Can statisticians beat surgeons at the planning of operations?

Joustra, P.E.; Meester, R.; van Ophem, J.C.M.

Can statisticians beat surgeons at the planning of operations?[1,2]

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
Faculty of Economics and Business
University of Amsterdam

January 2011

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. Our empirical results indicate that statistical methods often do better than surgeons. 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 under- or overestimating the actual length of operations.

---

[1]The comments of two anonymous referees are gratefully acknowledged. 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.

[2]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/).

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. 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.

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.

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 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.

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

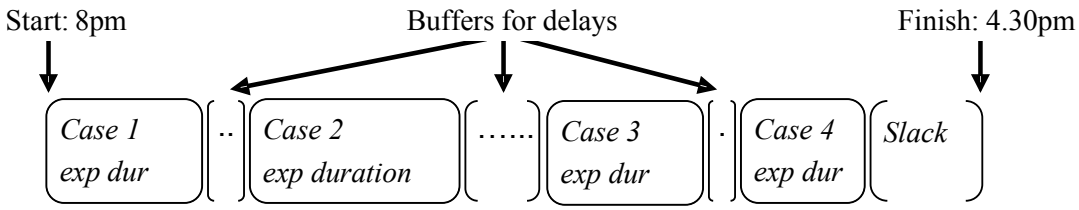
A daily OR program consists of elective cases and ambulatory cases. In this paper we define elective cases as all those cases that can be planned up to 10.30 am the day before, when the final planning has to be ready for the next day. Ambulatory cases are all cases coming through after that time. For 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 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[3] 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

[3]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.

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).

Figure 1. Graphical representation of daily planning.



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).

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 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 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 the 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:[4]

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

The Burr-hazard model is a '*mixture*' model and nests the Weibull and loglogistic hazard models. 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. The Weibull distribution is obtained by letting $\sigma^2\to 0$, thereby losing the unobserved heterogeneity part.[5] The loglogistic distribution is yet another special case that can be obtained by putting $\sigma^2=1$. Unobserved heterogeneity might be an important addition to the model because of

[4] In an earlier version of this paper (Joustra et al 2010) we also report results on the exponential hazard model and on an alternative specification of the piecewise constant hazard model.

[5] The exponential distribution is a special case of the Weibull distribution and can be obtained by setting $\alpha = 1$.

e.g. the occurrence of complications during surgery. Apart from the loglogistic and the Burr distribution, the generalized gamma (discussed below) distribution also allows for unobserved heterogeneity. All other distributions used in this analysis do not.

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:

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

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_i 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:

$$f(t) = \alpha \phi_i^{\alpha m} t^{\alpha m - 1} \exp -((\phi_i t)^{\alpha m}) / \Gamma(m). \tag{5}$$

where $\Gamma(m)$ is the gamma function. α (≥ 0), m (> 0), and ϕ_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 $m = 1$, to the lognormal density if $\alpha = 0$ and to the exponential density if both $\alpha = 1$ and $m = 1$.

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:

$$\log(t) = \beta' x_i + u_i \tag{6}$$

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 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 $\lambda_0(t)$ 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

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

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. We will use a time interval of 10 minutes.[6]

We estimate the predicted duration of an operation by the expected duration calculated from the ML-estimation. The expected durations are given by the following expressions:[7]

$$\begin{aligned} E(t_{Burr})&=\left(\phi_i^{\frac{-1}{\alpha}}\right)\frac{\Gamma(1+\frac{1}{\alpha})\Gamma(\frac{1}{\sigma^2}-\frac{1}{\alpha})}{\sigma^{2(1+\frac{1}{\alpha})}\Gamma(\frac{1}{\sigma^2}+1)} \\ E(t_{gen\Gamma})&=\frac{\Gamma(m+\frac{1}{\alpha})}{\phi_i\Gamma(m)} \end{aligned} \tag{8}$$

