Quality Technology & Quantitative Management

ISSN: (Print) 1684-3703 (Online) Journal homepage: www.tandfonline.com/journals/ttqm20

Conditional Value at Risk as a Measure for Waiting Time in Simulations of Hospital Units

Christian Dehlendorff, Murat Kulahci, Soren Merser & Klaus Kaae Andersen

To cite this article: Christian Dehlendorff, Murat Kulahci, Soren Merser & Klaus Kaae Andersen (2010) Conditional Value at Risk as a Measure for Waiting Time in Simulations of Hospital Units, Quality Technology & Quantitative Management, 7:3, 321-336, DOI: 10.1080/16843703.2010.11673235

To link to this article: https://doi.org/10.1080/16843703.2010.11673235

Quality Technology & Quantitative Management
Vol. 7, No. 3, pp. 321-336, 2010

QTQM

Conditional Value at Risk as a Measure for Waiting Time in Simulations of Hospital Units

Christian Dehlendorff[1], Murat Kulahci[1], Søren Merser[2] and Klaus Kaae Andersen[1]

[1]Informatics and Mathematical Modelling, Section for Statistics, Technical University of Denmark, Denmark
[2]Clinic of Orthopaedic Surgery, Frederiksberg Hospital, Denmark
(*Received* November 2008, *accepted* June 2009)

Abstract: The utility of conditional value at risk (*CVaR*) of a sample of waiting times as a measure for reducing long waiting times is evaluated with special focus on patient waiting times in a hospital. *CVaR* is the average of the longest waiting times, *i.e.*, a measure at the tail of the waiting time distribution. The presented results are based on a discrete event simulation (DES) model of an orthopedic surgical unit at a university hospital in Denmark. Our analysis shows that *CVaR* offers a highly reliable performance measure. The measure targets the longest waiting times and these are generally accepted to be the most problematic from the points of view of both the patients and the management. Moreover, *CVaR* can be seen as a compromise between the well known measures: average waiting time and the maximum waiting time.

Keywords: Conditional value at risk, health care, simulation, waiting time distribution.

1. Introduction

Simulation studies are widely used in health care applications due to the large number of uncertainties involved. The complexity of these systems together with the physical and legal constraints in the actual systems make simulation a very powerful tool for experimentation to serve as a basis for analytic optimization methods [4, 9].

Simulation models in health care applications are used both for optimization of existing facilities [8] and in planning new facilities [18]. Ferrin and McBroom [8] maximized hospital revenue by process improvements in the emergency departments. Length of stay (*LOS*), the number of patients leaving without receiving care, the percentage of admissions accepted and ambulance diversion hours were used as outcomes. Miller *et al.* [18] considered the merging of six emergency departments into one and focused on the average *LOS*. Their results show that the *LOS* can indeed be considerably reduced. They further show that the distribution of *LOS* is right-skewed with a long tail. Jun *et al.* [14] reviewed the health care simulation literature and concluded that simulation is often used to optimize allocations and as a tool in staff planning. They cited various studies related to patient scheduling and to staff sizing and planning. They also reported that many studies use trade-offs between the utilization of doctors, rooms etc. and patients' waiting times as outcomes.

Denton *et al.* [7] studied expected surgical suite waiting time, surgical suite idle time and total overtime and used a linear trade-off combination of these measures as a single measure. This linear combination is a cost measure which takes into account the discomfort of patient waiting time and considers it together with the lost revenue corresponding to idle surgical suite time and the cost of overtime.

Cayirli and Veral [5] reviewed out-patient scheduling and summarized a number of possible performance measures related to the quality of such systems. The time-based measures included the mean, the maximum and the frequency distribution of the waiting times. Their summary for the suggested performance measures showed that the majority of studies used mean waiting time, total costs of waiting, percentage of patient waiting less than a certain threshold, and the variation of waiting time.

The main objective in this article is to compare Conditional Value at Risk (*CVaR*) as a optimization measure for patients' waiting time with existing measures and to report on the performance of this new measure based on a specific case-study of an orthopedic surgical unit. The concept of *CVaR* is formally introduced in section 3.1 and originates from economics. *CVaR* was introduced by Rockafellar and Uryasev [21] as a measure to quantify a distribution of losses; typically in portfolio scenarios. The measure was introduced as an extension to Value at Risk (*VaR*), one of the most commonly used performance measures in portfolio management. The *CVaR* criterion focuses on the right tail of the loss distribution and provides a measure of the expected value of the highest losses. The *CVaR* criterion has been used in a wide variety of applications (see for example [1, 10] and [27]), but not in the context of our study. The suggested use of *CVaR* is for optimization of a given system's performance in terms of waiting time and is relevant in cases where the frequency of long waiting times is the primary concern.

In this article, a discrete event simulation model of an orthopedic surgical unit in Copenhagen, Denmark is presented as the case-study. The long term goal for the simulation study is to minimize the total waiting time, with special focus on long delays. In the case-study analysis of the uncertainties and behaviour of different performance measures including *CVaR* under various resource and simulation settings are presented. Moreover, *CVaR* is compared to other measures using this model as illustration. The article is structured in the following way: Section 2 describes the case-study. *CVaR* is defined in section 3 followed by section 4 where the performance measure is evaluated by considering the simulation model under different resource and simulation setups. Finally the key findings are summarized in section 5.

2. Simulation Model

In this section, we present our case-study for evaluating the performance of the *CVaR* waiting time criterion in the simulation of an orthopaedic surgery unit. The level of detail of the model is intentionally kept low, since our main objective is to use it as an illustration of the *CVaR* measure.

2.1. The Surgical Unit

As in much of the rest of the world, over the past decade the Danish public health care system has been subject to increasing demands for efficiency [14]. The system is now under considerable pressure for higher throughput in order to reduce waiting lists. Avoiding or reducing delays in the system is certainly one of the many options to reach this goal. Furthermore, fewer and/or shorter delays may also increase patient satisfaction, an issue that is central to today's quality and productivity improvement strategies in general.

The case-study is a surgical unit, which is part of an orthopedic department at a university hospital in Copenhagen, Denmark. The unit undertakes both acute and elective surgery and performs more than 4,600 operative procedures a year. While the patients come from various wards throughout the hospital, the main sources of incoming patients are the four stationary orthopaedic wards or the emergency care unit. The outpatients treated in

outpatient clinics are not considered in this model but the resources shared between outpatients clinics and the surgical unit are included. Also day-case surgery patients with short recovery times are included in the model.

