UvA Econometrics

Discussion Paper: 2010/06

Can statisticians beat surgeons at the planning of operations?

Paul Joustra, Reinier Meester and Hans van Ophem

www.feb.uva.nl/ke/UvA-Econometrics

Amsterdam School of Economics
Department of Quantitative Economics
Roetersstraat 11
1018 WB AMSTERDAM
The Netherlands

UvA Universiteit van Amsterdam

Can statisticians beat surgeons at the planning of operations?[1]

Paul Joustra
Academic Medical Center and Faculty of Economics and Business
University of Amsterdam

Reinier Meester
Faculty of Economics and Business
University of Amsterdam

and

Hans van Ophem[2]
Faculty of Economics and Business
University of Amsterdam

July 2010

Keywords: efficient planning of operations, duration models, cost reduction
JEL-codes: I10, I12

Abstract

The planning of operations in the Academic Medical Center is primarily based on the assessments of the length of the operation by the surgeons. We investigate whether duration models employing the information available at the moment the planning is made, offer a better alternative. We conclude that substantial cost reductions can be achieved by employing statistical techniques. This does not imply that the surgeons' predictions do not contain valuable information. This information is a key explanatory variable in our statistical models. What our conclusion does entail is that a correction of the predictions of surgeons is possible because they are often underestimating the actual length of operations.

[1] All ML-routines used in this paper are either performed by using standard routines from Stata or are carried out using R (free software, for information see http://www.r-project.org/).

[2] Corresponding author. Full address: Department of Quantitative Economics, Faculty of Economics and Econometrics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. Email: j.c.m.vanophem@uva.nl. Phone: +31 20 5254222. Fax: +31 20 5254349.

1. Introduction

Health care expenditures in western economies appear to be ever rising and are becoming a growing concern for both governments and residents. The burden to cover the costs invokes all the inventiveness of policy makers to come up with new ideas intended to decrease the rate of growth of, or even better, reduce these expenditures. The reason for the growing consumption of health care is threefold (cf. Chiappori et al (1998), Okunade and Murthy (2002), and Bago d'Uva and Jones (2009) for more thorough discussions): (1) the demographic shifts towards the geriatric age groups, (2) the ongoing development in medical care technology, and (3) the existence of large-scale health insurance schemes. Governments have trouble to influence the first two causes simply because they are largely out of their control or not popular due to elective reasons. In countries with publicly provided or financed health care system or insurance, governments have some direct control and many attempts have been undertaken to influence the growth rate of health care expenditures. Bago d'Uva and Jones (2009) give an extensive overview of the different methods European governments have used to regulate the demand for health care in order to slow down or even reduce health costs. Influencing the costs through the supply side usually takes the form of increasing the efficiency, cf. van Houdenhoven et al (2007) and Wullink et al (2007).

In this paper we will investigate whether it is possible to improve the efficiency of the planning of surgical operations at the Academic Medical Center (AMC) in Amsterdam, The Netherlands. In the present situation and in most hospitals, surgeons determine this planning to a large extent, cf. Dexter et al (2007) and Eijkemans et al (2010). They estimate the expected duration of an operation and based on this information the planning of the operating room (OR) is made. A question that should be raised is whether surgeons make their estimates in the interest of the hospital or on the basis of their own interests.

At the AMC, a large academic hospital in the Netherlands with 1200 beds and a budget of €728 million (2007), over 55.000 surgical operations where carried out in 2007 (Annual Accounts, 2007). The costs involved with operations are high. For example, according to a study by Macario et al (1995), OR costs make up for around 33 percent of the Stanford University Medical Center budget. Improvements in the planning of operations might therefore have a substantial impact in the reduction of the costs.

The difficulty of OR planning is balancing between schedules that are too wide and schedules that are too tight, while the duration of individual procedures listed in a schedule is often highly volatile and uncertain. If the planning is too wide there is a risk of empty OR time in between operations or at the end of the day. On the other hand, if the planning is too tight, OR cases will often cause overtime of OR personnel or even cancellations. Cancellations have to be avoided as much as possible in order to maintain a good level of patient satisfaction. On the other hand, the option to let the OR run overtime instead of canceling cases is costly and unpopular with OR personnel. Currently, the amount of overtime and cancellation of operations at the end of the day are a large problem at the AMC. Approximately 36% of programs ran late and average overtime resulting was around 50 minutes (Benchmarking OR, 2008). Only 4% of programs finished on time. It is for these reasons that OR management at the AMC seeks to improve the accuracy of daily OR planning and there appears to be plenty of scope.

More accurate prediction of individual OR case durations is one of the ways to reduce the current size of the problem of overtime and cancellation of operations. Here an OR case is defined as all that happens between entrance and exit of the OR by a patient. Generally, it consists of a pre-incision period for anesthesia induction and surgical preparations, the surgical procedure (possibly multiple) itself and the postsurgical period for anesthesia ‘deduction’. At most departments of the AMC, surgeons currently predict the duration of an operation at the intake of a patient based on their experience and preferences. The first element as such is no problem but the second might be driven by self-interest.

Unfortunately the surgeon's estimates of the case duration are not very accurate. For example, 18% of the ophthalmologic cases carried out in the AMC between 2003 and 2008 finished more than 15 minutes early and 34% finished more than 15 minutes later than planned. For other clinical specialties with longer procedures, these numbers are even larger. Since 2008, pilots have been running at the Neurosurgery and Gynecology departments to use also the historical averages per procedure per surgeon instead of personal predictions of surgeons alone. Previous investigation by Dexter et al (2007) indicates however that the historical average is unlikely to predict the variation in duration better than current predictions.

In our investigation we will predict the duration of operations on the basis of a number of different hazard models and we will compare the results with the predictions provided by surgeons. The predictions will be made on the basis of the ex ante information available, including the estimate of the duration by the surgeon. As such, using more complex statistical techniques is not a new idea, but thus far only the lognormal regression model appears to have been employed (Strum et al (2000a), Strum et al (2000b) and Eijkemans et al (2010)). Here we will use the exponential model, the Weibull model, the loglogistic model, the Burr or

Weibull-Gamma mixture model, the generalized Gamma model and the piecewise-constant hazard model as well.

We have data available of all ophthalmologic, neurosurgeric and gynecologic operations performed in the last twenty years in the AMC. Because the registration of case characteristics became more complete in 2003 only data from 2003 onwards are used. The remaining period is divided into a 'historical' or 'estimation' period (2003 – 2007), which is used for the estimation of econometric model, and a 'prediction' period (January – November 2008). The performance of the different prediction methods is compared within this out-of-sample prediction period. Not only do we consider the prediction on the individual case level, we will also investigate the performance of the different prediction techniques in terms of overtime, undertime and the number of cancellations for all the operations in the prediction period.[3]

In the next section the general problem of efficient OR planning and the relation with prediction of OR case duration is explained in more detail. Also, some relevant literature on prediction of individual case duration is reviewed. In section 3, we briefly discuss the statistical estimation methods and we will also discuss how the performance of the different methods will be evaluated. Section 4 contains a description of the available data and section 5 presents the empirical results. The conclusions are listed in section 6.

2. The planning of operations

A daily OR program consists of elective cases and ambulatory cases. In this paper we define elective cases as all those cases that can be planned up to 10.30 am the day before, when the final planning has to be ready for the next day. Ambulatory cases are all cases coming through after that time. For some specialties of the hospital like general surgery there are separate emergency rooms for ambulatory cases and these cases do not disturb regular planning. For other specialties however, like Ophthalmology, where cases are usually less urgent, there is no separate emergency room. For the last category of specialties, planning of elective cases is likely to be disturbed and delayed by the ambulatory cases coming through. Usually planners account for the possibility of ambulatory cases by leaving some spare time at the end of a daily program (see figure 1). For this reason we will ignore ambulatory cases. On top of that, for ambulatory cases no expected duration of the operation is recorded. Even though we do not consider ambulatory cases, a completely accurate planning of the OR capacity is

[3] We use the term undertime as the counterpart of overtime. Undertime, resulting in an underemployment of an operating room, will be considered to be a negative attribute by hospital managers as well as overtime and cancellations.

impossible due to randomness or unpredictable variability in case duration. For example unforeseen complications can occur during the surgical procedure. Moreover, the unpredictability of case durations is worse than average for the AMC, due to the academic nature of the hospital which attracts relatively many of the more rare or complex cases.

