text{ (correct estimate) if wait time in } w_2 \text{ is 46 to 60 minutes} \\ 0 \text{ (underestimate) if wait time in } w_2 \text{ is 61 to 90 minutes} \end{cases}$$

The results with underestimation precision as a metric, shown in Table 7(b) reveal that without attribute selection, classifying patients as ones who underestimate their waiting time as true outcome ($e_i = 0$, see Section 3) is 61.5% using DT. The results with overestimation precision as a metric, shown in Table 7(c) reveal that without attribute selection, classifying patients as ones who overestimate their waiting time as true outcome ($e_i = 2$, see Section 3) is 77.6% using Bayesian approaches.

The results using underestimation ROC area as a metric, shown in Table 7(d), reveal that CFS can boost the area under the ROC curve to 85.1%. The results using overestimation ROC area as metric, shown in Table 7(e) reveal that, similar to the improved area under the ROC curve of the underestimation results, CFS boosts the area under the ROC curve for overestimating waiting times.

Table 7. Overall accuracy (a), underestimation precision (b), overestimation precision (c), underestimation ROC area (d) and overestimation ROC area (e) for waiting area w_2.

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	66.5	65.3	65.3	68.2
BN	66.5	63.6	65.3	66.5
DT	65.9	64.2	62.5	**69.9**
Rule	61.4	62.5	65.9	65.9

(a)

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	60.0	55.6	54.5	56.7
BN	60.0	53.6	54.5	53.3
DT	**61.5**	43.2	37.5	60.0
Rule	38.5	50.0	58.3	58.3

(b)

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	**77.6**	74.4	66.3	73.7
BN	**77.6**	74.8	66.3	72.7
DT	74.6	73.1	64.2	73.3
Rule	64.2	64.8	66.7	66.7

(c)

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	80.4	65.9	75.2	**85.1**
BN	81.1	66.2	75.6	85.0
DT	81.3	58.7	69.6	84.2
Rule	55.2	57.6	58.6	58.6

(d)

Classifier	Attribute selection method			
	w/o	Relief-F	MB	CFS
NB	69.4	**75.4**	44.2	72.9
BN	69.1	75.1	43.7	72.6
DT	66.5	73.6	42.5	58.5
Rule	49.7	51.3	52.7	52.7

(e)

4.3. Simulation Model Setup, Validation and Results

We implemented a simulation model in Rockwell Arena (Wang, Guinet, Belaidi, and Bescombes (2009)) which represents the patient flow from the collaborating hospital shown in Figure 1 in Section 3.2. In the following, we will describe the parameters of the arrival and service process. Also, simulation parameters such as the number of replications are provided.

4.3.1. Arrival Process

The arrival pattern during one day is shown in Figure 4.

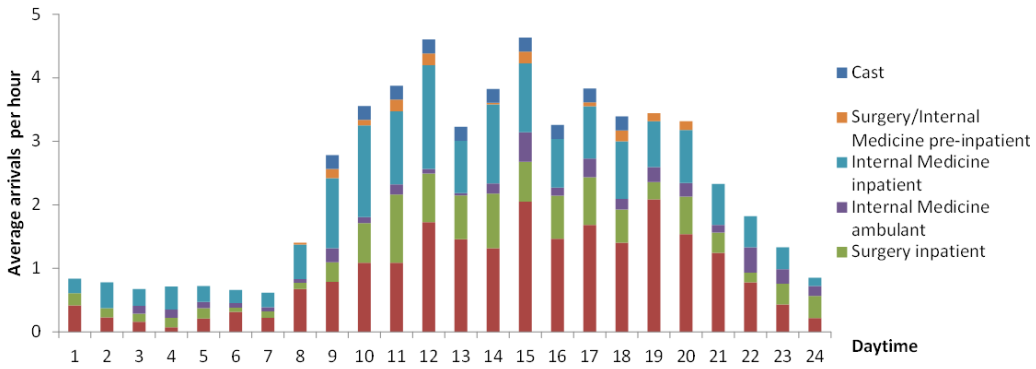

Figure 4. Patient arrivals broken down by time of day

The resource availabilities of medical staff is given in Table 8.

	Nurses	Surgeons	Internists
Resource 1	0am - 0pm	0am - 0pm	0am - 0pm
Resource 2	8am - 0pm	4am - 8am	4am - 8am
Resource 3	8am - 4pm		
Resource 4	1pm - 2:30pm		

Table 8. Staff schedule

We carried out a data analysis in order to obtain distributions for the arriving patients. All distributions and the distribution parameters were obtained using the Kolmogorov-Smirnof test (Hartung, Elpelt, and Klösener (1999)). Table 9 provides an overview of the distributions and their parameters for the triage assignment, X-Ray and CT resource requirement. We broke these distributions down by surgical and internal medicine patients.