In order to calculate the expected duration of the piecewise constant hazard model we need to introduce some notation first. Given the duration of the operation t_i and the length of the time interval Δt, we can calculate m_i as follows:

$$t_i=m_i\Delta t+(t_i-m_i\Delta t) \text{ where } 0\le t-m_i\Delta t<\Delta t \tag{9}$$

m_i defines the relevant j in eq. (7) for each observation. Using standard results on the relation between the hazard and the distribution function (cf. Cameron and Trivedi, 2005, p. 576-7) and partial integration, we get:

$$E(t_{PCH})=\frac{\exp\left(-\phi_i\left(\sum_{j=1}^{m_i}\exp(\alpha_j)\Delta t-\exp(\alpha_{m_i+1})m_i\Delta t\right)\right)}{\phi_i\exp(\alpha_{m_i+1})} \tag{10}$$

[6]We also investigated a 5 minutes time interval. The results did not improve compared to the model with the 10 minutes time interval.

[7] Cf. p. 68 of Lancaster (1990). The expected durations of the Weibull, loglogistic and lognormal hazards are special cases of the ones listed here.

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:[8]

- 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

Performance is optimal when an unknown 'loss function' is minimized. This loss function will depend on factors like the ones listed above. To evaluate the quality of the prediction methods we have to depend upon these factors in combination. This is unlikely to lead to clear cut and completely objective conclusions, but we believe that we are able to at least give a strong indication to what prediction method to prefer.

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 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

[8]If we denote the actual duration of operation i by t_i and the predicted duration by $\hat{t}_i$, the performance measured are calculated as follows: $MEAN = \sum \hat{t}_i / N$, $AD = \sum (\hat{t}_i - t_i)/N$, $ADD = \sum |\hat{t}_i - t_i|/N$, $rMSE = \sqrt{(\hat{t}_i - t_i)^2 / N}$, $UPx = \sum 1(\hat{t}_i - t_i > x)$, $OPx = \sum 1(\hat{t}_i - t_i < -x)$ and where the number of predicted operations is N and where l(condition) is 1 if the condition is true and 0 otherwise.

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.

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

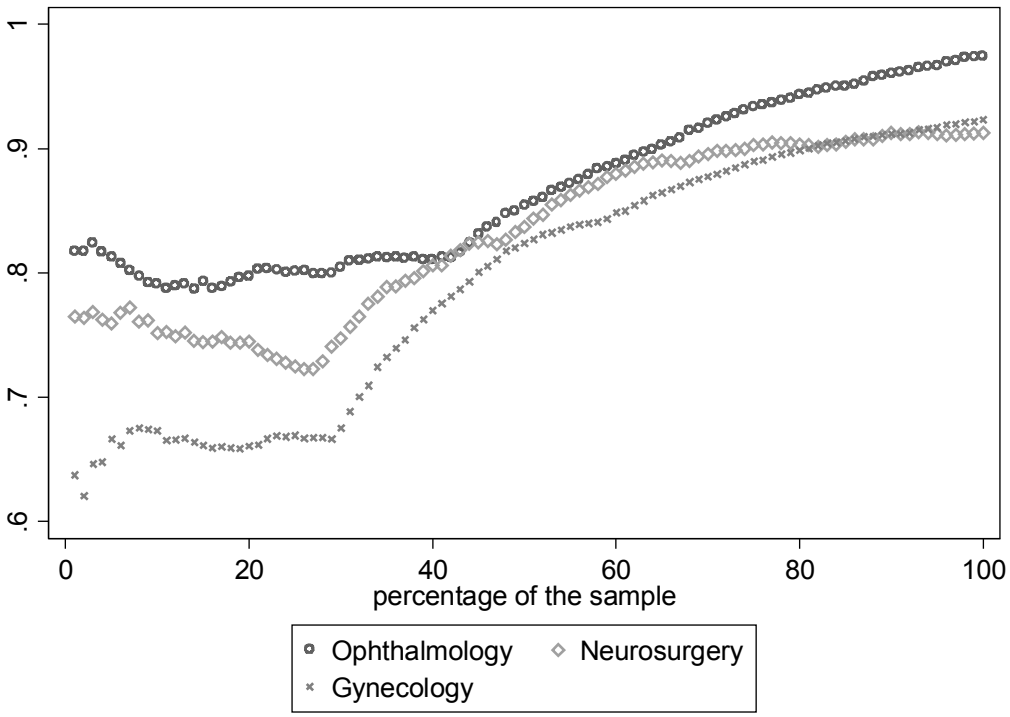
	Ophthalmology		Neurosurgery		Gynecology	
	Estimation sample	Prediction sample	Estimation sample	Prediction sample	Estimation sample	Prediction sample
Nr of obs	4092	1208	1863	423	3472	796
Actual duration						
Mean	75.6	72.0	245.0	217.4	110.5	109.7
Stand. dev.	41.3	37.0	178.2	183.2	97.2	93.9
Minimum	6	11	20	26	10	7
Maximum	735	397	1544	1115	863	775
Planned duration by the surgeon						
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.