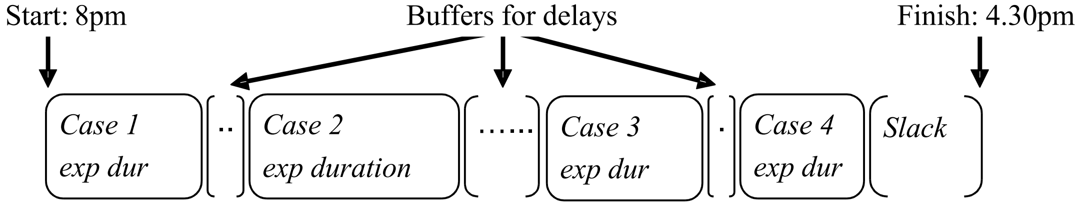
Because of the impossibility of completely accurate planning, optimal planning of OR capacity is a matter of balancing between several interrelated interests for the AMC. On the one hand, the hospital is reluctant to plan too tight or 'offensive', with the consequence that programs are likely to delay. As mentioned in the introduction this means that either cases have to be canceled[4] at the end of the program or that the OR runs overtime. The first result conflicts with the wish of the hospital to satisfy patients and the second result is not only costly but also unpopular among personnel. These problems can be avoided by leaving enough empty space at the end of the program, called 'slack', or by wide or 'defensive' planning (see figure l), but it is not hard to imagine that planning too defensive is not efficient either. If a case finishes earlier than planned, the next patient has to be prepared in advance in order to continue operating. Assuming that a patient is waiting in the preoperative waiting room no more than half an hour before he or she is scheduled to be operated, it is likely that no patient is ready to be operated after several cases have finished earlier than planned. In this case precious OR time is wasted while personnel waits for the next patient. More important even, if the entire program finishes earlier than planned, then there is almost certainly no patient at hand to fill the space remaining at the end of the day. So on the other side of the coin is the risk to plan too defensive and not fully exploit the OR capacity in between operations or al the end of the day.

Most specialties within the AMC currently tend to plan offensively. This explains the numbers presented in the introduction: 36% of programs ran late and the average overtime resulting was around 50 minutes (Benchmarking OR, 2008).

There are several ways to improve OR efficiency. A first way aims at reducing OR case duration by planning ‘straights’ of the same procedures. The idea is that surgeons or their assistants gain skillfulness during the straight resulting in reduced duration per case. This solution would have the positive effect that more procedures can be carried out on daily basis, but it does not directly address the problem of unpredictable variability in program duration (van Houdenhoven et al (2007)).

[4] In the AMC delays lead to cancellation of operations if the last operation(s) planned can not be started before 4 pm, the deadline to initiate a non-ambulatory case.

Figure 1. Graphical representation of daily planning.



Opposite to the solution of series of identical cases, is the solution of efficient portfolio selection. It is based on the idea that diversification in cases could reduce variability (risk) in the duration of an entire program. The theory originates from Nobel laureate Harry Markowitz, who intended it for asset portfolio construction and asset pricing in finance. In the hospital it could be applied by planning cases of similar variability next to each other. In theory the *idiosyncratic* risk of individual cases would then be partially offset, resulting in reduced variability in the duration of the entire program. Better diversification would yield better results (van Houdenhoven et al, 2007) as long as individual case durations are uncorrelated.

A third method to increase OR efficiency is to allow operating schedules to be more flexible. In the AMC the available OR time of a specific department is subdivided to individual surgeons at the beginning of the year and this subdivision is more or less fixed. For example, a surgeon always operates on Monday and Wednesday morning. More flexible schedules could improve daily and weekly planning because planners would be less constrained in finding the optimal daily portfolio of procedures.

Finally there is the solution of more accurate prediction of individual case duration, which is the central issue of this paper. This solution would first of all reduce the risk of individual cases finishing earlier or later than planned. Additionally, it is likely to reduce the risk or variability in an entire daily program as well however. This second effect would mean that less final slack is required in daily programs and therefore, that the OR can be used more efficiently without an increased risk of overtime and cancellations.

Currently there are two different methods to predict OR case durations at the AMC: prediction by surgeons and prediction using historical averages. The first method was used by Ophthalmology, Gynecology and Neurosurgery, and based solely on the experience and preferences of surgeons. For Ophthalmology, the surgeon writes an estimate of the duration of surgery at the intake form of a patient, accompanying a code for the most important surgical procedure. This estimate is supplemented by the planners of the department with a fixed amount of time for local or total anesthesia to determine the planned duration of an entire case. In 2008, the ophthalmologic surgeons underpredicted the case duration with less than 3 percent on average. The Ophthalmology department has neither an explicitly defensive nor offensive planning strategy. The 'imprecision' of planning measured in average absolute difference between planning and actual duration was nearly 29 percent however. Over all departments, most surgeons seem to underpredict case duration to avoid idle OR time resulting in offensive planning. Apart from an average tendency of underprediction of 17 percent AMC wide, predictions are generally imprecise with an average absolute difference between planned and actual duration of 36 percent (Benchmarking OR, 2008).

In 2008, the Gynecology and Neurosurgery departments started to plan OR cases using the historical average of the last ten 'similar' cases conducted by the same first surgeon as well. Here an historical case is regarded as similar if the main procedure that characterizes the newly accepted case was *at least* performed *within* the historical case. Whether additional procedures are carried out (or other specialties operated simultaneously) does not matter for regarding the case as similar. Since multiple procedures within a case occur quite frequently, approximately 25 percent of neurosurgery cases for example, it is evident that this method of estimation is often quite inaccurate. However, the historical average is only meant as a guiding figure. Ultimately surgeons and planners still decide on the actual time to be reserved for a case. Both Gynecology and Neurosurgery seem to have benefited from the new planning method because the inaccuracy of planning was approximately 16% lower in 2008 than in the five years prior to 2008.

The inaccuracy of prediction of OR case duration on the basis of the experience of surgeons or anesthesiologists or historical averages is discussed in Dexter et al (2007). They show that although using historical averages probably reduces underestimation of OR case duration, the larger problem of imprecision remains. In literature a number of alternative (statistical) methods have been suggested to predict OR case duration more accurately. The statistical distribution of the duration of surgery was investigated as early as 1963, when Rossiter and Reynolds (1963) noted that the distribution of the duration of surgery appears to fit a lognormal distribution well. An improvement of this method can be achieved by subdividing the data into more homogeneous subgroups (Dexter and Zhou (1998)). In Strum et al (2000a) the emphasis is on the appropriateness of the lognormal model (compared to the normal model) to describe case duration. It is considered category wise for categories with respect to Current Procedural Terminology (CPT) code and anesthesia type (general, local, monitored or total). They use a Friedman test to compare goodness-of-fit of the normal and the lognormal model and find that the lognormal model is preferable in 93 percent of cases. According to the authors, rejection of the lognormal model occurs if the subsample size is large, short procedure times are rounded or in case of outliers. The lesson of Strum et al (2000a), is not however, that the lognormal model is the most appropriate model overall to describe the distribution of case duration. In fact this topic has received little attention in literature at all and is therefore the most important topic of this paper.

In Strum et al (2003) earlier findings were supplemented by comparing the normal and the lognormal model for cases consisting of exactly two procedures, resulting in even higher preference of the lognormal model. Like in Strum et al (2003) and Eijkemans et al (2010), discussed below, cases with multiple procedures occur in the dataset of our investigation as well.

In Eijkemans et al (2010) a comparison is made between prediction of surgical duration by surgeons on the basis of historical averages and prediction on the basis of a lognormal regression model. The authors use five basic groups of regressors: operation characteristics, e.g. type of surgical procedure, session characteristics, e.g. the number of procedures, team characteristics such as experience of the team, patient characteristics such as age and Body Mass Index (BMI) and other characteristics such as the estimate of duration by the surgeon (without knowledge of an historical average). They find all categories except patient characteristics to contribute a considerable amount to the explanatory power of the model. Adding all explanatory variables significant at 30% they find an adjusted R-squared of 0.796. More importantly, the authors report a reduction in over- and underprediction of case duration by 19% and 17% respectively. Whereas Eijkemans et al (2010) applies only a lognormal regression model, they have more information on cases and therefore potential explanatory factors. In our investigation we apply several other methods, but less information is available from the information system. Also we have fewer observations available.

In the papers of Dexter and Zhou (1998), Strum et al (2000a) and Strum et al (2000b) it was identified that procedure, surgeon and anesthesia seem to be statistically significant explanatory factors for the duration of OR cases. Strum et al (2000b) and Strum et al (2003) estimate a lognormal regression model that they call ‘aggregate’ for the entire set of cases, in addition to fitting two-parameter lognormal or ‘individual’ models to subclasses of the data. As additional explanatory variables to CPT code and anesthesia technique they have the age of the patient, a variable indicating physical status (ASA), emergency and surgical specialty category as explanatory variables. They do not identify any of the additional factors to be statistically significant determinants of variability in duration, comparing differences in duration after tabulation with respect to the variables.

In Dexter et al (2008) a summary of articles is provided on explanatory factors for case duration. In this study first of all they explain differences in components of case duration by different medical conditions, different anatomic procedures used for the same medical condition and different approaches to achieve the same anatomic result. They too find that for prediction on the basis of the scheduled procedure(s), the operating personnel and anesthetic(s) considerable inaccuracy remains. Therefore they have searched for studies that use information from outside OR information systems such as medical records of surgeons, radiology pictures and patient demographics. They find little evidence however of these alternative explanatory factors significantly contributing to increased accuracy in prediction.