2.2. Model Description

The conceptual model is outlined in Figure 1. It consists of three main modules: (1) the incoming module with arrival and wards, (2) the surgical unit with preparation and operating rooms and (3) the recovery. Module 3 is linked back to module 1, since the patients return to the wards for final recovery and discharge.

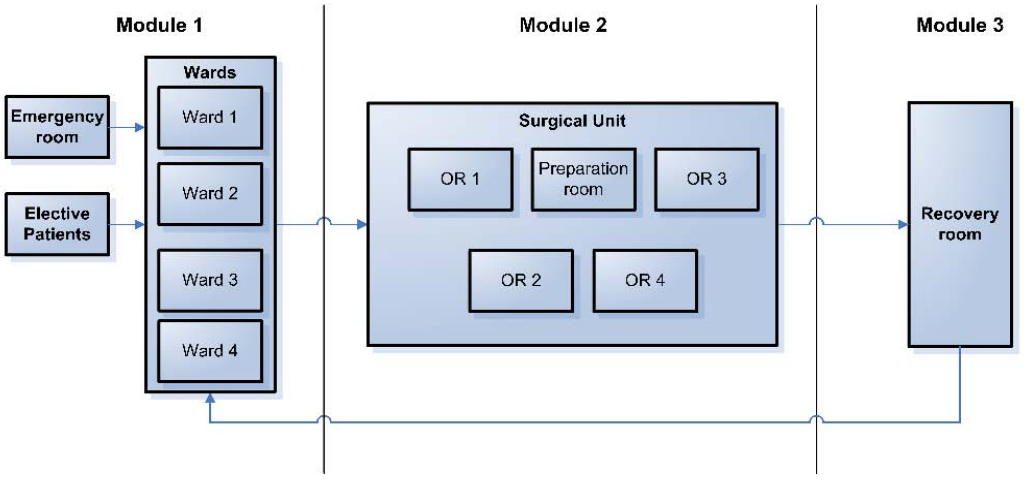

Figure 1. Conceptual model for an orthopedic surgery unit. The 3 modules are separated by vertical lines and the arrows indicate the patient flow.

The simulation model is implemented in Extend^{TM} version 6 [17] and controlled from a Microsoft Excel spreadsheet with a Visual Basic for application script. The patient flow is outlined in Figure 2. All patients are either acute or elective and are admitted to one of the four stationary wards from where the patients are collected when an operating theater is ready. Patients are then either sedated, sent to a preparation room and brought to the operating room or brought directly to the operating room for sedation and preparation. The patients are operated and hereafter attended to by an anesthesiologist before being moved to the recovery room. As the patients are moved out of the operating room, cleaning and preparation of the rooms for the next patients are started.

The resource constraints in the system are process related: available surgeons for the operation, a free recovery bed and an available porter for moving the patient to the recovery room, *etc*. These resources are controlled by a central mechanism controlled by different schedules, *e.g.*, more resources during regular hours. Sharing between different specialties is handled with the resource pools. In our model the resources include staff and physical facilities such as operating rooms and recovery beds. It should be noted that some resources such as surgeons, anesthesiologists, porters and recovery beds are shared with other departments or procedures not directly related to the surgical unit.

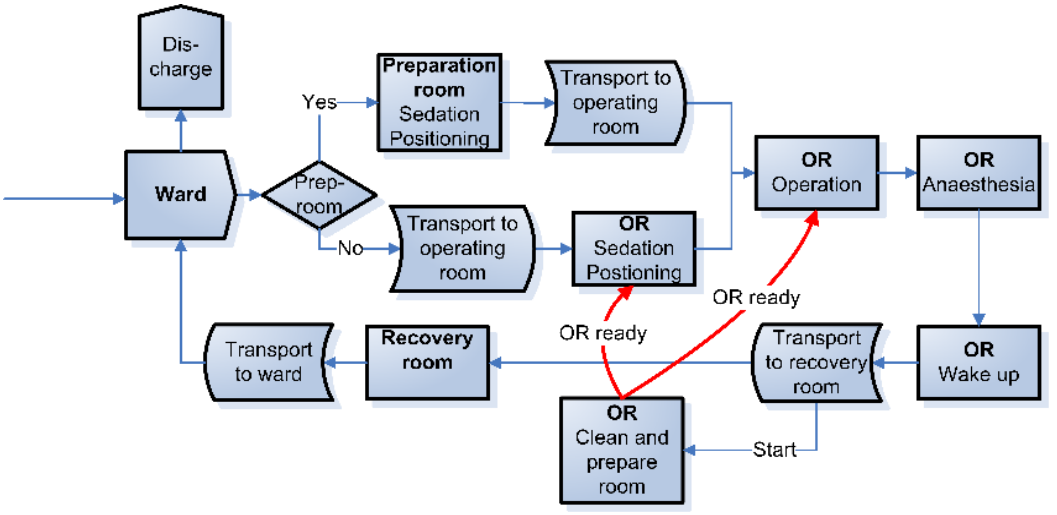

Figure 2. Process diagram for the patient flow through the system from ward to discharge.

2.3. Empirical Data

Prior to the simulation study, a simple registration of the time from patients' arrival atthe surgical unit until their departure to the recovery room was done by the staff for a period of 3 months. The initial data set held no information on subprocesses, which implied that a more elaborate registration system was needed. In the new registration system, the nurses at the surgical unit recorded the patient flow through the unit from the ward to the recovery room, *i.e.*, each subprocess was recorded over a period of 1 month.

The new data was validated on the data collected routinely by the staff prior to the simulation study by comparing the total time spent at the surgical unit recorded in the two data sets with a Kolmogorov-Smirnoff (K-S) goodness of fit test [6], which indicated no significant difference. Furthermore, tests for correlation [12, 2] between processes in the new data set indicated that the subprocess durations were statistically uncorrelated indicating that subprocesses could be modeled individually.

2.4. Validation and Verification

The model was inspected graphically by the management of the department to verify the patient routing and the procedures. Animation was included in the model to assist and simplify verification during the presentation of the model.

Model validation corresponding to patient volume and waiting time was carried out by comparing the simulation output with the observed data. All validation was carried out using graphical methods (QQ-plots, density plots and histograms) and formal statistical tests (K-S and Wilcoxon rank-sum tests [13]) with a significance level of 5%. A more elaborate validation was also carried out corresponding to the scheme outlined by Sargent [22] and although this concluded that the model was adequate, it is not presented in this article.

