Determining Optimum Operating Room Utilization

Donald C. Tyler, MD, MBA*, Caroline A. Pasquariello, MD*, and Chun-Hung Chen, PhD†

*Department of Anesthesiology and Critical Care Medicine, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania; and †Department of Systems Engineering & Operations Research, George Mason University, Fairfax, Virginia

Economic considerations suggest that it is desirable to keep operating rooms fully used when staffed, but the optimum utilization of an operating room (OR) is not known. We created a simulation of an OR to define optimum utilization. We set operational goals of having cases start within 15 min of the scheduled time and of having the cases end no more than 15 min past the scheduled end of the day. Within these goals, a utilization of 85% to 90% is the highest that can be achieved without delay or running late. Increasing the variability of case duration decreases the utilization that can be achieved within these targets.

(Anesth Analg 2003;96:1114–21)

Extensive operating room (OR) utilization is a goal of OR directors and hospital administrators. Unfortunately, the optimum level of utilization is not well defined, nor is it clear what tradeoffs are required to achieve optimum utilization. We sought to identify those factors that influence optimum utilization. The purpose of this article is to use a simple OR simulation to enumerate the important factors that must be considered in determining optimum utilization.

The classic definition of OR utilization is the sum of the time it takes to perform each surgical procedure (including preparation of the patient in the OR, anesthesia induction, and emergence) plus the total turnover time, divided by the time available (1,2). As an example, if the average "patient in to patient out" time for a herniorrhaphy is 45 min and the average turnover time is 15 min, then 10 herniorrhaphy cases can be performed in a 10-h period in that OR, for an OR utilization of 100%. With this definition, if cases extend beyond the scheduled end of the day, the time used after the scheduled end of the day is counted as utilization, even though the hospital may be paying overtime to provide the staffing.

Strum et al. (1) defined the concepts "overutilization" and "underutilization." Underutilization is defined as time during the scheduled hours of operation that is not used, and overutilization is defined as the time used by scheduled cases past the end of the scheduled time. With these concepts we can estimate the economic efficiency of an OR suite, as described in Methods.

The standard definition produces the actual utilization—the time that is actually used. Because it is necessary to know the actual case times to perform the calculation, utilization can never be known in advance. In this analysis, we also refer to the scheduled utilization, that is, the predicted utilization obtained when cases are scheduled.

To define what happens when utilization becomes excessive, consider one OR that is currently scheduled with inadequate utilization. We are asked to schedule additional cases, even if "overbooking" is necessary. As additional cases are added, a point is reached at which the expected case times go beyond the scheduled end of the day, thereby exceeding the allowed utilization. To schedule more cases, either the schedule can be allowed to extend beyond the end of the day, or the expected duration of each case can be arbitrarily shortened and more cases squeezed into the allotted time. Although these manipulations may result in maximum scheduled and actual utilization, the schedule may run late, and some cases may start later than scheduled. The ability to increase utilization, therefore, appears to be limited ultimately by the degree to which the schedule is allowed to extend beyond the end of day and by the delay that patients are asked to assume.

Other factors may also affect utilization. One is case duration. It is easy to see that shorter cases are easier to fit into a schedule than longer cases, because the

Accepted for publication November 19, 2002.

Address correspondence to Donald C. Tyler, MD, MBA, Department of Anesthesiology and Critical Care Medicine, Children's Hospital of Philadelphia, 34th St. and Civic Center Blvd., Philadelphia, PA 19104. Address e-mail to Tyler@email.chop.edu. Reprints will not be available from the authors.

DOI: 10.1213/01.ANE.0000050561.41552.A6

0003-2999/03

remaining time available for scheduling may not accommodate a longer case. If 1 h remains unbooked and a 2-h case needs to be scheduled, either the case must be allowed to run over by 1 h or the 1 h is left empty. Dexter et al. (3,4) have discussed strategies to ensure maximum scheduled utilization.

Variability of case duration also makes it difficult to predict actual utilization. Even for straightforward, common operations, actual case time is uncertain. Each patient is different, and the actual time for a given operation cannot be predicted. This fact means that in a series of cases that are scheduled to follow, the actual start time of cases after the first case cannot be determined in advance.