Parameter	Surgical		Internal medicine	
	ambulatory	inpatient	ambulatory	inpatient
Triage	DISC(0.11,3,1.0,1)	DISC(0.4,4,1.0,2)	DISC(0.33,3,1.0,1)	DISC(0.61,4,1.0,2)
X-Ray	DISC(0.53,0,1.0,1)	DISC(0.4,0,1.0,1)	DISC(0.87,0,1.0,1)	DISC(0.37,0,1.0,1)
CT	DISC(0.99,0,1.0,1)	DISC(0.83,0,1.0,1)	DISC(0.99,0,1.0,1)	DISC(0.8,0,1.0,1)

Table 9. Distributions and their parameters for triage and radiology resources

For example, DISC(0.11,3,1.0,1) means that we have a discrete distribution in which 11% of the arriving ambulatory patients are assigned to triage category 3 while the rest of the patients in this group are assigned to triage category 1. The results for estimating the service time distributions are shown in Table 10.

Patient type	Examinations	
	Nurse	Physician
Surgical patient	WEIB(0.494, 0.965)	BETA(0.754, 1.62207)
Internal medicine patient	1.97*BETA(0.999, 2.09)	BETA(1.02, 1.63319)

Table 10. Average examination durations [minutes]

For example, WEIB(0.494, 0.965) means that we have a scale of $\lambda = 0.494$ and a shape of $k = 0.965$. Besides these parameters, we set the X-Ray and CT examination durations to an ERLA(0.0391, 3) and 0.1+EXPO(0.2) distribution, respectively.

4.4. Replication Length, Number of Replications and Warmup Time

Having implemented the patient flow logic and having determined the distributions for the arrivals and service durations, we now have to determine the following three parameters for running the simulation experiments: Replication length, replication number and warm up time. For the replication length we run 24 hours from 0:00 a.m. until 12:00 p.m. To determine the replication number R_θ for resource type θ based on a sample standard deviation of the waiting time $S = 50$, samples for resource type θ and a half-width of $\epsilon = 10\%$, we employed the following equation: $R_\theta \geq \left(\frac{z_{\frac{\alpha}{2}} \cdot S}{\epsilon}\right)$ (Banks, Carson, Nelson, and Nicol (2001)). In doing so, the replication numbers come up to $R_{\text{room}} = 414$, $R_{\text{physician}} = 100$ and $R_{\text{Nurse}} = 168$ for the room, physician and nurse waiting times, respectively. We decided to use the maximum and rounded up to a replication number of $R^* = 500$.

4.5. Validation of the Simulation Model

The results of the simulation model validation are provided in Figure 5. The figure reveals that the waiting time obtained by surveying patients in the hospital's ED and the ones reported from the simulation model correlate very well. On average, the difference between the waiting time determined by simulation and the empirical waiting times is approximately 5.0%.

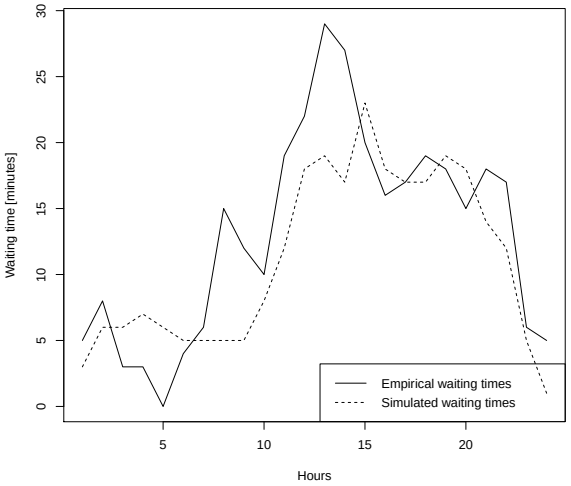

Figure 5. Simulation model validation

4.6. Scenario Analysis of Different Shift Schedules

To provide recommendations how patients who overestimate waiting times can be reduced or balanced across the w_1 and w_2, we performed simulation runs with nine different scenarios. The parameters changed are the personnel resources: Nurses, surgeons and internists. According to the collaborating hospital and because of their system to schedule staff, these are the only feasible workforce changes. All scenarios are provided by Table 11.