The model parameters were calibrated on the individual processes and queuing times, and finally validated on the total duration defined as the time from the patient leaving the ward to the time the patient is moved to the recovery room. Figure 3 shows that the model tends to mimic the system's overall behavior, which was confirmed with K-S and Wilcoxon

tests indicating no statistical difference. The throughput, mixture of patients and distribution of patients per day were validated as a part of the tuning and calibration process.

The incoming rate of elective patients per day was shown to fit a discretized triangular distribution, which was also validated by a K-S test. The acute patients were assumed to have exponentially distributed inter-arrival times. K-S tests indicated that the distribution of acute patients per day and the ratio of elective to acute patients were modeled adequately. The acute incoming rate was much more volatile compared to the one for elective patients. The coefficient of variation (CV), which is defined as the standard deviation divided by the mean, was 2.5 times higher for the acute patients compared to the elective patients. In both cases the variation in the observed data set was large with CV greater than 90 %.

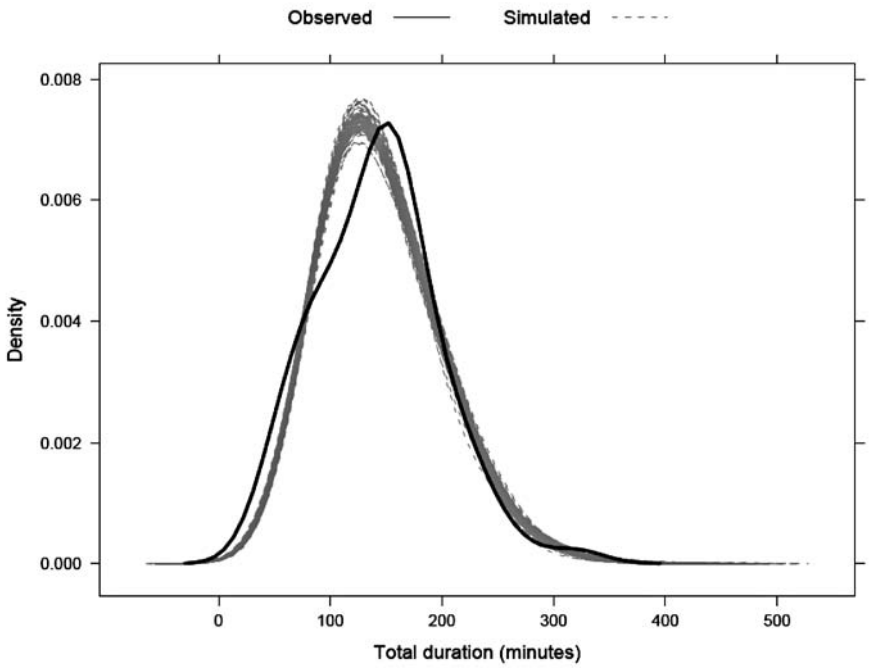

Figure 3. Estimated density functions for observed (black solid line) and 100 simulation runs (gray dotted lines) for total time at surgical unit.

3. Performance Measures

One of the most essential issues in any simulation study is to define sound and reliable performance measures [19]. Each simulation run is summarized in a set of measures, which characterizes the overall performance of the system. Often more than one measure is investigated in order to quantify the objectives of the study, *e.g.*, avoiding long waiting times while keeping a certain level of patient throughput. In this paper Conditional Value at Risk is introduced as a waiting time measure targeting the longest waiting times and compared to other existing measures.

3.1. Conditional Value at Risk

Conditional Value at Risk is a concept originating from finance as a measure of risk [21, 15, 16]. For a distribution of waiting times, T, $CVaR$ is defined as the expected value

of the $(1-\alpha)$-tail of T, *i.e.*, given as

$$CVaR_\alpha(T) = E[T \mid T > q_\alpha], \tag{1}$$

where q_α is the α-quantile, where $P(T \le q_\alpha) = \alpha$. For a sample of simulated waiting times, $T_x = \{t_{x1}, \ldots, t_{xN}\}$ (obtained from the x^{th} run), the $CVaR_\alpha(T_x)$ is estimated by

$$CVaR_\alpha(T_x) = \frac{1}{1-\alpha}\left[\left(\frac{i_\alpha}{N} - \alpha\right) t_{xi_\alpha} + \sum_{i=i_\alpha+1}^{N} \frac{t_{xi}}{N}\right] \tag{2}$$

with $t_{x1} \le t_{x2} \le \ldots \le t_{xN}$, i_α is the index satisfying $i_a / N \ge \alpha > (i_\alpha - 1)/N$, t_{xi_α} is the α-quantile and in economics denoted as the Value at Risk (*VaR*). *VaR* is seen to be indifferent to the shape of the $(1-\alpha)$-tail, *i.e.*, a given *VaR* value covers situations from short $(1-\alpha)$-tails to long $(1-\alpha)$-tails. In most applications of *CVaR* the estimate is based on the $(1-\alpha)100\% = 5\%$ longest waiting times and in the following *CVaR* is therefore estimated by equation (2) with $\alpha = 0.95$.

For waiting times the *VaR* waiting time is the value of the α-quantile of the total waiting times, *e.g.*, for $\alpha = 0.95$, 95% of the patients have a total waiting time less than or equal to *VaR*. *CVaR* is the average of the 5% longest waiting times, *i.e.*, a measure about the tail of the waiting time distribution. It is seen that *CVaR* is at least as large as *VaR* and that the difference indicates the length of the tail, hence the two measures are correlated. *CVaR* is seen to be more sensitive to samples with very long waiting times compared to *VaR*. However, Webby *et al.* [27] noted that *CVaR*, as opposed to *VaR*, is more stable with changes in the α-value. This can be explained by the fact that *CVaR* is an average of the tail, whereas *VaR* is the quantile defining the tail. The quantile is likely to jump with a small sample, whereas the average will shrink this effect.

The rationale for introducing *CVaR* waiting time measure is that it is a well known measure of risk in finance. It fits well in an optimization framework with the objective of minimizing the overall waiting time while controlling the risk of experiencing very long waiting times. The tail of the waiting time distribution in these studies is quite important since as shown by Bielen and Demoulin [3], in terms of patient satisfaction, waiting time influences satisfaction negatively. That is, longer waiting times decrease patient satisfaction significantly. Using the average waiting time inherently imply that the distribution of the waiting times is unimportant as long as the overall waiting time is low. This is, however, not in accordance with patient satisfaction and quality perception. On the other hand the maximum waiting time may be a too risk averse measure and could potentially confound good settings with bad settings since it is based on only the most extreme observation.