To examine how factors such as case duration and the variability of case duration affect actual OR utilization, we created a computer simulation of a simple OR. Within this simulation, operational goals can be defined and the effect of case variables examined. To appeal to surgeons and patients, the OR has to run on time, and to keep costs down, it has to be efficient, with maximum utilization and few overtime salary costs. Thus, we defined the goals of this simulated OR to be that cases should start within 15 min of the scheduled start time and that the last case should finish within 15 min of the scheduled end of the day.

Methods

We started with the simplest situation: one OR performing the same procedure repeatedly. In developing the simulation, we made several assumptions. The OR has defined hours of operation, from 7:15 AM until 5:30 PM, and there are no emergencies. Case duration varies from case to case and is described as a probability distribution. The scheduled start time of each case is limited by patient arrival; therefore, cases cannot start earlier than scheduled.

To create the model, we used data from the ORs of the Children's Hospital of Philadelphia. The times for all cases of adenoidectomy with or without tonsillectomy (T&A) for the period January 1 to June 30 1999 were collected. A software program (Arena; Rockwell Automation, Milwaukee WI) was used to find a theoretical distribution that could best fit the collected real data. More specifically, Arena is able to identify a statistical distribution and its variables so that the square error value of the distribution and histogram, which is a measure of the quality of the distribution's match to the data, is minimized. The best fitted distribution is a log-normal distribution with a mean case duration of 48 min and an SD of 15 min. To take into account the size of the SD in relation to the magnitude of the case times, we used the coefficient of variation, which is defined as the SD of the case time divided by the mean.

Using these case times, we created an OR simulation. Cases in the simulated OR were scheduled on the basis of a mean case time of 48 min with a turnover time of 20 min; that is, the next case was scheduled to start 20 min after the previous one finished. Initially we studied days with a utilization of approximately 50% and determined the start time for each case after the first one for the day. We determined "delay" for each case, which was calculated by subtracting the scheduled start time from the actual start time. We also determined "late," which was determined by subtracting 5:30 PM from the end time of the last case of the day if the time was past 5:30. For each situation studied, cases were assigned to the room, and the simulation was run 50,000 times to determine mean start and finish times.

After data were obtained with the baseline utilization, another case was added to the schedule, and the simulation was repeated. Thus, the number of cases performed for the day was increased to achieve increased utilization. Additional cases were added until the scheduled utilization exceeded 100%.

To investigate how the variability of case times affects the ability to achieve maximum utilization, additional simulations were run in the same way, except that smaller and larger coefficients of variation of case duration were used. We also performed an analysis by using the concepts of overutilization and underutilization. To generate data, we used a simulation based on an Excel (Microsoft, Redmond, WA) spreadsheet. The simulation was run, and downtime (unused time during scheduled hours) and late time (time after the end of the day) were calculated. The inefficiency of utilization was calculated by using minutes as the quantitative measure of inefficiency. We assumed that overtime pay is 1.5 times regular pay and that it is 1.5 times as expensive to run the OR with overtime. The inefficiency factor was calculated by adding unused time during the day to 1.5 times the time used after the end of the day. The minutes of inefficiency were plotted against actual utilization for cases with different coefficients of variation.

To examine the effect of our assumption that cases cannot start early because patients have not arrived, we allowed cases to start up to 30 min early and repeated the simulation. To simplify data presentation in this exercise, we determined the number of 48 $\pm$ 15 min cases that could be performed without exceeding the goals of a 15-min delay and 15 min late.

We then examined the effect of turnover time. Simulations were run with turnover times of 0, 10, and 30 min, and the results were compared with those with a turnover time of 20 min. We also simulated a randomly selected turnover time of 20 $\pm$ 10 min.

Our hospital recently opened an ambulatory surgery center (ASU), staffed only by attending physicians. We created a simulation that demonstrated the