Scenario	Nurses	Surgeons	Internists
Scenario 1	4pm - 10pm		
Scenario 2a		8am - 4pm	
Scenario 2b			8am - 4pm
Scenario 3a		12(noon) - 4pm	8am - 12(noon)
Scenario 3b		8am - 12(noon)	12(noon) - 4pm
Scenario 4a		12(noon) - 4pm	
Scenario 4b			8am - 12(noon)
Combination 1	4pm - 10pm	12(noon) - 4pm	
Combination 2	4pm - 10pm	12(noon) - 4pm	8am - 12(noon)

Table 11. Additional Personal Resources per Scenario

For example, in scenario 1, 2a and 2b, one additional employee is staffed for six and eight hours for nurses and physicians, respectively. Similarly, 4a and 4b take one new physician for a four hour shift into account. The other scenarios consider a multiple

resource increase where combinations 1 and 2 combine scenarios previously introduced. Combination 1 (C1) is composed of scenario 1 and 4a whereas combination 2 (C2) covers the changes of option 1 and 3a.

Table 12 shows the simulation results where the total average waiting time, the waiting time in each waiting area and the average number of patients who overstimate waiting time as classified by the decision tree learner are reported and broken down by each of the scenarios.

	Avg. Waiting Time [h]			Avg. # Overestimaters	
	Total	w_1	w_2	w_1	w_2
Base case	0.3046	0.2469	0.0577	1.9	10.9
1	0.2289	0.1899	0.0389	2.0	10.9
2a	0.2054	0.1225	0.0828	1.4	23.1
2b	0.2438	0.1904	0.0533	1.4	12.8
3a	0.1901	0.1224	0.0676	1.4	19.6
3b	0.2291	0.1598	0.0692	**1.2**	18.6
4a	0.2190	0.1261	0.0928	1.4	32.3
4b	0.2668	0.2120	0.0547	1.4	12.8
C1	0.1583	0.1115	0.0467	1.8	10.9
C2	**0.1277**	**0.1006**	**0.0270**	2.2	**6.6**

Table 12. Changes to Base Case - Results

The figures reveal that scenario C2 reduces the actual waiting times from 0.3046 hours to 0.1277 hours, on average. Also, waiting times in both areas drop substantially using the staffing levels in this scenario. Another observation is that this scenario performs best for reducing patients who overestimate waiting times in the treatment room. However, if reducing the number of patients in the waiting area is the goal, then scenario 3b performs better. A more detailed analysis of the comparison between scenario 1 and the base case shows that adding more nurses does not substantially change the waiting time overestimation in both areas. However, when adding more human resources, the impact is that patients are pulled from the waiting area into the rooms because patients are seen quicker. As a consequence, the average waiting time in w_2 may increase and thus overestimation of waiting time increases which is shown in scenario 4a.

5. Summary and Conclusions

In this paper, we have evaluated attribute selection, classification techniques and discrete-event simulation modelling to understand and influence patients' behaviour to under-, correctly and overestimate waiting times in an Emergency Department. We have shown that the set of patient attributes can be reduced to a set of highly rele-

vant ones. These attributes can be divided into two categories – those associated with objectively collected data such as actual waiting time, and those based on subjective information such as perceived waiting time, indicating the significance of behavioural data in understanding and improving individuals' services (White (2016)) provided by clinicians vs. automated systems. Using the selected subset of attributes, we have compared four different classification techniques on overall accuracy, precision and ROC area for the true outcome under-, correctly and overestimating waiting times. We broke down our analysis by focusing on two waiting areas. While precision for overestimating waiting time is approximately 70% in the waiting area of the ED, the precision of overestimating waiting time in the treatment room yields approximately 78%. Linking the decision tree learner with a discrete-event simulation model revealed that not only actual waiting times can be evaluated using different staffing patterns but also the number of patients who overestimate waiting times can be influenced and ultimately reduced in the different waiting areas.

Our attribute selection results demonstrate that providing information for individual patients on their remaining waiting time is important. This could be implemented into the emergency department's computer system by a screen that automatically shows and updates the estimate on the remaining waiting time. Another area of future work is to incorporate additional behavioural, contextual and cognitive factors into the set of attributes, machine learning and DES approaches.

Acknowledgement

The authors sincerely thank the anonymous referees for their careful review and excellent suggestions for improvement of this paper.