The benefits of using *CVaR* as performance measure are that it is easy to compute, easy to interpret and targets the long waiting times. As mentioned above, if the mean waiting time (denoted risk neutral) is used, an increase in the longest waiting times can be overlooked since a shift in the tail may be averaged out by the rest of the distribution. On the other hand, using the maximum waiting time (risk averse) may corrupt the results, since a single long waiting time may be an outlier in an otherwise well performing setup. *CVaR* can be seen as a compromise between the average waiting time $(\alpha = 0)$ and the maximum waiting time $(1 - 1/N < \alpha < 1)$, with $(1-\alpha)$ reflecting the risk of long waiting times. Hence a low α corresponds to a high risk of overlooking long waiting times since the importance of these is low.

3.2. Other Measures

Other measures have been suggested in the health care literature, which are discussed in the following. Tang *et al.* [26] presented mean residual life, *i.e.*, the expected residual life time given that a unit has lived a certain amount of time. In terms of waiting time this is equivalent to the expected residual waiting time having waited a certain amount of time. Length of additional stay (LAS) is another metric for measuring waiting times, Silber *et al.* [24] defined it as the remaining length of stay (LOS) after the transition point at which the stay becomes prolonged. A stay may become prolonged at the first time point, x, where the probability of a total length of stay of $x+y$ is greater than the probability for a LOS of y from the beginning. The test for the prolonging point is done with the Hollander-Proschan test [11]. LAS is seen to be the mean residual life at the point where the stay becomes prolonged. The rationale behind LAS is that if a stay is prolonged it is more likely to be associated with a complicated case [24].

Both LAS (the MRL at the prolongation point) and MRL are similar to the $CVaR$ measure. However, $CVaR$ is the expected waiting time of the $(1-\alpha)100\%$ longest waiting times, whereas mean residual life at the α-quantile is the expected remaining waiting time after having waited t_{xi_α} minutes. Silber *et al.* [24] suggested using the point at which a stay becomes prolonged as the choice for α. This imply that for different setups the corresponding $LAS's$ (or $MRL's$) are the average residual waiting times for the prolonged stays, *i.e.*, for different α-values. Furthermore, the scale is different depending on the setting: in one case it may the residual waiting time after having waited 30 minutes while in another it may be the residual waiting time after having waited 60 minutes. For LAS and MRL in general unlike for $CVaR$ the interpretation is seen to be dependent on the distribution. This implies that the scale and interpretation are maintained for different settings, which makes it suited for use in optimization. Moreover, the distribution of waiting times may be on time, i.e. no prolongation point is present, which implies that the LAS concept breaks down.

From a quality point of view the waiting time may be more interesting than the residual waiting time, since the patient's perception of the quality of the treatment is related to his/her total waiting time and not the residual waiting time after having already waited for x minutes. In terms of waiting times the length of additional stay may not be as important as for the length of a hospital stay, since the waiting time indicates something about the system's performance and not of the severity of the operation or complications for the individual patient. Moreover, the waiting time is the time between activities and hence complicated cases have longer activity times and more difficult recovery, which do not influence the waiting time. Silber *et al.* [24] used the LAS as an indicator of health care outcomes and the measure is hence not targeted at evaluating a system's performance. The LAS framework does not seem to be well suited for evaluating waiting times, whereas it is highly relevant for seeking complicated hospital stays.

4. Case Study

This section presents the performance measures by applying them to output from the simulation model presented in section 2. The measures are initially examined under the existing setup in terms of the variation from run to run and the sensitivity to length and number of runs. They are then considered under different resource settings. The proposed measure, *CVaR*, is analyzed and compared to other well known measures presented in section 4.3.

4.1. Simulation Setup

The simulation model is run for at least 300,000 minutes (see section 4.4). This corresponds to 30 weeks with a warm-up period of 10,080 minutes (1 week) for each run. In each run different performance measures are obtained as described in section 4.3. These measures are summarized by their minimum, maximum, average and coefficient of variation (sample standard deviation in % of the average) across runs.

4.2. Analysis Methods

The results from the simulation model are analyzed using statistical test methods. Wilcoxon two-sample tests [13] are used to compare two samples in terms of their location. The test is a non-parametric test. Comparing two samples in terms of their distributions is done with Kolmogorov-Smirnoff two-sample test [6], which is also a non-parametric method. Here we compare the empirical distributions and test whether they can be assumed to be identical. Significance of correlation coefficients is tested based on Spearman's rho [12, 2], a non-parametric approach based on ranks. The main rationale for using non-parametric tests is that they do not rely on specific distribution assumptions and are robust against outliers. All data analysis was done in R version 2.7.1 [20].

Densities functions are estimated with the density procedure from the stats-package and plotted with the densityplot function from the lattice-package in R [20, 23] using the default values. The defaults are a Gaussian kernel with a bandwidth, $h = 0.9n^{-1/5}$ $\min[\hat{\sigma}_x, IQR_x / 1.34]$, where x is the sample, which has sample standard deviation $\hat{\sigma}_x$, inter-quartile range IQR_x and sample size n (Silverman's rule-of-thumb) [25].

4.3. Performance Measures

The main focus of the simulation study is on the waiting times defined as the time wasted between processes and is measured in minutes. For each patient a number of waiting times are identified: waiting time before the surgeon talks to the patient before sedation, waiting time for the anesthesiologist, waiting time before operating room is ready and waiting time for a porter and a free recovery bed, etc. The total waiting time for the j^{th} patient in the i^{th} simulation run, t_{ij}, is estimated as the sum of K sub waiting times, t_{ijk}. The waiting time measures considered in this article are

- Average waiting time, $\overline{WT}$
- Maximum waiting time, MWT
- Conditional Value at Risk, $CVaR$, waiting time, $CVaR(WT)$
- Value at Risk, VaR, waiting time $VaR(WT)$

Additionally total throughput (total number of patients treated, TT) and percentage of elective patients treated outside regular hours, $EOUT$, are considered. These measures are included in the simulation study to ensure that the throughput remains the same and the elective patients are treated outside regular hours, hence without creating additional costs due to overtime. The average and maximum waiting times are frequently used measures to quantify the waiting time [5]. VaR is included to highlight the additional information contained in our main measure, $CVaR$, and to illustrate its close relationship to $CVaR$.