3. Statistical methods

The variable of interest is the duration of an operation. The natural method of analysis of durations is hazard models. Lancaster (1990) and Cameron and Trivedi (2005) give an extensive overview of these models. Since our objective is not so much the understanding of the contributing factors to the duration of operations but to get optimal predictions of the duration and since there are no clues to which model to use, we will apply a broad range of hazard models and simply evaluate important sample statistics to see what hazard model is the optimal one and whether we can outperform the predictions of surgeons. As stated before we will estimate the model on part of the available data (about 80% of the data) and make predictions on the remaining part (about 20% of the data). We will consider the following duration models:

- the exponential hazard model
- the lognormal hazard model
- the Weibull hazard model
- the log-logistic hazard model
- the Burr or Weibull-gamma mixture hazard model
- the generalized gamma hazard model
- the piecewise constant hazard (PCH) model.

The Burr-hazard model is a '*mixture*' model and contains a number of the other models listed above. Originally the Burr stems from allowing for a gamma distributed unobserved heterogeneity in the Weibull model. The Weibull hazard belongs to the class of proportional hazard specifications and this means that the hazard function can be written as:

$$\lambda(t|x,\theta)=\lambda_0(t,\psi)\cdot\phi(x_i,\beta) \tag{1}$$

where t denotes the duration, x_i is a vector of explanatory variables and $\theta = (\psi,\beta)$ are unknown parameters. The usual choice on the specification of is $\exp(\beta' x_i)$. Allowing for unobserved heterogeneity means that an error is added to this last specification:

$$\phi(x_i,\beta)=\exp(\beta' x_i)\cdot\varepsilon_i=\phi_i\cdot\varepsilon_i \tag{2}$$

Under the assumption of a gamma-distrubuted ε_i and using the Weibull hazard, the Burr hazard model results. The cumulative distribution function of the Burr is

$$F(t|x_i,\theta)=1-(1+\sigma^2\phi_i t^\alpha)^{(-1/\sigma^2)} \tag{3}$$

where $\alpha > 0$. σ^2 reflects the variance of the unobserved heterogeneity term ε_i. This distribution function contains as special cases the Weibull distribution for $\sigma^2\rightarrow 0$ and the exponential distribution by setting also $\alpha = 1$. The log-logistic distribution is yet another special case that can be obtained by putting $\sigma^2=1$.

The lognormal hazard model is already applied by Sturm et al (2000b) and Eijkemans et al (2010). It assumes that the natural logarithm of duration is normally distributed with mean $\beta' x$ and variance σ^2. The model is most intuitively presented as a linear regression model:

$$\log(t)=\beta' x_i+u_i \tag{4}$$

where u_i is normally distributed with mean 0 and variance σ^2. This model can be estimated with OLS and this might explain the popularity of this model in the literature.

The generalized gamma family of models belongs to a different class of models than the previous models described, namely the class of Accelerated Failure Time (AFT) models. This means the model can be expressed as follows:

$$\log(t) = -\log(\lambda(\beta' x_i)) + u_i \qquad (5)$$

where in this case $u_i = w_i/\alpha$ and exp(w_i) is Gamma(m) distributed and $\lambda(\beta' x_i)$ is the hazard function (Lancaster, 1990, p.38). The *u* term is a disturbance term that allows for unobserved heterogeneity. The distribution of the disturbance term implies that the generalized gamma family of models is characterized by the following density function:

$$f(t) = \alpha \phi_i^{\alpha m} t^{\alpha m - 1} \exp - ((\phi_i t)^{\alpha m}) / \Gamma(m). \qquad (6)$$

where Γ(m) is the gamma function. α, m, and $\phi_i > 0$ are the parameters of the model. Regressors are brought in by letting $\phi(x_i, \beta) = \exp(\beta' x_i)$. The density reduces to the Weibull density if $\alpha = 1$, to the two-parameter gamma density if $m = 1$ and to the exponential density if both $\alpha = 1$ and $m = 1$.

The piecewise constant hazard model belongs to the class of proportional hazard characterized by (1). The main characteristic of the piecewise constant hazard model is that it allows the baseline hazard λ_0(t) to be a step function so that this hazard is constant in prespecified time intervals. In this sense it is a generalization of the standard exponential model for which the hazard is restricted to be constant across the entire range of t. So, in the piecewise constant hazard model we have

$$\lambda_0(t, \psi) = \exp(\alpha_j) \text{ if } c_{j-1} \leq t < c_j \text{ for j = 1,..,M} \qquad (7)$$

where $c_0 = 0$ and $c_M = \infty$ and the other thresholds are specified, but the α_j's have to be estimated. As before, regressors are brought in by letting $\phi(x_i, \beta) = \exp(\beta' x_i)$ in (1). Depending on how small the intervals are taken over which the hazard is assumed to be constant, the model can be made as flexible as needed but at the cost of introducing additional parameters that have to be estimated.

Prediction performance measures

To evaluate the predictions for the durations of operations following from the above listed models and stated by the surgeons we will consider the following performance measures:

- TOTAL: the total of the estimated operation time needed to process all operations of the prediction period according to the OR planning[5]
- MEAN: the mean of the estimated operation time
- AD: the average difference between prediction and actual duration
- AAD: the average absolute difference between prediction and actual duration
- rMSE: the root mean squared error
- UPx: the proportion underprediction by more than x = 10, 20 and 30 minutes
- OPx: the proportion overprediction by more than x = 10, 20 and 30 minutes
- LOSS: LOSS(prediction method) = abs[AD(prediction method)] + AAD(prediction method).

Performance is optimal when an unknown 'loss function' is minimized. The choice for the symmetric and rather straightforward loss function above, that is quite similar to the Mean Squared Error (MSE) for small losses, is probably not optimal for the AMC. Our loss function is defined this way because no formal research has been done into the actual losses that result when cases end for example 10, 20 or 30 minutes late or early. Only when these losses are calculated or estimated a really sensible loss function can be defined. For example one could imagine that larger penalties are given to delays than to early finishes.

An implication of our choice of loss function, i.e. that we minimize AD in addition to AAD, is that we prefer predictions centered *on average* around actual duration to predictions centered around some higher or lower percentile of duration. On average, we choose therefore neither for offensive nor for defensive planning but for a neutral planning instead. In fact we are willing to give up some of the accuracy for the sake of minimal AD in absolute terms. The optimal method from our perspective is not necessarily the most accurate in terms of AAD.

4. Data

The AMC has started registration of case duration and some characteristics as early as 1988. In this investigation we have decided however to use the data from 2003 onwards. The first reason is that so much has changed in the OR and in operation technology since 1988 that the early information is not likely to be relevant for current case duration prediction. What is

[5] We decided not to present the deviation of the total time needed and the deviation of the means for the planned and actual durations because we believe that the presented figures are of interest themselves. A comparison with the actual total time and mean can be done by using the information from Table 1.

more, many case characteristics that are available through the OR information system today, were not registered until 2003. We retrieved information on operations performed by three different specialties: Ophthalmology, Neurosurgery and Gynecology. The selection of specialties allows for the investigation of a wide variety of OR cases that is more or less representative for the AMC. Neurosurgical cases are generally very complex and demanding and accordingly have the longest average duration as well as the largest spread in duration. Many unpredictable complications can occur during a case. Ophthalmologic cases are usually shorter and less unpredictable. Gynecology combines the extremes of Ophthalmology and Neurosurgery, consisting of many very short procedures as well as relatively many of the more complicated and especially long-lasting cases. Together these specialties make up for an interesting and widespread collection of cases to investigate statistically.

Because of the unavailability of the surgeon's expected duration of the operation, we had to discard all ambulatory operations. Apart from lacking this information, other information on unexpected or emergency operations is often not available as well. As a result, the case duration of ambulatory cases will be much harder to predict. As discussed before, the AMC solves the occurrence of ambulatory operations by allowing for some slack in the daily operation schedule or by using the emergency room.

Sample statistics on the actual and planned duration of the estimation and prediction samples can be found in Table 1. For Ophthalmology the data set resulting from the selection of procedures consists of 5299 observations of which 1208 (22.8%) lie in the prediction period of approximately 11 months. The average duration is 75.6 minutes with a minimum of 6 and a maximum of 735 minutes. Around 95% of the cases last no longer than 2 hours. The average planned duration is right on the spot. The standard deviation of the planning is quite a bit lower than that of the actual duration. These figures grossly reflect the character of ophthalmologic procedures: they are of short duration and duration is relatively easy to predict. The nature of the operations of Neurosurgery is very different than those of Ophthalmology. First of all, the dataset consists of only 2286 observations in total of which 423 (18.5%) lie in the prediction period. The 95th percentile is now greater than 500 minutes, whereas average duration is 245 minutes. Especially the right tail of the distribution is spread out much more for Neurosurgery therefore than for Ophthalmology. The planned duration appears to systematically underestimate the actual duration. The difference between planned total duration and actual total duration of all operations in the estimation sample is almost 30%. The planned spread is also substantially smaller than the actual spread. The underprediction of the duration of operations appears to be systematic. Gynecology entails a combination of short procedures and very long procedures, although not as long as the longest neurosurgeric procedures. Because of this combination, the average duration of 111 minutes lies somewhere in between. The 95th percentile is near 300 minutes. The spread also lies somewhere in the middle. Also for Gynecology the planned duration differs considerably from the actual duration and again there appears to be an underprediction. The total number of observations is 4268 and 796 (18.7%) observations lie in the prediction period. Note that the sample statistics differ for the samples distinguished but the conclusions drawn before hold also for the prediction sample.