Appendix A. Attributes evaluated

Table A1 provides a detailed overview about all attributes available for our study. The data was collected at the five data collection points shown in Figure 2. Two questionnaires were created and used for data collection points C and E. The data which was collected at point A was accessed through the IT-based hospital information system. Physicians and nurses manually documented time stamps at point B and D. In both waiting areas, clocks were removed; however, patients were allowed to wear their watches. Note that for waiting area w_2 we didn't collect information about 'The contact with the staff in this waiting area was nice.' because patients waiting in the treatment room cannot interact with staff that is responsible for the management of the waiting area. The same holds true for question/attribute: 'The staff informed me about my waiting situation.' In addition, we have chosen not to collect data about 'Did you occupy yourself with other things while waiting in the treatment room?' because the patient fills out the questionnaire in that room and therefore, we expect biased answers because of filling out our questionnaire.

Attribute	**Data type**	**Distinct attribute values or bins**	**Collected at data collection point (see Figure 2)** A	B	C	D	E
Age	nominal	10 (e.g. 0–9.6 years)	✓				

Attribute	**Data type**	**Distinct attribute values or bins**	**Collected at data collection point (see Figure 2)** A	B	C	D	E
Gender	nominal	2 (male, female)	✓				
Health insurance type	nominal	2 (private, statutory)	✓				
Specialty	nominal	2 (surgical, internal medicine)	✓				
Triage level	nominal	8 (e.g. 3 – urgent)	✓				
Type of admission	nominal	2 (outpatient, inpatient)	✓				
Type of arrival	nominal	2 (ambulance, walk-in)	✓				
Weekday	nominal	7 e.g. Monday	✓				
Waiting time in waiting area w_1	nominal	11 (e.g. 0–2 minutes)		✓			
Did you have company while waiting in the waiting area?	nominal	2 (yes or no)			✓		
Did you look at your watch while waiting in the waiting area?	nominal	2 (yes or no)			✓		
Did you occupy yourself with other things while waiting in the waiting area?	nominal	2 (yes or no)			✓		
Do you agree that patients who arrive after you are treated before you?	nominal	2 (yes or no)			✓		
How long do you estimate your waiting time in waiting area?	nominal	11 (e.g. 0–2 minutes)			✓		
How was your perception on waiting time in waiting area (w_1)?	nominal	5 (e.g. 'very fast')			✓		
In the waiting area, I felt calm and unhurried.	nominal	5 (e.g. completely agree)			✓		
The ambiance in the waiting area is pleasant.	nominal	5 (e.g. completely agree)			✓		
The contact with the staff in the waiting area was nice.	nominal	5 (e.g. completely agree)			✓		
The staff informed me about my waiting situation.	nominal	5 (e.g. completely agree)			✓		
Treatment time	nominal	10 e.g. (0–5 minutes)				✓	
Waiting time in the treatment room	nominal	11 (e.g. 0–2 minutes)				✓	
Did you look at your watch while you were waiting in the treatment room?	nominal	2 (yes or no)					✓

Attribute	Data type	Distinct attribute values or bins	Collected at data collection point (see Figure 2) A	B	C	D	E
How long do you estimate your waiting time in the treatment room?	nominal	11 (e.g. 0–2 minutes)					✓
How was your perception on waiting time in the treatment room?	nominal	5 (e.g. very fast)					✓
In the treatment room, I felt calm and unhurried.	nominal	5 (e.g. completely agree)					✓

Table A1.: Attributes assessed for classifying under-, correct and overestimation of waiting time

References

Antonides, G., Verhoef, P. C., & van Aalst, M. (2002). Consumer perception and evaluation of waiting time: A field experiment. *Journal of Consumer Psychology*, *12*(3), 193–202.

Banks, J., Carson, J. S., Nelson, B. L., & Nicol, D. (2001). *Discrete-event system simulation* (Vol. 3). Prentice-Hall International Series in Industrial and Systems Engineering, Upper Saddle River, New Jersey.

Boudreaux, E. D., & O'Hea, E. L. (2004). Patient satisfaction in the emergency department: A review of the literature and implications for practice. *The Journal of Emergency Medicine*, *26*(1), 13–26.

Brailsford, S., & Schmidt, B. (2003). Towards incorporating human behaviour in models of health care systems: An approach using discrete event simulation. *European Journal of Operational Research*, *150*(1), 19-31.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168).

Crawford, E. A., Parikh, P. J., Kong, N., & Thakar, C. V. (2013). Analyzing discharge strategies during acute care a discrete-event simulation study. *Medical Decision Making*, *34*(2), 231-241.