4.4. Run Length and Sample Size Analysis

The first example consists of simulations on the system at its current configuration. Here, the main objective is to examine the performance measures under different run lengths

and numbers of repetitions (runs). Table 1 shows the summary for three types of simulation runs for the system as it is: (1) 30-weeks simulation repeated over 100 runs, (2) 30-weeks simulation repeated over 200 runs and (3) 60-weeks simulation repeated over 60 runs.

From the first block in Table 1 it is seen that the total waiting times are highly skewed with an average $\overline{WT}$ of around 31 minutes, a $95\,\%$ quantile of around 61 and a maximum of 111 minutes. It is seen from the CV column in the first block in Table 1 that the most varying measure is the MWT ($CV = 11.3\%$) followed by $EOUT$ ($CV = 9.2\,\%$). The remaining four measures are comparable in terms of coefficient of variation ($1.5\% \leq CV \leq 2.5\%$).

Table 1. Summary for performance measures over runs, *e.g.*, the minimum, maximum, average and CV of total throughput for three types of simulation setups. The *Min*-entry for the first row *e.g.*, summarizes the minimum $\overline{WT}$ of the 100 runs, Max the maximum, $Mean$ the average and CV the standard deviation in percent of the mean. The unit for the waiting time statistics are minutes, the unit for $EOUT$ is percent and TT is measured in number of patients.

	Min	Max	Mean	CV(%)
	30 weeks, 100 runs, 3 porters			
$\overline{WT}$	30.03	32.21	30.97	1.52
MWT	89.00	157.88	111.25	11.34
TT	1635	1797	1711	2.02
$EOUT$	8.25	12.69	10.15	9.22
$CVaR$	67.98	77.47	71.17	2.26
VaR	58.05	64.01	60.95	1.92
	30 weeks, 200 runs, 3 porters			
$\overline{WT}$	29.69	32.29	30.98	1.49
MWT	89.00	163.36	111.92	11.48
TT	1615	1827	1715	2.15
$EOUT$	8.25	12.97	10.36	9.15
$CVaR$	67.58	78.09	71.36	2.30
VaR	58.05	64.40	60.94	1.99
	60 weeks, 60 runs, 3 porters			
$\overline{WT}$	30.21	21.52	30.91	0.94
MWT	94.30	153.97	118.57	10.27
TT	3347	3599	3468	1.82
$EOUT$	8.95	11.73	10.51	5.91
$CVaR$	67.90	73.35	71.17	1.43
VaR	58.96	62.16	60.69	1.25

Figure 4 illustrates the evolution of the CV's as the number of runs is increased. It can be seen that all CV's are stabilized after 70 runs, however subdivided into the two groups as described previously. It can also be seen that the two upper curves take more runs to settle in compared to the bottom four. Clearly the maximum waiting time is a measure

highly dependent on the simulation run, since it is the most extreme observation in each run. The average waiting time is as expected the least varying measure, whereas the *CVaR* and *VaR* are seen to vary almost equally much. Figure 4 indicates that the four best performing measures have stabilized after 30-40 repetitions.

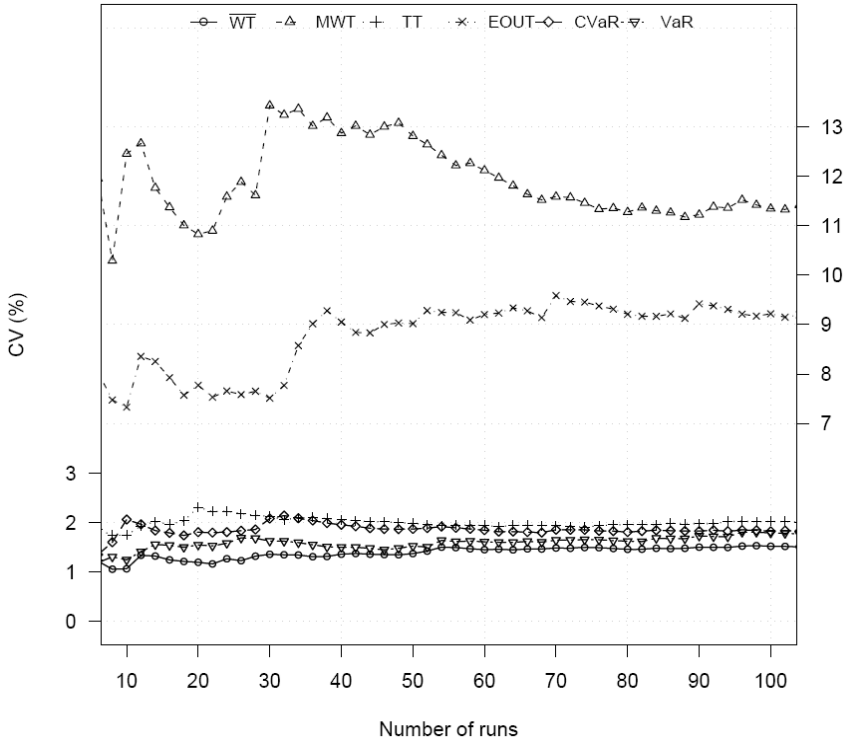

Figure 4. Coefficient of variation as function of included runs for 6 performance measures.

Figure 5 shows that a run-length of 300,000 minutes (30 weeks) seems to be adequate for obtaining a low *CV* for 5 out of 6 measures (no significant improvements hereafter). *EOUT* is seen to be improving by more than 2%-points from 300,000 minutes to 600,000. Simulating 30 weeks repeated 60 times is a good trade-off between simulation time and precision for *MWT*, which leads to an approximate half width of a 95% confidence interval for the average of *MWT* corresponding to 2.7% of its estimated value. For $\overline{WT}$, *TT*, *VaR* and *CVaR* considerably fewer repetitions are needed. In fact Figure 4 suggests that fewer than 20 repetitions will be sufficient.

In the 100 run simulation of 30 weeks each *CVaR* is significantly correlated with *VaR* (as expected), *MWT* and $\overline{WT}$. Moreover, *VaR* is significantly correlated with $\overline{WT}$, whereas *TT* is correlated both with *EOUT* and $\overline{WT}$. The correlations are all positive, which implies that higher throughput is associated with longer waiting times. The *VaR* is seen to be uncorrelated with the *MWT*, whereas *CVaR* is. This in fact fits well with the definition of *CVaR* and *VaR*. The connection between *CVaR* and $\overline{WT}$ and *MWT* was shown in section 3.1.