Table 1: Sample statistics on the actual and planned duration of operations.

	Ophthalmology		Neurosurgery		Gynecology	
	Estimation sample	Prediction sample	Estimation sample	Prediction sample	Estimation sample	Prediction sample
Nr of obs	4092	1208	1863	423	3472	796
Actual duration						
Total	309355	86976	456435	91960	383656	87321
Mean	75.6	72.0	245.0	217.4	110.5	109.7
Stand. dev.	41.3	37.0	178.2	183.2	97.2	93.9
Minimum	6	11	20	26	10	7
Maximum	735	397	1544	1115	863	775
Planned duration by the surgeon						
Total	308128	87097	351921	78128	326021	82068
Mean	75.3	72.1	188.9	184.7	93.9	103.1
Stand. dev.	30.5	25.8	108.7	148.9	83.0	83.8
Minimum	10	15	15	30	5	15
Maximum	330	300	660	784	507	426

Unit of measurement of all sample statistics: minutes.

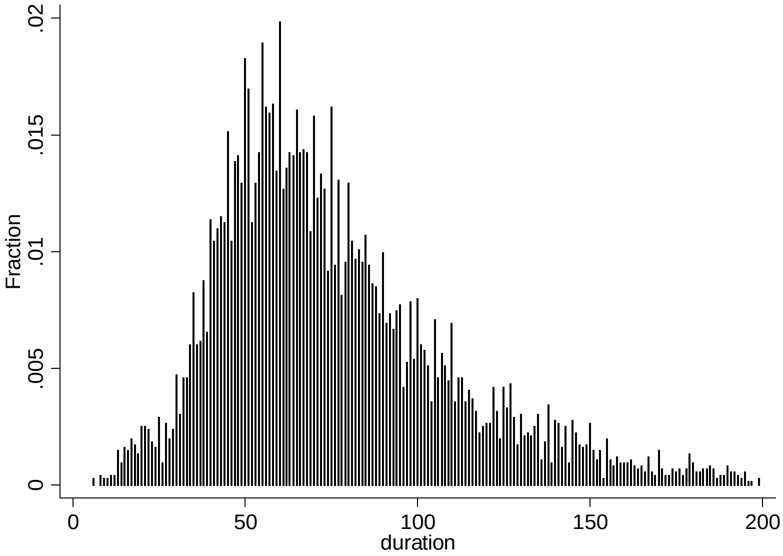
Apart from the distributional assumptions underlying any econometric regression model**,** the dependent variables of the model are the most important factors to explain (or describe) the differences in case duration. Since our efforts are aimed at predicting operation durations as good as possible we will include all information available to us, but only if this information was available before the operation was scheduled. A complete list of the variables used can be found in the appendix. The explanatory variables can be divided into a number of categories.

Following Eijkemans et al (2010), the explanatory variables are distinguished in five categories: operation characteristics (e.g. type of surgical procedure), session characteristics (e.g. the number of surgical procedures), team characteristics (experience of the team), patient characteristics (health condition indicators) and other case characteristics (the predicted duration of the operation by the surgeon). In the first instance, the predicted duration of the operation by the surgeons appears to be a peculiar explanatory variable to use since it seems to be at odds with the objective of this investigation. However, what we are interested in is to predict the duration of operations as good as we can with the use to statistical techniques and on top of that evaluate whether the use of such methods has the potential to improve the predictions as given by surgeons. As such these expectations are likely to contain very valuable information for the prediction of case duration, although, since the surgeons have their own incentives, these expectations might be biased. The figures in Table 1 illustrate that the surgeons tend to underestimate especially in the neurological and gynecological specialties. Note, however, that in all three cases the performance of the surgeons is better in the prediction than in the estimation period. Whether this is a coincidence or not is not clear, although we know that the AMC has put more emphasis on the importance of good estimation of operation duration in latter years and our prediction sample consists of the more recent observations.[6] The question is to what extent are the surgeon's estimations of the length of operations biased and whether other information has some explanatory power such that we can improve on the predictions. Note that the information we have additional to the expectation of the surgeon, is available to the surgeon as well.

Unfortunately, we experience a significant amount of missing values. To solve this problem we replaced the missing values by the average of the variable (in case that an average has a meaning) or by zero values (in the case of e.g. dummies). In each of these cases a separate binary variable is generated that is equal to 1 for the missing information. Especially the group of patient characteristics is registered very irregularly and the discrete variables indicating health are nearly constant at zero (no complications). As a result, these particular variables are expected to have limited explanatory power.

[6] As we have stated before, from the beginning of 2008 the departments of Neurosurgery and Gynecology also use information on the historical average duration per surgeon in the planning of operations.

Figure 2: Spike plot of ophthalmologic operation duration.



A complication in the data available is the prevalence of measurement errors both in the dependent variable is in at least one important explanatory variable. The measurement error in case duration is caused by the fact that operating personnel tends to round off operating room durations to a five minute precision level. For example quite distinguished peaks are seen in the spike plot of Ophthalmology every five minutes compared to relative lows in between (Figure 2), especially around an hour. Another indication can be found in Table 1. The minimum and maximum planned durations are all factors of five minutes. The rounding errors might have an effect on the performance of the continuous prediction methods in this paper. We have experimented with rounding off predictions to a five minute precision level and we concluded that the rounding off does not appear to have a systematic effect.

The second variable that is known to be subject to measurement error is the first surgeon. The first surgeon reported a priori is not always the one who is actually performing the surgery. Although the first surgeon is the one responsible for the operation, the second surgeon or an assistant surgeon may be taking all or part o the action. If this is the case it is no longer possible to determine the correct effect of a surgeon on duration. Moreover, other parameter estimates might be biased as well. Unfortunately there is little that can be done about this flaw. Evidently, our predictions as well as current AMC predictions could have benefited to some extent from correct information concerning the surgeon.

Another complication is the fact that part of the cases consists of multiple procedures. For a rough sketch, approximately 29% of ophthalmologic cases, 27% percent of gynecologic cases and 25% of neurosurgeric cases between 2003 and 2008 consisted of 2 to maximally 8 procedures. To make the final insight into the applicability of statistical methods as complete as possible, we deliberately consider these cases as well. For the multiple-procedure cases we have chosen to use only the main procedure and the total number of procedures within the case as explanatory variables, instead of using all information and adding each performed procedure. The latter approach is not expected to deliver better results because the additional time required for extra procedures is usually less than the time required for the procedure if it stands by itself. The most important explanation for this difference is the fact that multiple procedures usually overlap in time. The second approach would introduce a measurement difficulty that would not be solved easily. At least many more explanatory variables would be required. The former approach, also taken by Houdenhoven (2007), is preferred mainly because the corresponding parsimony is expected to weigh more heavily on prediction performance than the loss of information attached to it.

5 Empirical results

We estimate the duration of an operation for the three specialties Ophthalmology, Neurosurgery and Gynecology separately with several hazard specifications and with the use of all information available at the moment operations are scheduled. We do not strive to get a model that is capable of explaining the duration but we are interested in the best prediction possible. As a result we decided to plug in all information available to us. To investigate the quality of a duration model we split up our three samples into two parts: (1) an estimation subsample, on which the model is estimated, containing about 80% of the complete sample and (2) a prediction subsample, on which we predict durations, containing about 20%.[7] The following hazard models are estimated: the exponential (Exp) hazard, the lognormal (Lnorm) hazard, the Weibull (Weibull) hazard, the loglogistic (Loglog) hazard, the Burr (Burr) hazard, the generalized gamma (GenΓ) hazard, the piecewise constant (PCH5: with five minutes intervals. PCH10: with ten minutes intervals) hazard.

The estimation results will not be discussed in detail. We will only present some common features across the three specialties. The estimated prediction of the length of the operation tends to be underestimated by the surgeons. This result is stronger within the neurosurgical and gynecological specialties. In all estimations the surgeon's expectation contributes significantly to the model. Other strongly significant variables are the number of

[7] The subsample sizes are approximate because the actual division of the sample was based on a date.

surgical procedures performed during the operation, characteristics of the first surgeon and the type of operation. Patient characteristics do not appear to have a strong impact.

Table 2: Prediction measures Ophthalmology (1208 operations)

	Surg	Exp	Lnorm	Weibull	Loglog	Burr	GenΓ	PCH10	PCH5
TOTAL	87097	86952	86674	88136	86432	86529	86734	87761	87773
MEAN	72.13	71.98	71.75	72.96	71.55	71.63	71.80	72.65	72.66
AD	0.13	-0.02	-0.25	0.96	-0.35	-0.38	-0.20	0.65	0.66
AAD	18.62	15.51	15.47	16.25	15.34	15.35	15.46	16.15	16.09
rMSE	25.81	23.05	23.05	23.68	22.99	23.00	23.04	23.85	23.76
UP10	0.26	0.23	0.23	0.22	0.24	0.23	0.23	0.23	0.23
UP20	0.17	0.14	0.13	0.13	0.14	0.14	0.14	0.14	0.14
UP30	0.10	0.07	0.07	0.07	0.08	0.08	0.07	0.08	0.08
OP10	0.38	0.31	0.30	0.33	0.29	0.29	0.30	0.33	0.33
OP20	0.17	0.13	0.12	0.14	0.12	0.12	0.13	0.14	0.13
OP30	0.06	0.05	0.05	0.06	0.04	0.04	0.05	0.06	0.06
LOSS	18.75	15.53	15.72	17.21	15.69	15.73	15.66	16.80	16.75

Shaded entries represent the best result across the row. The predicted duration and actual duration are measured in minutes.