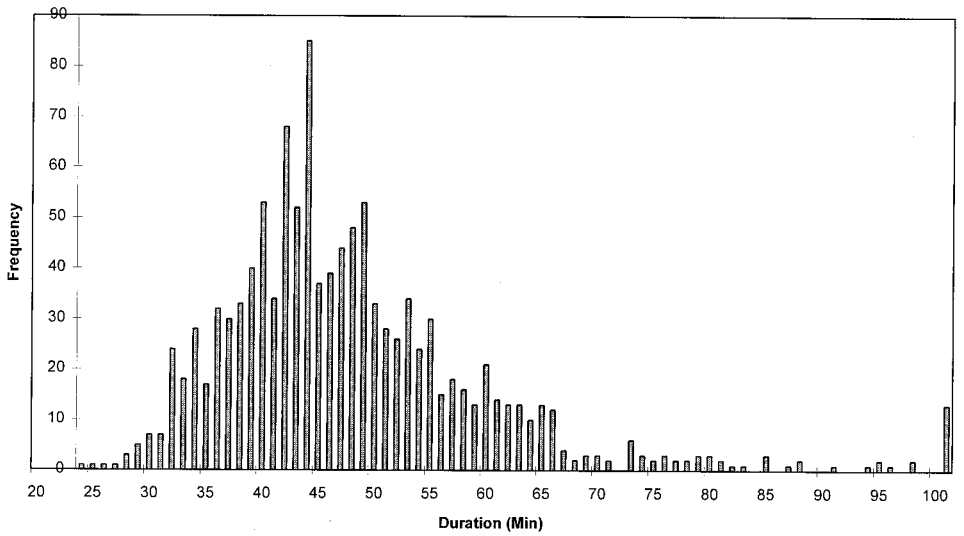

Figure 1. Distribution of case times for adenoidectomy and adenoidectomy with tonsillectomy.

effect of moving to a different venue healthy patients with procedures performed by experienced practitioners. With these patients moved to the ASU, the main OR was left with sicker patients whose procedures were performed by the same practitioners but with the addition of residents and fellows. For all simulations, we created an OR that performed five cases of myringotomy and tubes (BMT) and five cases of T&A, with a scheduled utilization of approximately 90%. For the baseline situation, the mean and SD of the case times of the total group of cases were used. We then used the mean and SD of case times from the ASU and from the main OR without the ASU patients for the additional simulations. Scheduling was based on mean case times for the procedures lumped together. In each simulation, we measured delay, late time, and over- and underutilization.

To evaluate a more complex situation, we simulated an OR performing four cases of varying length, namely, 1, 2, 3, and 3 h. Turnover time was 20 min. For each case, we used a coefficient of variation equal to that of our original data, namely, 0.31, thus resulting in a moderate degree of variability of case times. The simulation was run with a scheduled utilization of 100%, performing the shorter cases first. To vary utilization, we eliminated first the 1-h case, then the 2-h case, for scheduled utilizations of approximately 87% and 77%. Simulations were run in each of these levels of utilization. We also performed the simulation for all four cases, performing the longer cases first.

Results

There were 1162 cases with a duration of 48 ± 15 (SD) min. The coefficient of variation was 0.31. The best-fitted distribution was a log-normal distribution (Fig. 1). When utilization increased more than approximately 90%–95%, both late minutes and delay increased beyond the target of 15 min (Fig. 2).

Figure 3 shows the effect of changing the coefficient of variation, that is, increasing or decreasing the variability of case duration, on late minutes and on average delay. Decreasing the coefficient of variation (smaller SD of case duration) allows increased utilization without exceeding the target for late minutes (Fig. 3A). Increasing the variability of case duration, however, decreases the utilization possible within the target for late minutes. A smaller variability of duration allows increased utilization, whereas increased variability allows less utilization within the target for patient delay (Fig. 3B). Efficiency increases as the coefficient of variation decreases, and overall the OR is most efficient with a utilization between 85% and 95% (Fig. 4).

When cases were allowed to start up to 30 min before their scheduled start time, the effect was to increase the utilization possible while staying within the guidelines for late times and delay of cases (data not shown). The change was enough that nine cases instead of eight could be performed in a day without violating the targets for late and delay minutes.

As turnover time changes, there is a difference in the number of cases that can be performed within the guidelines for late and delay minutes. With a 10-min turnover, nine cases could be performed; with a 20-min turnover, eight cases could be performed; and with a 30-min turnover, only seven cases could be performed (Table 1). When turnover was randomly chosen from a distribution of 20 ± 10 min, seven cases could be performed; this was limited by the degree to which patient start time was delayed.