Fügener, A., Schiffels, S., & Kolisch, R. (2017). Overutilization and underutilization of operating rooms - insights from behavioral health care operations management. *Health Care Management Science*, *20*(1), 115–128.

Gartner, D. (2015). Scheduling the hospital-wide flow of elective patients. *Springer Lecture Notes in Economics and Mathematical Systems*. (Heidelberg)

Gartner, D., Kolisch, R., Neill, D. B., & Padman, R. (2015). Machine learning approaches for early DRG classification and resource allocation. *INFORMS Journal on Computing*, *27*(4), 718–734.

Hall, M., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, *15*(6), 1437-1447.

Hartung, J., Elpelt, B., & Klösener, K.-H. (1999). *Statistik: Lehr- und handbuch der angewandten statistik* (Vol. 12). Oldenbourg, München.

Hedges, J. R. (2002). Satisfied patients exiting the emergency department (speed) study. *Academic Emergency Medicine*, *9*(1), 15–21.

Hong, W., Hess, T. J., & Hardin, A. (2013). When filling the wait makes it feel longer: a paradigm shift perspective for managing online delay. *Mis Quarterly*, *37*(2), 383–406.

Karnon, J., Stahl, J., Brennan, A., Caro, J. J., Mar, J., & Möller, J. (2012). Modeling using discrete event simulation a report of the ISPOR-SMDM modeling good research practices task force. *Medical Decision Making*, *32*(5), 701–711.

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *Science and information conference (sai), 2014* (pp. 372–378).

Landi, S., Ivaldi, E., & Testi, A. (2018). Socioeconomic status and waiting times for health services: An international literature review and evidence from the italian national health system. *Health Policy*, *122*(4), 334-351.

Luo, W., J. Liberatore, M., L. Nydick, R., B. Chung, Q., & Sloane, E. (2004). Impact of process change on customer perception of waiting time: A field study. *Omega*, *32*(1), 77–83.

Nanda, U., Chanaud, C., Nelson, M., Zhu, X., Bajema, R., & Jansen, B. H. (2012). Impact of visual art on patient behavior in the emergency department waiting room. *The Journal of Emergency Medicine*, *43*(1), 172–181.

Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, *21*(1).

Ramsey, J. (2006). *A PC-style Markov blanket search for high dimensional datasets* (Tech. Rep.). Carnegie Mellon University.

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, *53*(1), 23-69.

Sadat Zadeh, S. M. T., Anwar, T., & Basirat, M. (2012). A survey on application of artificial intelligence for bus arrival time prediction. *Journal of Theoretical and Applied Information Technology*, *46*(1), 516-525.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517.

Shaikh, S. B., Witting, M. D., Winters, M. E., Brodeur, M. N., & Jerrard, D. A. (2013). Support for a waiting room time tracker: A Survey of patients waiting in an urban ED. *The Journal of Emergency Medicine*, *44*(1), 225–229.

Shunko, M., Niederhoff, J., & Rosokha, Y. (2018). Humans are not machines: The behavioral impact of queueing design on service time. *Management Science*, *64*(1), 453-473.

Soremekun, O. A., Takayesu, J. K., & Bohan, S. J. (2011). Framework for analyzing wait times and other factors that impact patient satisfaction in the emergency department. *The Journal of Emergency Medicine*, *41*(6), 686–692.

Thompson, D. A., Yarnold, P. R., Adams, S. L., & Spacone, A. B. (1996). How accurate are waiting time perceptions of patients in the emergency department? *Annals of Emergency Medicine*, *28*(6), 652–656.

Thompson, D. A., Yarnold, P. R., Williams, D. R., & Adams, S. L. (1996). Effects of actual waiting time, perceived waiting time, information delivery, and expressive quality on patient satisfaction in the emergency department. *Annals of Emergency Medicine*, *28*(6), 657–665.

Wang, T., Guinet, A., Belaidi, A., & Bescombes, B. (2009). Modelling and simulation of emergency services with ARIS and Arena. Case study: The emergency department of Saint Joseph and Saint Luc Hospital. *Production Planning & Control*, *20*(6), 484-495.

Welch, S. J. (2009). Twenty years of patient satisfaction research applied to the emergency department: A qualitative review. *American Journal of Medical Quality*, *25*(1), 64–72.

White, L. (2016). Behavioural operational research: Towards a framework for understanding behaviour in or interventions. *European Journal of Operational Research*, *249*(3), 827-841.