Table 2 presents the prediction measures for Ophthalmology. The definition of the measures is discussed at the end of section 3. In the second column information is listed on the prediction of the surgeons (surg). The other columns present prediction measures with respect to the indicated hazard specifications. The results show that, whatever prediction method used, the total number of minutes necessary to do the 1208 operations is closely approximating the actual total duration. With respect to the other prediction measures we can conclude that the statistical techniques do in most cases better than the surgeons. The differences are not very large but apart for the average deviation between the predicted and actual duration of the operations in the prediction sample (AD) and the percentage of predictions with a difference of more than +30 minutes between the predicted and actual duration of the operation[8], the statistical techniques always do better. Note that maximizing a likelihood function does not imply that the best predictions will be found. The results with respect to the Burr hazard are in some instances worse than those of nested models like the Weibull, loglogistic and exponential hazard. Especially the loglogistic model appears to

[8] In this case the actual duration is smaller than the predicted duration.

perform well. Futhermore, the often used lognormal specification is certainly not one of the best statistical methods to use. For none of the measures it performs best. Finally, note that based on the average deviation between prediction and reality, all methods are very accurate. The average fault is in all cases less than a minute. In absolute deviations the error is much larger ranging from about 15.3 minutes (loglogistic hazard) to 18.6 minutes (surgeons).

Table 3: Prediction measures Neurosurgery (423 operations)

	Surg	Exp	Lnorm	Weibull	Loglog	Burr	GenΓ	PCH5	PCH10
TOTAL	78128	100348	98669	105839	97700	97853	-	-	122382
MEAN	184.67	237.23	233.26	250.21	230.97	231.33	-	-	289.32
AD	-32.70	19.86	15.88	32.83	13.60	13.96	-	-	71.94
AAD	68.29	60.77	58.46	70.53	56.15	56.23	-	-	105.90
MSE	103.14	110.21	104.94	135.81	99.19	99.14	-	-	454.30
UP10	0.51	0.34	0.34	0.30	0.34	0.34	-	-	0.30
UP20	0.44	0.25	0.26	0.22	0.25	0.25	-	-	0.21
UP30	0.40	0.18	0.20	0.16	0.21	0.21	-	-	0.16
OP10	0.33	0.49	0.46	0.51	0.47	0.47	-	-	0.52
OP20	0.25	0.38	0.35	0.43	0.35	0.35	-	-	0.42
OP30	0.19	0.29	0.28	0.35	0.27	0.27	-	-	0.36
LOSS	141.40	80.63	74.34	103.36	69.75	70.19	-	-	177.84

Shaded entries represent the best result across the row. No convergence was achieved for the "-" entries. The predicted duration and actual duration are measured in minutes.

Table 3 presents the same prediction measures for Neurosurgery. The conclusions are more or less in line with Ophthalmology, although neurosurgeons underpredict the duration of their operations seriously. A striking result is that the statistical methods appear to overpredict the duration, although in a much less serious manner. Part of the explanation might be the large deviations between the mean duration of operation in the estimation and prediction sample for Neurosurgery. On top of that, the standard deviations shows a reversed pattern (cf. Table 1). Surgeons underestimate the duration of neurological by more than half an hour on average. Statistical methods overestimate durations. The best result is found for the loglogistic-hazard with an overprediction of 14 minutes. The Weibull and piecewise-constant hazard perform even worse than the surgeons. The absolute average deviations are closer but still most statistical methods outperform the surgeons. We can also add that in this case the prediction measures are far worse than in the case of Ophthalmology. This conclusion is not surprising.

As we have noted before neurosurgical operations tend to be much longer than ophthalmological operations. The surgeons appear to score good at the overprediction percentages, but, of course, this is a result of the strong tendency to underpredict of surgeons. The Weibull and the loglogistic models seem to obtain the best scores and the scores are quite similar, where we prefer the loglogistic model. It scores much better on the AD, AAD, MSE and LOSS measures than the Weibull does. The Weibull scores best on the measures reflecting underprediction by more than 10, 20 and 30 minutes, but does a bad job in the prediction of actual durations of operations. Again, the lognormal hazard does not distinguish itself as a particularly good method to use.

Table 4: Prediction measures Gynecology (796 operations)

	Surg	Exp	Lnorm	Weibull	Loglog	Burr	GenΓ	PCH5	PCH10
TOTAL	82068	111615	84495	75270	84161	85260	-	78876	78605
MEAN	103.06	140.22	106.15	94.56	105.73	107.11	-	99.09	98.75
AD	-6.62	25.47	-3.54	-15.13	-3.95	-2.58	-	-10.60	-10.93
AAD	26.02	23.11	22.63	29.05	22.50	22.55	-	29.62	29.48
MSE	45.78	43.01	42.47	48.29	42.40	42.33	-	60.23	59.19
UP10	0.38	0.32	0.29	0.44	0.29	0.28	-	0.39	0.39
UP20	0.25	0.21	0.20	0.32	0.20	0.19	-	0.28	0.28
UP30	0.17	0.15	0.14	0.24	0.14	0.13	-	0.21	0.21
OP10	0.25	0.28	0.31	0.23	0.31	0.32	-	0.23	0.25
OP20	0.14	0.13	0.14	0.12	0.12	0.14	-	0.11	0.14
OP30	0.08	0.07	0.07	0.06	0.07	0.07	-	0.08	0.08
LOSS	32.64	48.58	26.17	44.18	26.45	25.13	-	40.22	40.41

Shaded entries represent the best result across the row. No convergence was achieved for the "-" entries. The predicted duration and actual duration are measured in minutes.

Table 4 present the results for Gynecology. As we argued before, the durations of operations in this specialty are somewhere in between the previous specialties considered. In this case again the surgeons are clearly outperformed by the statistical methods. Whatever performance measure considered, we can always find at least three statistical methods with a better score. The best predictions are found for the Burr hazard. Note the relative good performance of the semiparametric hazard with the five minutes time interval (PCH5). It scores best for two measures (OP10 and OP20). The loglogistic hazard performs almost as good as the Burr.

The planning of operations.

Looking at individual operations, as we do in Tables 2, 3 and, 4, does give information on the quality of the prediction methods but does not show the full and most interesting picture. In most cases more than one operation is scheduled every day and it might be that mispredictions of the duration of individual operations lead to less misprediction or even stronger mis-prediction of the entire day. In order to investigate this, it would be optimal to employ the actual planning algorithm of the AMC. Unfortunately, this is far to complex to be employed in our cases. For example, in the actual planning degree of urgency of operations is taken into account and this information is not entered in the information system and therefore, not available to us. Many other elements of the necessary information to make this planning are not available to us as well. To get an idea about the quality of the prediction methods we decided to adopt a very simple planning method. We use the prediction samples with the operations arranged according to the actual operation date and time, and simply plan the operations according to the predicted duration of the operation. After having created a fictitious operation schedule in the way, we confronted the schedule with the actual durations of the operations and calculated some performance measures. As far as we can see this is a straightforward and fair way of evaluating the different planning methods. If it favors any of the methods it will be the one based on the surgeon's evaluations since the order of the operations is determined on the basis of these expectations.

We adapt four simple planning strategies. For all strategies we impose that at least one operation is scheduled every day. In this way we allow for operations with an expected duration beyond the operation time available per day. In the first strategy, panel A in Tables 5 up to 10, we plan up to eight hours per day and overtime is never allowed, except for the first operation that day. In the second strategy (panel B) we allow for some slack at the end of the day by only planning for six operating hours. Overtime is not allowed. In the third and the fourth strategy (panels C and D) we do allow for overtime, but only to a limited degree, either in a relative or absolute manner. Overtime is allowed if it suffices the following condition: the expected duration of the marginal operation minus time left that day, relative to the time left that day is smaller than 1. This means e.g. that an operation expected to last 60 minutes will not be scheduled if less than 30 minutes operating time is left for that day. In panel D overtime is only allowed if the expected duration of the marginal operation minus time left that day is smaller than 60 minutes.[9]

[9] We investigated other strategies as well. We changed the number of available operating hours, the overtime criteria, allowed for slack between operations etc. No substantial deviations were found. In all cases the our basic findings were the same.

The performance methods we use are the number of days necessary to perform all operations according to the prediction method used (denoted by 'Days'), the number of minutes of and days with idle time of the operation room (denoted by 'Undertime (min)' and 'Undertime (days)') , the number of minutes of and days with overplanning of the operation room (denoted by 'Overtime (min)' and 'Overtime (days)') and the number of times an operation had to be canceled (denoted by 'Cancellations (days)') or the cancellations measured in minutes ('Cancellations (min)'). Operations are canceled if the expected duration of the last scheduled operation minus the time left until the end of the day exceeds 60 minutes and if the expected duration of the last scheduled operation minus time left that day, relative to the time left that day is smaller than 0.5.[10]

We only report the results for the predicted duration of operations as made by the surgeons, the predicted duration on the basis of the lognormal hazard (since this is the most commonly used hazard function in the literature) and the most promising (according to Tables 2, 3 and 4) statistical methods (i.e. the Weibull, the loglogistic and the Burr hazard).