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 in the estimation period 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. Although the sample statistics differ for two periods distinguished, the conclusions drawn before hold also for the prediction sample.

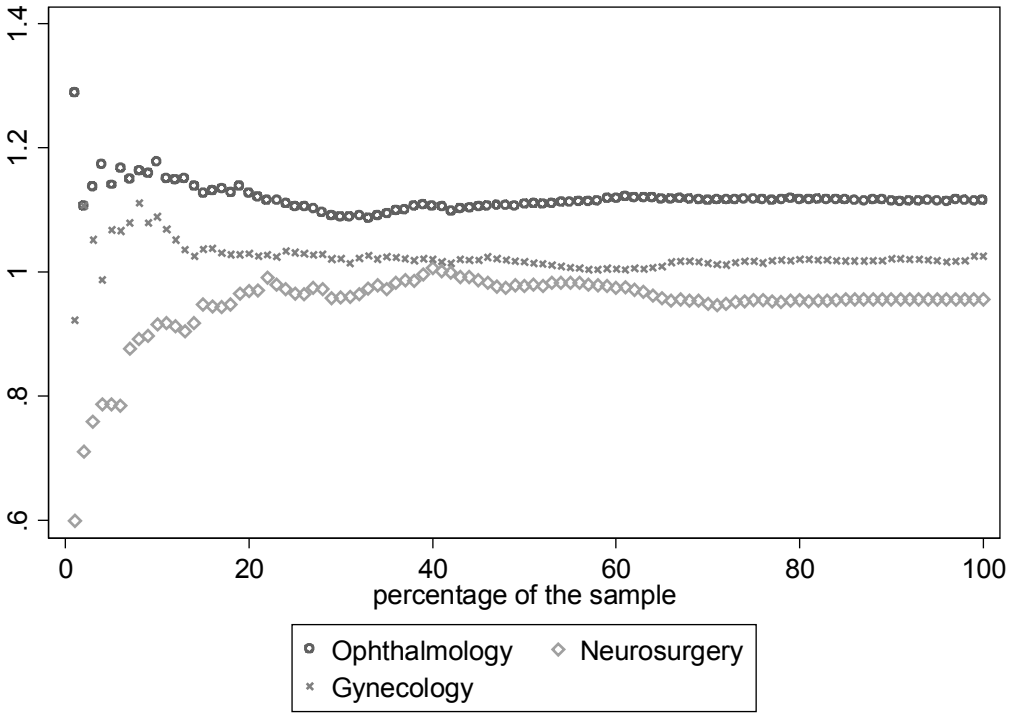
Figure 2: The average fraction of the predicted and actual durations of operations as a function of sample size of the estimation and prediction period.



In terms of underprediction, the surgeons of Ophthalmology and Gynecology the surgeons appear to predict the durations of the operations somewhat better in the prediction

sample than in the estimation sample. Figure 2 reveals that this is also true for Neurosurgery. In this figure we have depicted the averages of the fraction of the predicted and actual durations calculated cumulatively for each integer percentage of the complete sample ordered across time. So for the combined estimation and prediction sample, the average fraction is calculated for the first 1%, the first 2% up to 100% of the sample and depicted in Figure 2. If we consider the first 30% of the sample, we find that the mean of the average fraction is more or less constant for each of the three specialties but from that point on the average fraction starts moving towards 1.[9] The predicted duration by the surgeons divided by the actual duration varies from about 0.65 (Gynecology) to 0.8 (Ophthalmology). For the full sample we find fractions of about 0.95 (Ophthalmology) and 0.9 (Neurosurgery and Gynecology) This indicates that the quality of the surgeon's predictions are better in the prediction sample than in the estimation sample.