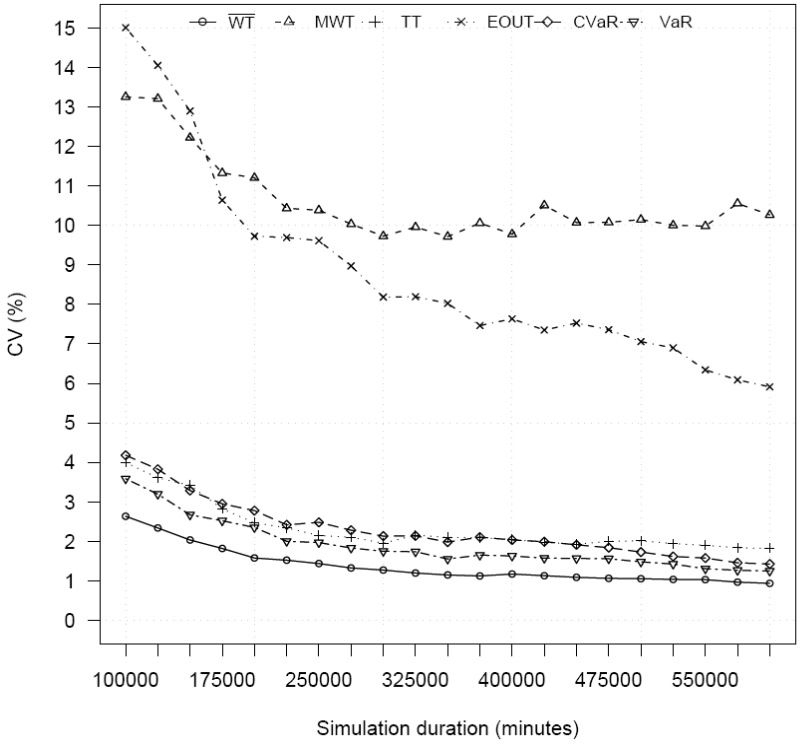

Figure 5. Coefficient of variation for 60 runs with varying run lengths for the 6 performance measures.

4.5. Sensitivity to Changes in Resource Allocation

The sensitivities of the measures to changes in resource allocation are analyzed by changing the number of porters at the surgical unit in regular hours. Three porters are available in regular hours in the current system described in section 4.4. This number is set to 1, 2 and 4 in the following analysis. The porters are a relatively less costly resource to adjust than the number of surgeons, nurses and operation rooms. The expectations are that lowering the number of porters will increase the waiting times and decrease the throughput or increase the percentage of patients being treated outside regular hours. Hence increasing the number of porters may enable an increase in the incoming flow of patients without increasing the waiting times if the remaining resources are underutilized in the current setup.

Table 2 summarizes the results from 60 runs of 30 weeks for three different settings of porters. It can be seen that having 2 or 4 porters are equivalent with the results for 3 porters in Table 1, whereas having 1 porter clearly increases the waiting times in terms of the average, *CVaR* and *VaR* waiting time. In the top part of Figure 6 the associated estimated density functions indicate that 2-4 porters lead to equivalent waiting time distributions, whereas the 1 porter distribution seems to differ.

With 1 porter it is observed that all measures besides the total throughput are changed significantly (Wilcoxon two-sample test [13]: $p < 0.001$) compared to having 3 porters. The patients wait longer on average (8.56% increase on average), have higher maximum waiting times (8.41 % increase on average), more patients are treated outside regular hours (19.41

% increase on average) and *CVaR* and *VaR* are increased significantly (7.53% and 6.97%, respectively). Figure 6 shows that the primary change from 2-4 porters to 1 porter is a heavier tail. This is reflected in the *CVaR* in Table 1 and 2, which show that the increase is around 2 times the increase in the average waiting time. The top part of Figure 6 shows that the estimated density function with 1 porter is flatter around the peak and has a thicker tail, which increase the *CVaR* more than $\overline{WT}$. The increase by 5 minutes in *CVaR* from 3 to 1 porter corresponds to an increase in waiting time for the approximately 85 patients with the 5% longest waiting times of 7 hours. In our simulation study the difference in *CVaR* is statistical significant, but the practical importance of the increase may be limited.

Table 2. Summary for performance measures over runs for three different configurations as in Table 1. The unit for the waiting time statistics are minutes, the unit for *EOUT* is percent and *TT* is measured in number of patients.

	Min	Max	Mean	*CV*(%)
	30 weeks, 60 runs, 4 porters			
$\overline{WT}$	29.85	31.93	30.89	1.56
MWT	92.58	161.64	113.01	12.17
TT	1609	1812	1710	2.45
EOUT	6.87	12.69	10.48	9.99
CVaR	66.97	74.24	71.17	2.44
VaR	58.09	63.35	60.67	1.97
	30 weeks, 60 runs, 2 porters			
$\overline{WT}$	30.08	32.34	31.16	1.42
MWT	87.70	139.94	110.49	10.27
TT	1629	1815	1718	2.42
EOUT	8.38	13.05	10.88	8.11
CVaR	67.79	75.33	71.13	2.54
VaR	57.99	63.14	60.92	1.89
	30 weeks, 60 runs, 1 porters			
$\overline{WT}$	32.70	34.42	33.62	1.16
MWT	97.88	151.27	120.01	10.86
TT	1625	1815	1715	2.41
EOUT	10.36	14.08	12.12	6.51
CVaR	71.78	80.61	76.53	2.37
VaR	62.40	67.65	65.20	1.92

Adding an extra porter does not shorten the waiting times (top block in Table 2), the situation is comparable with the original 3 porter setting. The performance measures were not significantly different. The lowest p-value is obtained for *VaR* with a p-value of 0.18. Figure 7 furthermore shows that increasing the number of elective patients leads to a significantly worse performance compared to both the 3 and 4 porter situation (for all measures other than *MWT*). The bottom part of Figure 6 indicates that the patients are waiting longer on average as the incoming rate is increased and that the tail of the waiting time distribution has the same length (*MWT* the same) but is heavier (*VaR* and *CVaR* increased).