Tables 5, 6 and 7 present some characteristics of the complete planning of the operations in the prediction period for the Ophthalmology, Neurosurgery and Gynecology specialties. As we discussed before, we present results for four different planning strategies and these are given in panels A to D. An important indicator of the quality of the planning is the number of days necessary to program all operations. For Ophthalmology the performance in this respect of the planning based on the surgeon's indication of the duration of the operations or the one based on the statistical methods is very comparable. For the strategies not allowing for limited overtime the surgeons appear to do a little better, whereas if overtime is allowed the statistical methods have the advantage. Things are markedly different for Neurosurgery and Gynecology. The surgeons appear to do much better than the statistical methods. In the case of an eight hour operation day and no overtime the 423 neurosurgical operations can be planned in only 205 whereas the best statistical method needs 29 days more. However account has to be taken of the fact that neurosurgeons severely underestimate the duration of the operations. Only 78128 minutes are planned in by them, whereas the best statistical methods plans in 97853 minutes (Table 3). Afterwards the actual duration of the 423 operations turned out to be 91960 minutes. To make a fair comparison we should therefore adjust the number of days needed by the neurosurgeons with 17.7% (=(91960 -

[10] Changing the cancellation policy by putting the relative factor to 1, something that appears to be more in line with the way we allow for limited overtime, although in the cancellation policy both conditions need to hold, does not have a consequential impact on the conclusions. Slightly more operations will be canceled and that is true whatever prediction method used and more or less with the same factor of proportionality.

78128)/ 78128) and this increases the number of days necessary to perform all 423 operations to 242 days, a result that is much more in line with the statistical methods. If we make the similar adjustments in the case of the other planning strategies the same conclusion prevails. The correction factor for Gynecology is 6.4%. So the surgeons' number of planned days have to be increased to 221, 303, 181 and 191 days in order to get a fair comparison. The numbers do not compare favorably with the statistical methods, although we also should make a similar, but smaller, correction for the statistical methods because these under- or overestimate the actual duration of the gynecological operations as well. Note that the Weibull hazard predicts the least number of days necessary. Table 3 reveals that this is the only statistical method that severely underestimates the actual duration of the operations.

Another consequence of the underestimation of the duration of operations is that the score on vacant operation rooms, measured by undertime, is relatively good whereas the score on overtime and cancellations is relatively bad. A brief glance on tables 6 and 7 indeed reveals that the surgeons score usually better on the undertime indicator, but worse on the overtime and cancellation indicators. A surprise is that the same conclusion holds for Ophthalmology, even though, the predictions on the duration of the operations are right on the spot. Note that the undertime results across the three different specialties are quite different. This is not explained by a substantial difference in total operations times, but is explained by the nature of the operations. In Ophthalmology the average duration of an operation is much shorter than in Neurosurgery, whereas Gynecology is somewhere in between. The same explanation applies to the relative differences in the overtime and cancellation indicators. The general conclusion has to be that there exists some trade off between the quality indicators undertime, overtime and cancellations. As such, this is not a surprise but it can be clearly found in Tables 5 to 7. In order to evaluate the quality of the planning procedures we need to introduce a cost function that weighs the different quality indicators. This issue will be discussed in due course.

If we compare the different strategies some foreseeable observations can be made. Allowing for more flexibility, either by having more operation hours available or accepting overtime to a limited extent, decreases the number of days planned and undertime. Obviously overtime will be higher if we allow for it but there is a different impact of the method used. For Ophthalmology using the absolute criterion creates more overtime than using the relative

Table 5: Planning 1208 operations Ophthalmology

		Surgeon	Lnorm	Weibull	Loglog	Burr
A	Days planned	197	199	203	200	200
	Undertime (min)	10021	10778	12214	11096	11183
	Undertime (days)	144	144	153	143	145
	Overtime (min)	2441	2238	1754	2076	2163
	Overtime (days)	53	54	49	56	54
	Cancellations (min)	1200	964	368	765	982
	Cancellations (#)	15	12	5	9	12
B	Days planned	269	273	279	272	272
	Undertime (min)	12627	13465	15251	13275	13275
	Undertime (days)	204	212	217	204	204
	Overtime (min)	2767	2165	1791	2335	2335
	Overtime (days)	62	59	58	66	66
	Cancellations (min)	977	512	217	576	576
	Cancellations (#)	14	8	4	9	9
C	Days planned	184	181	184	181	181
	Undertime (min)	5847	4271	5478	4412	4412
	Undertime (days)	101	93	101	95	95
	Overtime (min)	4507	4371	4138	4512	4512
	Overtime (days)	82	87	81	85	85
	Cancellations (min)	2364	1914	2311	1858	1860
	Cancellations (#)	28	24	29	21	21
D	Days planned	176	174	179	174	175
	Undertime (min)	4022	2908	4395	2970	3403
	Undertime (days)	76	58	82	57	63
	Overtime (min)	6522	6368	5455	6430	6383
	Overtime (days)	98	116	96	116	112
	Cancellations (min)	3546	2778	2492	2701	3021
	Cancellations (#)	45	37	36	38	41

A: Available operating hours = 8, no overtime scheduled.
B: Available operating hours = 6, no overtime scheduled.
C: Available operating hours = 8, overtime scheduled with relative criterion.
D: Available operating hours = 8, overtime scheduled with minutes criterion.

Table 6: Planning 423 operations Neurosurgery

		Surgeon	Lnorm	Weibull	Loglog	Burr
A	Days planned	205	238	241	234	234
	Undertime (min)	19314	30704	31876	28939	28939
	Undertime (days)	120	182	188	176	176
	Overtime (min)	12864	8414	8146	8569	8569
	Overtime (days)	85	56	53	58	58
	Cancellations (min)	7178	1435	1590	1385	1386
	Cancellations (#)	40	7	6	7	7
B	Days planned	252	282	290	282	282
	Undertime (min)	17454	24469	27034	24419	24419
	Undertime (days)	139	185	196	185	185
	Overtime (min)	18684	14899	14584	14849	14849
	Overtime (days)	111	96	93	95	95
	Cancellations (min)	9276	6263	6005	5918	5931
	Cancellations (#)	44	15	13	14	14
C	Days planned	158	191	199	189	190
	Undertime (min)	3868	12411	15825	11500	11583
	Undertime (days)	49	100	118	97	99
	Overtime (min)	19978	12681	12255	12730	12333
	Overtime (days)	109	91	80	91	90
	Cancellations (min)	11145	3124	2719	3763	3444
	Cancellations (#)	48	12	15	14	13
D	Days planned	182	215	222	214	215
	Undertime (min)	10475	21486	24331	21064	21493
	Undertime (days)	87	117	125	117	118
	Overtime (min)	15065	10236	9721	10294	10243
	Overtime (days)	95	78	71	78	77
	Cancellations (min)	7713	1985	2186	2307	2307
	Cancellations (#)	46	11	11	12	12

A: Available operating hours = 8, no overtime scheduled.
B: Available operating hours = 6, no overtime scheduled.
C: Available operating hours = 8, overtime scheduled with relative criterion.
D: Available operating hours = 8, overtime scheduled with minutes criterion.

Table 7: Planning 796 operations Gynecology

		Surgeon	Lnorm	Weibull	Loglog	Burr
A	Days planned	207	213	184	210	216
	Undertime (min)	16407	17970	9390	16444	19226
	Undertime (days)	139	165	89	159	168
	Overtime (min)	4356	3039	8379	2953	2855
	Overtime (days)	67	48	95	51	48
	Cancellations (min)	1543	1223	3674	1175	1059
	Cancellations (#)	17	13	43	13	13
B	Days planned	284	288	254	288	294
	Undertime (min)	20575	20205	12437	20225	21977
	Undertime (days)	197	218	142	215	233
	Overtime (min)	5644	3834	8306	3854	3446
	Overtime (days)	85	68	111	71	59
	Cancellations (min)	1827	642	3589	636	478
	Cancellations (#)	22	10	43	10	8
C	Days planned	170	175	158	174	176
	Undertime (min)	3860	4726	2537	4287	4780
	Undertime (days)	65	77	38	74	79
	Overtime (min)	9569	8035	14006	8076	7609
	Overtime (days)	105	98	120	100	97
	Cancelations (min)	4566	3269	7669	3344	2800
	Cancelations (#)	42	28	70	26	24
D	Days planned	179	184	160	185	188
	Undertime (min)	7591	8315	3638	8651	9431
	Undertime (days)	80	86	38	85	94
	Overtime (min)	8980	7304	14147	7160	6500
	Overtime (days)	99	98	122	100	94
	Cancellations (min)	3311	3269	7669	3344	2800
	Cancellations (#)	46	38	79	38	31

A: Available operating hours = 8, no overtime scheduled.
B: Available operating hours = 6, no overtime scheduled.
C: Available operating hours = 8, overtime scheduled with relative criterion.
D: Available operating hours = 8, overtime scheduled with minutes criterion.

criterion. For Neurosurgery and Gynecology it is the other way round. Again this can be attributed to the different mean length of operations. The number of cancellations increases considerably if overtime is allowed. It is hard to decide on what strategy is the best one. As we saw before, low amounts of undertime is accompanied by a relative large amount of over time and many cancellations. To make a decision we need to combine these measures in a cost function.