Figure 3: The average fraction of the predicted and actual durations of operations as a function of sample size of the prediction period.



Since the estimation sample is much larger than the prediction sample, the steady increase of the fraction predicted duration/actual duration, also reflects that in the prediction

[9] The 30% sample size corresponds to the period October 2004. Apparently, something changed in the prediction methods of the surgeons for all three disciplines. The exact reason for this is unknown to us, but one way or another surgeons were stimulated to make better predictions.

sample, the surgeon's predictions are even better than might be concluded from Figure 2. To investigate this further, consider Figure 3. It represents exactly the same information as used in Figure 2, but now only for the prediction period. From about 20% of the sample, the fractions are remarkably stable at about 1.12 (Ophthalmology), 0.95 (Neurosurgery) and 1.02 (Gynecology). Using this measure we even find that for two out of three specialties, the surgeons overpredict the duration of operations.[10] Again, the overall conclusion has to be that the surgeon's predictions have become much better in time. As a result our statistical prediction methods have to compete with the relatively better predictions of the surgeons. The improvement of the surgeon's predictions is likely to be due to the AMC putting more emphasis on the importance of good estimation of operation duration in latter years.[11]

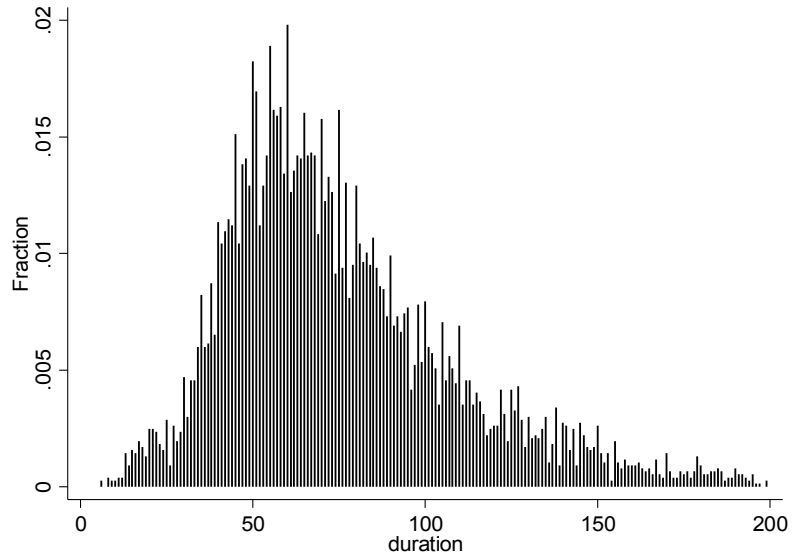
Apart from the distributional assumptions underlying any econometric regression model, the dependent variables of the model are the most important factors to explain (or describe) the differences in case duration. Since our efforts are aimed at predicting operation durations as good as possible we will 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, these expectations appear to be biased (see Table 1 and Figures 2 and 3). Note that the surgeon shares all the information we have, but has even more information because some information on e.g. the urgency of the operation and on the patient's health is not recorded.

[10]This conclusion appear to contradict the Table 1, where we find underprediction for Neurosurgery and Gynecology. However, if we weigh with duration the fractions are almost identical to the ones that can be deduced from Table 1. This indicates that larger prediction errors are made for longer durations.

[11] E.g., 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.

There are a few problems in the data that we need to discuss here. 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.

Figure 4: Spike plot of ophthalmologic operation duration.



Another complication in the data available is the prevalence of measurement errors both in the dependent variable and 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 4), 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.

Another 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 of 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.