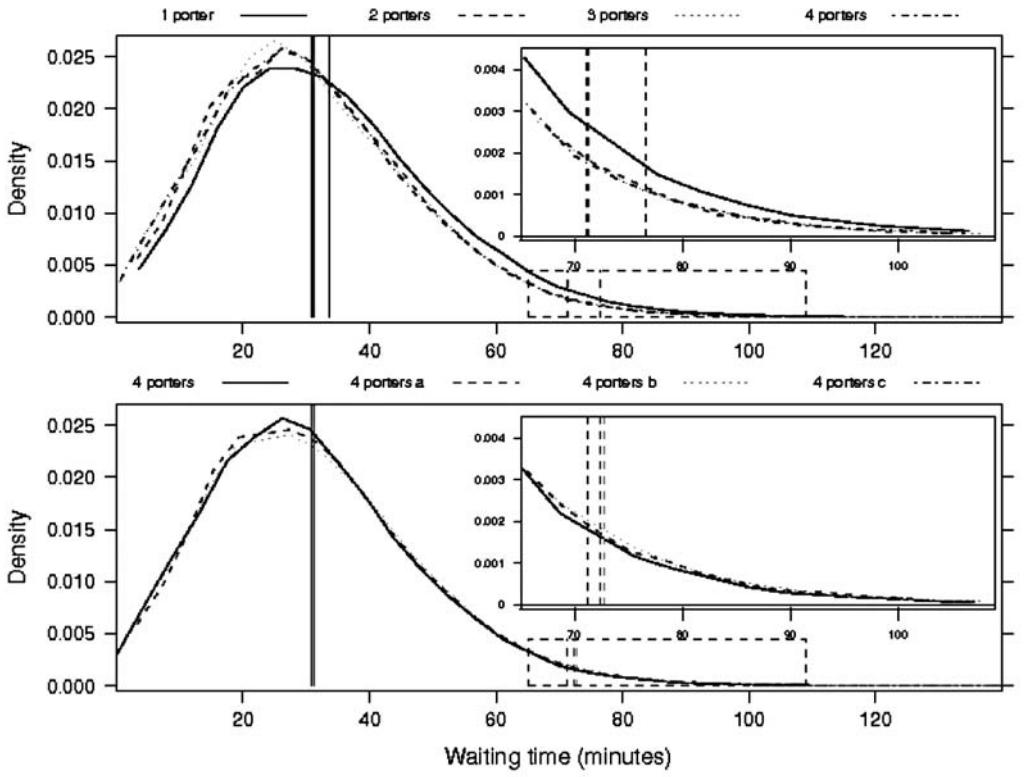

Figure 6. Estimated densities for 1, 2, 3 and 4 poters (top) and 4 poters with different patient load (bottom). Vertical lines correspond to mean waiting time (solid lines) and $CVaR$ value (dashed lines). Poters 4a, 4b and 4c correspond to 4 poters with 7%, 14% and 29% more elective patients, respectively. The dashed area in the lower right of each panel is highlighted in the upper right.

It can be seen that $CVaR$ has a higher absolute increase compared to $\overline{WT}$ for the 3 vs. 1 porter comparison, showing that the 5 % longest waiting times are increased the most. For increased patient input MWT does not increase, whereas $CVaR$ and VaR do. This shows that using the MWT as criterion for judging the waiting time performance is a poor choice as it may not pick up differences in the waiting time distribution due the large uncertainty on this measure of the extreme. Moreover, the MWT does not consider the shape of the waiting time distributions, which may differ in the thickness of the tails but have the same MWT. It is seen that $CVaR$ picks up the change in the distribution of waiting times by using information from the whole tail rather than relying on the most extreme observation in each run.

5. Conclusions

The analysis of simulation studies needs reliable performance measures to answer the relevant research questions. In this article $CVaR$ is suggested as a measure of the tail distribution of waiting times for a surgical unit with the objective of avoiding long waiting times. Our analysis shows that $CVaR$ is a reliable measure that is specific to the tail. Moreover, $CVaR$ can be seen as a compromise between the risk neutral average waiting time

and the risk averse maximum waiting time. The results presented in this article show that using the maximum waiting time is a poor choice since it is highly variable and ignores changes in the shape of the waiting time distribution.

The average waiting time is not always representative for the waiting times, since such distributions often are skewed and long waiting times may potentially be more problematic from the points of view of patients and management. The *VaR* criterion is a measure of a quantile in the distribution but is indifferent to the tail distribution and does not quantify the tail distribution. In terms of quality management with patient satisfaction as outcome *CVaR* is highly relevant since it quantifies the problematic long waiting times. Moreover, the *CVaR* criteria is more stable compared to *VaR* with respect to the chosen α-level since it is a sample average. It has nice properties as it is easy to compute and interpret and it is robust. *CVaR* of the waiting times may therefore be a relevant outcome in many quality improvement studies within health care with the objective of reducing the risk of long waiting times.