The differences between the statistical methods considered are large in some instances. The deviations between the results on the basis the loglogistic and the Burr hazard are quite similar, but especially the results of the Weibull hazard are quite different. Especially for Gynecology, the results of the Weibull are very poor. The undertime-score is very good but at the expense of large overtime and many cancellations. The popular lognormal distribution does do better but is not as good as the loglogistic or the Burr.

To make an assessment about the quality of the prediction methods a straightforward way to proceed is define a cost function that combines the quality measures in a single quality measure. Apart from Pandit and Carey (2006), no attempts in this direction appear to have been made, although also Stepaniak et al (2009) and Stepaniak et al (2010) do mention this possibility. The quality measures we will consider are undertime, overtime and the number of cancellations.[11] We will ignore the number of days necessary to program all operations of our prediction sample since this is heavily influenced by the underestimation of the duration of the operations. As we have shown, if we correct for this underestimation, the number of days necessary are quite similar across the prediction methods. Assuming a linear cost function, we have:

$$c = undertime + \gamma_1 \, overtime + \gamma_2 \, cancellations \tag{8}$$

where γ_1 and γ_2 are positive weights. The problem now is to determine these weights. In the optimal situation, hospital managers would give us the information necessary to determine te weights to allow us to make an objective comparison of the prediction and planning methods. Unfortunately we do not have such information and we have to rely on our potentially subjective instincts. We propose to use two sets of weights. The first one, which we will not justify because of its objective nature, is to put γ_1 and γ_2 both equal to 1. In the second cost function we assume that $\gamma_1 = 1.5$ and $\gamma_2 = 2$. Although, given the information we have, it impossible to justify the exact magnitude of these weights, we do believe that $1 \leq \gamma_1 \leq \gamma_2$.

[11] Pandit and Carey (2006) only consider overtime and cancellations.

The problem with undertime is that the operating room is possibly vacant for some time, but since there is no time pressure, it is unlikely that there will be repercussions on the quality of the operation. In the case of substantial undertime, fewer operations will be scheduled than in the optimal situation, and this might have financial consequences to the hospital as well. Depending on the demand for operations, the number of operation rooms available and the method of planning of operations, undertime in the case of a particular specialty might also have consequences on the planning of the operations of other specialties. An advantage of undertime is that emergency operations are more easily accommodated. The consequence of overtime are more severe. Since an operation can not be stopped halfway the operation there is no other option than to proceed. The result of overtime will be the postponement or even cancellation of other operations, a reduction of the quality of the operations due to the time pressure and additional financial costs because the operation staff has to prolong there working day. The first disadvantage is comparable to the main disadvantage of undertime, but the others are additional. Consequently we believe that $\gamma_1 > 1$. Cancellations affect the reputation of hospitals and more importantly the mental well being of the patients. On top of that, if an operation is canceled, it usually will have to be rescheduled within a couple of days. This will cause additional strain on the operating schedule that might result in overtime or the necessity to put extra slack in the schedule. It is quite hard to weigh the reputation and mental effects with the more financial consequences but we believe, but in this case it is basically only belief, that γ_2 is even higher than γ_1. Finally, we measure the costs in minutes. The alternative of measuring in days gives very similar results.

Table 8: Relative cost measures Ophthalmology

		Surgeon	Lnorm	Weibull	Loglog	Burr
A	$\gamma_1 = \gamma_2 = 1$	1.000	1.023	1.049	1.020	1.049
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.999	0.969	0.979	1.019
B	$\gamma_1 = \gamma_2 = 1$	1.000	0.986	1.054	0.989	0.989
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.947	0.981	0.957	0.957
C	$\gamma_1 = \gamma_2 = 1$	1.000	0.830	0.938	0.848	0.848
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.845	0.941	0.859	0.860
D	$\gamma_1 = \gamma_2 = 1$	1.000	0.855	0.876	0.859	0.909
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.862	0.840	0.862	0.910

A: Available operating hours = 8, no overtime scheduled.
B: Available operating hours = 6, no overtime scheduled.
C: Available operating hours = 8, overtime scheduled with relative criterion.
D: Available operating hours = 8, overtime scheduled with minutes criterion.

If our evaluation of the relative importance of the three arguments of the cost function is correct, the equal weight cost function will favor the planning based on the surgeons' expectation of the length of the operations for the neurosurgical and gynecological specialties. The alternative specification will favor the statistical methods for these disciplines. Tables 8 (Ophthalmology), 9 (Neurosurgery) and 10 (Gynecology) present the relative costs of the planning methods for the four different planning strategies we considered earlier. Since we have a relative measure, we normalize on the costs following from planning according to the surgeons' assessments of the duration of the operations.

For Ophthalmology (Table 8) we find that for the planning strategies that do not allow for overtime the differences in costs across the methods are small. In some cases the surgeons do better but in other cases several statistical methods do better. In the planning strategies that do allow for overtime, the statistical methods outperform the surgeons. In that case, a cost reduction of more about 15% can be achieved. Note that there are no large differences between the two cost functions we employ. Furthermore, note the relative good performance of the planning based on the predictions of the lognormal distribution.

Table 9: Relative cost measures Neurosurgery

		Surgeon	Lnorm	Weibull	Loglog	Burr
A	$\gamma_1 = \gamma_2 = 1$	1.000	1.030	1.057	0.988	0.988
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.872	0.893	0.841	0.841
B	$\gamma_1 = \gamma_2 = 1$	1.000	1.011	1.049	0.995	0.995
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.936	0.951	0.914	0.914
C	$\gamma_1 = \gamma_2 = 1$	1.000	0.806	0.909	0.800	0.782
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.671	0.742	0.679	0.659
D	$\gamma_1 = \gamma_2 = 1$	1.000	1.014	1.090	1.012	1.024
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.841	0.893	0.848	0.855

A: Available operating hours = 8, no overtime scheduled.
B: Available operating hours = 6, no overtime scheduled.
C: Available operating hours = 8, overtime scheduled with relative criterion.
D: Available operating hours = 8, no overtime scheduled with minutes criterion.

For Neurosurgery the choice of the cost function does matter. This is due to the heavy underprediction of the length of operations. Whatever planning strategy and what ever cost function that is used there is allows a statistical method with lower costs. Especially for the planning methods not allowing for overtime and the equal-weight cost function the differences are really small, but for the other methods the differences are quite large. A cost reduction of 10% or more is not exceptional.

As we experienced earlier, the results for Gynecology lie somewhere in between. Again there is always a statistical method with lower costs than making a planning on the basis of the surgeons' expectations. The bad performance of the Weibull hazard is again striking. The loglogistic hazard performs very well. It indicates that a cost reduction of 5.7% t to 16.4% is possible. Also in this case the choice for the weights in the cost function appear to be non-consequential. Whatever the weights, a cost reduction of 6% or more is possible by applying a statistical method.

Table 10: Relative cost measures Gynecology

		Surgeon	Lnorm	Weibull	Loglog	Burr
A	$\gamma_1 = \gamma_2 = 1$	1.000	0.997	0.961	0.922	1.037
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.960	1.126	0.892	0.985
B	$\gamma_1 = \gamma_2 = 1$	1.000	0.880	0.868	0.881	0.924
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.833	0.981	0.834	0.860
C	$\gamma_1 = \gamma_2 = 1$	1.000	0.891	1.346	0.873	0.844
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.853	1.422	0.844	0.797
D	$\gamma_1 = \gamma_2 = 1$	1.000	0.934	1.242	0.943	0.940
	$\gamma_1 = 1.5;\ \gamma_2 = 2$	1.000	0.909	1.397	0.913	0.893

A: Available operating hours = 8, no overtime scheduled.
B: Available operating hours = 6, no overtime scheduled.
C: Available operating hours = 8, overtime scheduled with relative criterion.
D: Available operating hours = 8, no overtime scheduled with minutes criterion.

6. Conclusion

We have investigated the planning of operations in the Academic Medical Center for three different specialties. At present, the operations are scheduled according to the surgeon's estimation of the case duration. The average length of the operations performed by the Ophthalmology, Neurosurgery and Gynecology departments are quite different and in general we see that the longer an operation lasts the more difficult it is for the surgeon to predict the length of the operation correctly. Moreover especially in the Neurosurgery department and to a lesser extent in the Gynecology department, the surgeons seriously underpredict the duration of operations. We have investigated the potential of several statistical methods to see whether they do a better job than the surgeons with respect to predicting the duration of operations correctly. In many cases this appears to be the case. Moreover in the future, the prediction period can be extended and the statistical estimations will probably be even more accurate.

In the literature the lognormal model is proposed as an adequate method to represent the duration of operations. From our investigation it follows that this choice, especially for longer durations, is not the optimal one. Especially the Burr distribution, or its special case the loglogistic distribution, appears to be more suitable in many situations.