A final 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 Van 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%.[12]

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

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 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	Lnorm	Weibull	Loglog	Burr	GenΓ	PCH10
MEAN	72.13	71.75	72.96	71.55	71.63	71.80	72.65
AD	0.13	-0.25	0.96	-0.35	-0.38	-0.20	0.65
AAD	18.62	15.47	16.25	15.34	15.35	15.46	16.15
rMSE	25.81	23.05	23.68	22.99	23.00	23.04	23.85
UP10	0.26	0.23	0.22	0.24	0.23	0.23	0.23
UP20	0.17	0.13	0.13	0.14	0.14	0.14	0.14
UP30	0.10	0.07	0.07	0.08	0.08	0.07	0.08
OP10	0.38	0.30	0.33	0.29	0.29	0.30	0.33
OP20	0.17	0.12	0.14	0.12	0.12	0.13	0.14
OP30	0.06	0.05	0.06	0.04	0.04	0.05	0.06

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.[13] The definition of the measures as well as the models applied are discussed in 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 first of all that all prediction methods are quire accurate in terms of the average duration predicted, where the surgeons score best. With respect to the other prediction measures, the differences are more pronounced and always in favor of the statistical prediction methods. For the absolute deviations, the prediction error is ranging from about 15.3 minutes (loglogistic hazard) to 18.6 minutes (surgeons), a difference of nearly 18%. In terms of this measure, two models distinguish themselves favorably: the Burr and the nested loglogistic hazard model. The differences with generalized gamma and the lognormal are relatively small. With respect to the under- and overprediction, the hierarchy of the results are very similar, although the

[13] We will only present estimation results based on the generalized gamma distribution for Ophthalmology. For the other specialties, α was estimated to be negative and as a result E(t) can not be calculated.

results are even closer. The Weibull may perform quite well in terms of underprediction, but this result is offset by the relatively poor performance with respect to overprediction. 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 and loglogistic hazard. Overall, the loglogistic model appears to perform best.

Table 3: Prediction measures Neurosurgery (423 operations)

	Surg	Lnorm	Weibull	Loglog	Burr	PCH10
MEAN	184.67	233.26	250.21	230.97	231.33	289.32
AD	-32.70	15.88	32.83	13.60	13.96	71.94
AAD	68.29	58.46	70.53	56.15	56.23	105.90
MSE	103.14	104.94	135.81	99.19	99.14	454.30
UP10	0.51	0.34	0.30	0.34	0.34	0.30
UP20	0.44	0.26	0.22	0.25	0.25	0.21
UP30	0.40	0.20	0.16	0.21	0.21	0.16
OP10	0.33	0.46	0.51	0.47	0.47	0.52
OP20	0.25	0.35	0.43	0.35	0.35	0.42

Shaded entries represent the best result across the row. 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. Surgeons underestimate the duration of neurological by more than half an hour on average or with about 15%. A striking result is that the statistical methods appear to overpredict the duration in our estimations, although in a much less serious manner than the underprediction of the surgeons. Part of the explanation might be the large difference 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). The best result obtained is for the loglogistic model yielding an overprediction of the total operation time of on average 14 minutes or with 6.2%. The Weibull and piecewise-constant hazard perform even worse than the surgeons in this respect. The absolute average deviations are closer, but still most statistical methods outperform the surgeons considerably. Here the difference between the most accurate models and the planning of surgeons is approximately 12 minutes or 18%. As for Ophthalmology, the Burr, the loglogistic and the lognormal model appear to outperform the other methods. In terms of under- and overprediction the results are very

similar as was encountered before. In the opposite direction, the surgeons appear to score very well at the overprediction percentages, but this is a result of the strong tendency to underpredict of surgeons. Overall, the Burr and the loglogistic models seem to obtain the best scores and their scores are quite similar. The Weibull model might be preferred if underprediction is considered to be a very serious error. As before, the lognormal hazard stays somewhat behind on the Burr and the loglogistic model.