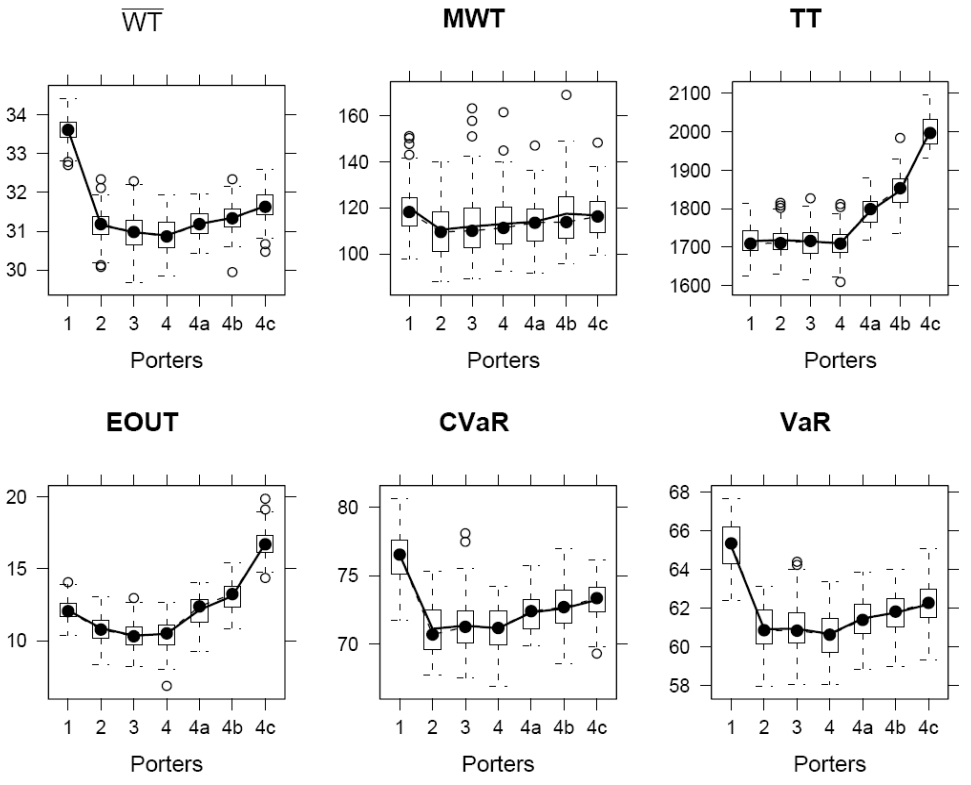

Figure 7. Box plots for comparing performance criteria for different resource settings. 4a, 4b and 4c correspond to 4 porters with 7%, 14% and 29% more elective patients, respectively.

References

1. Alexander, S., Coleman, T. and Li, Y. (2006). Minimizing VAR and CVAR for a portfolio of derivatives. *Journal of Banking and Finance*, 30(2), 583-605.

2. Best, D. and Roberts, D. (1975). Algorithm as 89: The upper tail probabilities of spearman's rho. *Applied Statistics*, 24, 377-79.

3. Bielen, F. and Demoulin, N. (2007). Waiting time influence on the satisfaction-loyalty relationship in services. *Managing Service Quality*, 17(2), 174-193.

4. Brailsford, S. C. (2007). Tutorial: Advances and challenges in healthcare simulation modelling. *Proceedings of the 2007 Winter Simulation Conference*, 1436-1448.

5. Cayirli, T. and Veral, E. (2004). Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4), 519-49.

6. Conover, W. J. (1971). *Practical Nonparametric Statistics*. New York: John Wiley & Sons, 295-301 (one-sample Kolmogorov test), 309-314 (two-sample Smirnov test).

7. Denton, B. T., Rahman, A. S., Nelson, H. and Bailey, A. C. (2006). Simulation of a multiple operationg room surgical suite. *Proceedings of the 2006 Winter Simulation Conference*, 414-424.

8. Ferrin, D. M. and McBroom, D. L. (2007). Maximizing hospital financial impact and emergence department throughput with simulation. *Proceedings of the 2007 Winter Simulation Conference*, 1566-1573.

9. Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Roberts, K., Coyle, E., Bevan, G. and Palmer, S. (2003). Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health Medicine*, 25(4), 325-35.

10. Garca-Gonzlez, J., Parrilla, E. and Mateo, A. (2007). Risk-averse profit-based optimal scheduling of a hydro-chain in the day-ahead electricity market. *European Journal of Operational Research*, 181(3), 1354-1369.

11. Hollander, M. and Proschan, F. (1972). Testing whether new is better than used. *The Annals of Mathematical Statistics*, 78(4), 1136-1146.

12. Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. New York: John Wiley & Sons, 185-94.

13. Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. New York: John Wiley & Sons, 27-33 (one-sample), 68-75 (two-sample).

14. Jun, J., Jacobson, S. and Swisher, J. (1999). Application of discrete-event simulation in health care clinics: a survey. *Journal of the Operational Research Society*, 50(2), 109-123.

15. Kibzun, A. and Kuznetsov, E. (2003). Comparison of VAR and CVAR criteria. *Automation and Remote Control*, 64(7), 153-164.

16. Kibzun, A. I. and Kuznetsov, E. A. (2006). Analysis of criteria VAR and CVAR. *Journal of Banking & Finance*, 30(2), 779-796.

17. Krahl, D. (2002). The extend simulation environment. *Proceedings of the 2002 Winter Simulation Conference*, 205-213.

18. Miller, M., Ferrin, D., Ashby, M., Flynn, T. and Shahi, N. (2007). Merging six emergency departments into one: A simulation approach. *Proceedings of the 2007 Winter Simulation Conference*, 1574-1578.

19. Nakayama, M. K. (2006). Output analysis for simulations. *Proceedings of the 2006 Winter Simulation Conference*, 36-46.

20. R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

21. Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26, 1443-1471.

22. Sargent, R. G. (1998). Verification and validation of simulation models. *Proceedings of the 1998 Winter Simulation Conference*, 121-130.

23. Sarkar, D. (2009). *Lattice: Lattice Graphics*. R Package Version 0, 17-22.

24. Silber, J. H., Rosenbaum, P. R., Koziol, L. F., Sutaria, N., Marsh, R. R. and Even-Shoshan, O. (1999). Quality and outcomes of care - conditional length of stay. *Health Services Research*, 34(12), 349-363.

25. Silverman, B. W. (1986). *Density Estimation*. Chapman and Hall, 48.

26. Tang, L., Lu, Y. and Chew, E. (1999). Mean residual life of lifetime distributions. *IEEE Transactions on Reliability*, 48(1), 73-78.

27. Webby, R., Adamson, P., Boland, J., Howlett, P., Metcalfe, A. and Piantadosi, J. (2007). The mekong-applications of value at risk (VAR) and conditional value at risk (CVAR) simulation to the benefits, costs and consequences of water resources development in a large river basin. *Ecological Modelling*, 201(1), 89-96.

Authors' Biographies:

Christian Dehlendorff is a Ph.D. student in Informatics and Mathematical Modeling at the Technical University of Denmark. He has a M.Sc. in Engineering within data analysis and statistics. His research interests are within design of experiments and computer experiments.

Murat Kulahci is an Associate Professor in Informatics and Mathematical Modeling at the Technical University of Denmark. His research interests include design of experiments, statistical process control, and financial engineering. He is a member of the American Statistical Association, European Network of Business and Industrial Statistics (ENBIS), and the Institute of Operations Research and the Management Sciences.

Søren Merser is a surgeon (MD) at Clinic of Orthopedic Surgery at Frederiksberg Hospital, Denmark. He is a member of Danish Orthopedic Society and his primary research interest is on-line quality control in hospital units.

Klaus K. Andersen is an Associate Professor in Informatics and Mathematical Modeling at the Technical University of Denmark. He has a Ph.D. in time series analysis and his research interests are within design of experiments and statistical consulting.