Due to the complexity of the planning algorithm used by the AMC we were unable to apply it directly to our results. We created four alternative planning strategies that we use to quantify the effect of more accurate predictions of case durations on undertime, overtime and cancellations. Whatever strategy is used, significant cost reductions appear to be possible. Also, the specific functional form of the cost function utilized does not appear to be very important.

We did not engage in further fine tuning of the statistical methods. For instance, it might be worthwhile to define subclasses of expected case durations and to optimize per subclass. We could distinguish short/medium/long expected durations, according to frequencies of types of operations or according to the number of procedures in the operation. Dexter and Zhou (1998) indicates that this is a useful way to proceed. A brief investigation on our own data has shown us that there indeed is some potential here.

Finally, we want to state that the surgeons' expectations of the case duration is far from worthless. This expectation is an important explanatory variable in our statistical models. Our recommendation, therefore, is not to use statistical methods exclusively, but only in combination with information supplied by the surgeon.

References

Bago d'Uva T, Jones AM Health care utilization in Europe: New evidence from the ECHP. Journal of Health Economics 2009; 28; 265-279.

Benchmarking OR. Benchmarking: Een Kwestie van Leren, digital publication on URL: www.benchmarking-ok.nl; 2008.

Cameron AC, Trivedi PK. Microeconometrics, Cambridge University Press; 2005.

Chiappori P-A, Durand F, Geoffard P-Y. Moral hazard and the demand for physician services: Firste lessons from a French natural experiment. European Economic Review 1998; 42; 488-511.

Dexter F, Zhou J. Method to Assist in the Scheduling of Add-on Surgical Cases. Anesthesiology 1998; 89; 1228-1232.

Dexter F, Macario A, Ledolter J. Identification and Systematic Underestimation (bias) of Case Durations During Case Scheduling Would Not Markedly Reduce Over-utilized Operating Room Time. Journal of Clinical Anesthesiology 2007; 19; 198-203.

Dexter F, Dexter EU, Masursky D, Nussmeier NA. Systematic Review of General Thoraric Surgery Articles to Identify Predictors of Operating Room Case Duration. Anaesthesia & Analgesia 2008; 106; 1232-1241.

Eijkemans MJC, van Houdenhoven M, Nguyen T, BoersmaE, Steyerberg EW, Kazemier G. Predicting the Unpredictable: A New Prediction Model for Operating Room Times Using Individual Characteristics and the Surgeon's Estimate. Anesthesiology 2010; 12; 41-49.

Lancaster T. The Econometric Analysis of Transition Data. Cambridge University Press; 1990..

Macario A, Vites TS, Dunn B, McDonald T. Where Are the Costs in Perioperative Care?: Analysis of Hospital Costs and Charges for Inpatient Surgical Care. Anaesthesiology 1995; 83;1138-1144.

Okunade AA, Murthy VNR. Technology as a 'major driver' of health care costs: a cointegration analysis of the Newhouse conjecture. Journal of Health Economics 2002; 21; 147-159.

Pandit JJ, Carey A. Estimating the Duration of Common Elective Operations: Implications for Operating List Management. Anesthesia 2006; 1; 768-776.

Rossiter CE, Reynolds JA. Automatic Monitoring of the Time Waited in Out-patient Departments. Med Care 1963; 1; 218-225.

Stepaniak PS, Heij C, Mannaerts GH, de Quelerij M, de Vries G. Modeling Procedure and Surgical Times for Current Procedural Terminology-Anesthesia-Surgeon Combinations and Evaluation in Terms of Case-Duration Prediction and Operating Room Efficiency: a Multicenter Study. Anesthesia & Analgesia 2009; 109; 1232-1245.

Stepaniak, PS, Heijand C, de Vries G. Modeling and Prediction of Surgical Procedure Times. Statistica Neerlandica 2010; 64; 1-18.

Strum, DP, May JH, Vargas LG. Modeling the Uncertainty of Surgical Procedure Times. Anesthesiology 2000a; 94; 1160-1167.

Strum DP, Sampson AR, May JH, Vargas LG. Surgeon and Type of Anaesthesia Predict Variability in Surgical Procedure Times. Anesthesiology 2000b; 92; 1454-1466.

Strum DP, May JH, Sampson AR, Vargas LG, Sprangler WE. Estimating Times of Surgeries with Two Component Procedures. Anesthesiology 2003; 98; 232- 240.

Van Houdenhoven M, van Oostrum JM, Hans EW, Wullink G, Kazemier G. Improving Operating Room Efficiency by Applying Bin-Packing and Portfolio Techniques to Surgical Case Scheduling. Anesthesia & Analgesia 2007; 105; 707-714.

Wullink, GM, van Houdenhoven M, Hans EW, van Oostrum JM, van der Lans M, Kazemier G. Closing Emergency Operating Rooms Improves Efficiency. Journal of Medical Systems 2007; 31; 543-546.

Appendix:
The explanatory variables used in the estimation of the durations.

The explanatory variables can be categorized in five groups.

Operation characteristics:

- *Procedure* (*x* times). This dummy variable is equal to 1 for the procedure it is named after. For each procedure that is investigated there is one variable like this.
- *surgeon* (*x times*). This binary variable is equal to 1 if *surgeon* is the first surgeon of a case. Each operating staff member or senior assistant that was still operating in 2008 has a separate variable. (Co-)Assistants are therefore not included as well as retired or departed staff, for the sake of parsimony. Their inclusion is required in theory to determine the correct effect of the other surgeons on duration. In practise however we have not noticed any positive effect of their inclusion on prediction.
- *Anaescode*. This categoric variable indicates the type of anaesthetic and is 0 if anaesthesia was monitored or no technique was reported in OKPlus. Furthermore, it is 1 if anaesthetics are inducted locally, 2 if anaesthetics are inducted regionally and 3 if anaesthetics are inducted totally. Obviously duration increases with *anaescode*.
- *Monitor*. It is a binary variable equal to 1 if anaesthesia was monitored.

Session characteristics:

- *No_anaes*. This is a binary variable equal to 1 if no anaesthesiology is reported (excluding the initial period of January 2003 till October 2004 for which a separate variable is defined). It is generated to exploit potential information about the duration of a case present in the fact that the type of anaesthesia is not reported. First of all no report could simply mean that no anaesthetics were inducted. Perhaps other reasons exist as well however.
- *No_anaesreg*. It is a binary variable equal to 1 for the initial period of January 2003 until October 2004 in which anaesthesiology was not reported at all.
- *Totprocs*. This is the total number of surgical procedures within a single case. It is the only variable used together with the previous to describe the surgical part of a case. Second and third procedures are left unidentified thereby, mainly for the sake of parsimony (see the discussion in section 3.3).

Team characteristics:

- *Experience*. This variable is defined only for Neurosurgery to separate personnel into four classes of experience, 1 the least experienced until 4 most experienced. It may perhaps serve as a parsimonious replacement of the surgeon dummy-variables. The specialty has divided personnel over these *static* classes itself, not using strict definitions for each class.
- *Age_oper*. The inclusion of the age of the surgeon is intended to capture the time-effect in experience of an surgeon and the influence thereof on duration. An surgeon is likely to become faster, especially in the beginning of his career (see Houdenhoven (2007). *Age_oper* is zero if the age of an surgeon is missing.
- *No_age*. This is a binary variable equal to 1 if *age_oper* is missing.
- *D_oper2*. This is a binary variable equal to 1 if a second surgeon is present during a case.

Patient characteristics:

- *Compli_code*, *Pulmon_code*, *cardia_code*, *allerg_code*, *gencond_code*. These are four categoric variables indicating the medical condition of a patient in 3 levels. These characteristics are registered by and of special interest for anaesthesiologists. The variables are set equal to zero if not reported.
- *No_compl*. This is a binary variable equal to 1 if the above information is missing. Either all four variables are reported or they are not.
- *Sober*. This binary variable is equal to 1 if a patient is sober. Again, this is information used by anaesthesiologists.
- *Asacode*. This is a variable indicating the condition (ASA) of the patient from 1 (good) to 5 (lethal).
- *No_asa*: This binary variable is 1 if *asacode* is missing.
- *Age_patient*.
- *Weight*. The weight of the patient is set equal to average weight if missing.

Other characteristics:

- *Location*. This is a binary variable designed to discriminate between cases on the '*daily*' and the clinical OR. It is equal to 1 for cases conducted in the clinical OR.
- *Dur_pl*. This is planned case duration. It is included because it reflects the beliefs of surgeons about the duration (even if surgeons tend to underpredict structurally). It may therefore contain information the surgeon has that is not reported. A drawback of the inclusion of this variable is that it allows surgeons to influence predictions. New models would have to be estimated every now and then to neutralize this effect.
- *First*. This is a binary variable equal to 1 if a case start between 7.50am and 8.10am, meaning the case is the initial case of the day. Initial cases often delay because part of the OR personnel is late. The variable allows for such an effect.
- *Time*. This is a count variable counting the days between operating and the 1st of January 2003. This variable is included to capture time-trends in OR case duration induced by technological progress for example.