Table 4: Prediction measures Gynecology (796 operations)

	Surg	Lnorm	Weibull	Loglog	Burr	PCH10
MEAN	103.06	106.15	94.56	105.73	107.11	98.75
AD	-6.62	-3.54	-15.13	-3.95	-2.58	-10.93
AAD	26.02	22.63	29.05	22.50	22.55	29.48
MSE	45.78	42.47	48.29	42.40	42.33	59.19
UP10	0.38	0.29	0.44	0.29	0.28	0.39
UP20	0.25	0.20	0.32	0.20	0.19	0.28
UP30	0.17	0.14	0.24	0.14	0.13	0.21
OP10	0.25	0.31	0.23	0.31	0.32	0.25
OP20	0.14	0.14	0.12	0.12	0.14	0.14
OP30	0.08	0.07	0.06	0.07	0.07	0.08

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. Only with respect to the first overprediction class, surgeons perform relatively well. With respect to all other measures, the predictions by the surgeons are outperformed by at least three statistical methods. In terms of AAD, the difference between the most accurate method (loglogistic) and the planning of the surgeons is approximately 3.5 minutes or 13%. Overall, the best predictions are found for the Burr hazard model. The loglogistic hazard performs almost as good as the Burr.

An illustration of 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 misprediction 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 too 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 and apply it to Gynecology.[14] 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 that 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 actual order of the operations is determined on the basis of these expectations.

We will adopt one simple planning strategy: we plan up to six hours per day and overtime is never allowed, except for the first operation that day. We limited planned operations to six hours to allow for some slack at the end of the day.[15]

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 with idle time of the operation room (denoted by 'Undertime') , the number of minutes of with overplanning of the operation room (denoted by 'Overtime') and the number of times an operation had to be canceled (denoted by 'Cancellations'). 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.[16]

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 loglogistic and the Burr hazard).

[14]Some alternative planning strategies are considered in Joustra et al (2010). The main conclusions do not differ from the ones presented here. Results for the other specialties can be found there as well.
[15]We also investigated a planning based on 7 and 8 hours a day. In that case the conclusions are quite similar.
[16] Changing the cancellation policy by putting e.g. the relative factor to 1, does not have a consequential impact on the conclusions.

Table 5 presents some characteristics of the complete planning of the operations in the prediction period for Gynecology specialty. An important indicator of the quality of the planning is the number of days necessary to program all operations. The surgeons appear to do a little better than the statistical methods, except for the planning based on the Weibull predictions. However, for a fair comparison, account should be taken of the relative large amount of overtime generated by the surgeon's predictions. To win 4 days, surgeons increase overtime by about 1800 minutes, or about 5 full 6-hour days. With respect to undertime three out of four statistical methods perform better than the surgeons, although the difference is not large. The number of cancellations is best for the Burr distribution, although the lognormal and loglogistic score more or less the same.

Table 5: The planning of operations for Gynecology (796 observations)

	Surgeon	Lnorm	Loglog	Burr
Days planned	284	288	288	294
Undertime	20575	20205	20225	21977
Overtime	5644	3834	3854	3446
Cancellations	22	10	10	8

To actually make an assessment about the quality of the prediction methods a straightforward way to proceed is to define a cost function that combines the quality measures in a single cost 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. Assuming a linear cost function, we have:[17]

$$c = undertime + \gamma_1 overtime + \gamma_2 cancellations + \gamma_3 days\ \ planned \qquad (11)$$

where γ_1, γ_2 and γ_3 are non-negative weights. The problem now is to determine these weights. In the optimal situation, hospital managers would give us the information necessary to determine the weights to allow us to make an objective comparison of the prediction and planning methods. Unfortunately we do not have such information. What we can conclude is that it is quite likely that some statistical methods result in lower costs because they score better at three out of four elements cost function (11). The planning based on the lognormal and the loglogistic score better on undertime, overtime and cancellations than the planning

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

based on the surgeon's predictions and score not much worse than surgeon's with respect to the number of days planned. In percentages the two statistical methods score 1% better at undertime, 45% better at overtime and 120% better at cancellations while scoring 2% worse at the number of days planned. The conclusion that cost reduction can be achieved by using statistical methods in the planning of operations does not seem to be unrealistic.

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 prediction method, although the differences are not very large.. The Burr distribution, or its special case the loglogistic distribution, appears to perform slightly better. Both these distributions allow for unobserved heterogeneity.

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 vital. 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.

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, Boersma E, 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.

Joustra P, Meester R, Van Ophem, H. Can Statisticians Beat Surgeons at the Planning of Operations, Discussion paper 2010/06, UvA-Econometrics, Amsterdam School of Economics, University of Amsterdam 2010.

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.

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 Van 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.