For the simulation of the effects of opening the ASUs, the case time for BMT for the combined group

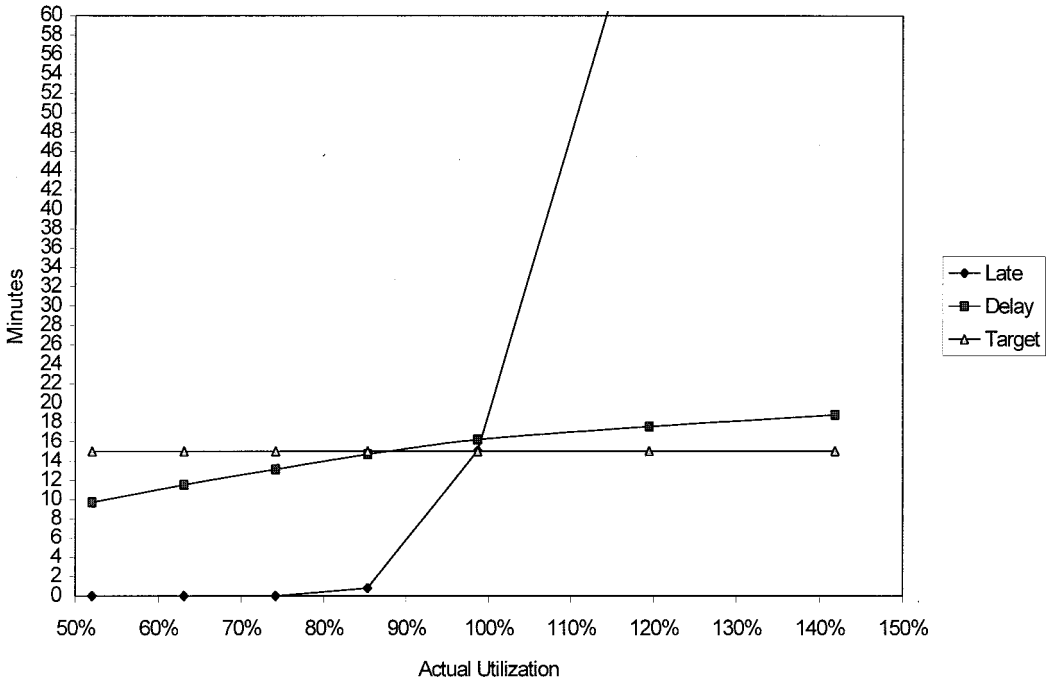

Figure 2. Late and delay times with increasing utilization.

was 15.4 ± 11.6 min, and for T&A it was 43.8 ± 12.3 min. After the ASUs opened, BMT time was 19.0 ± 15.9 min and 12.7 ± 5.6 min at the main hospital and the ASU, respectively, whereas the times for T&A were 48.3 ± 14.6 min and 40.7 ± 11.1 min, respectively. Before the ASU patients were removed from the schedule, the main hospital was able to perform five BMTs and five T&As with a utilization of 89% within the guidelines for late and delay minutes (Table 2). When the ASU patients were removed from the schedule, the main hospital was not able to perform the same number of cases in a day, whereas the ASU was able to finish within the guidelines.

When longer cases were mixed with shorter ones, the maximum utilization that could be achieved within the guidelines was approximately 85% (Table 3). Performing the shortest cases first resulted in less patient delay compared with performing the longer ones first.

Discussion

The use of simulation allows for the evaluation of various factors that affect OR utilization without the necessity of gathering data over a long period of time. In addition, simulation allows for the manipulation of one factor at a time in a way that would not be possible in a real-life situation. We started with the simplest possible situation—one procedure being performed repeatedly in one OR, with a constant, short mean case duration. Although the real-life situation is much more complicated, some conclusions can be drawn to help define the optimum utilization of an OR.

To approach an answer to the question of what is the optimum utilization of an OR, we need to define certain operational variables. In this model, we defined goals for average patient delay and an acceptable time past the end of the day. To avoid having patients arrive early and wait a long time, we defined an average patient wait of 15 minutes as the goal. We also wanted to avoid using overtime, so we defined the goal of finishing the schedule within 15 minutes of the scheduled end of the day. With these criteria, in the simple situation described, optimum utilization was approximately 90%. Increased utilization resulted in increased patient delay and overtime beyond our defined limits.

If case times could be predicted accurately, it would be relatively easy to schedule an OR, but the actual case duration is not known in advance. The variability of case times, as measured by the coefficient of variation, is thus a major factor to be considered if one wishes to achieve maximum utilization. Decreasing the variability of case times allows increased utilization to be achieved. As seen in our model, when the variability increases, the predictable optimum utilization decreases. As the variability of case times increases, it is more difficult to schedule times accurately and more difficult to achieve maximum utilization without making patients wait or having the schedule extend past the end of the day.

Because the simulation selects the case duration from a probability distribution for each run, the nature

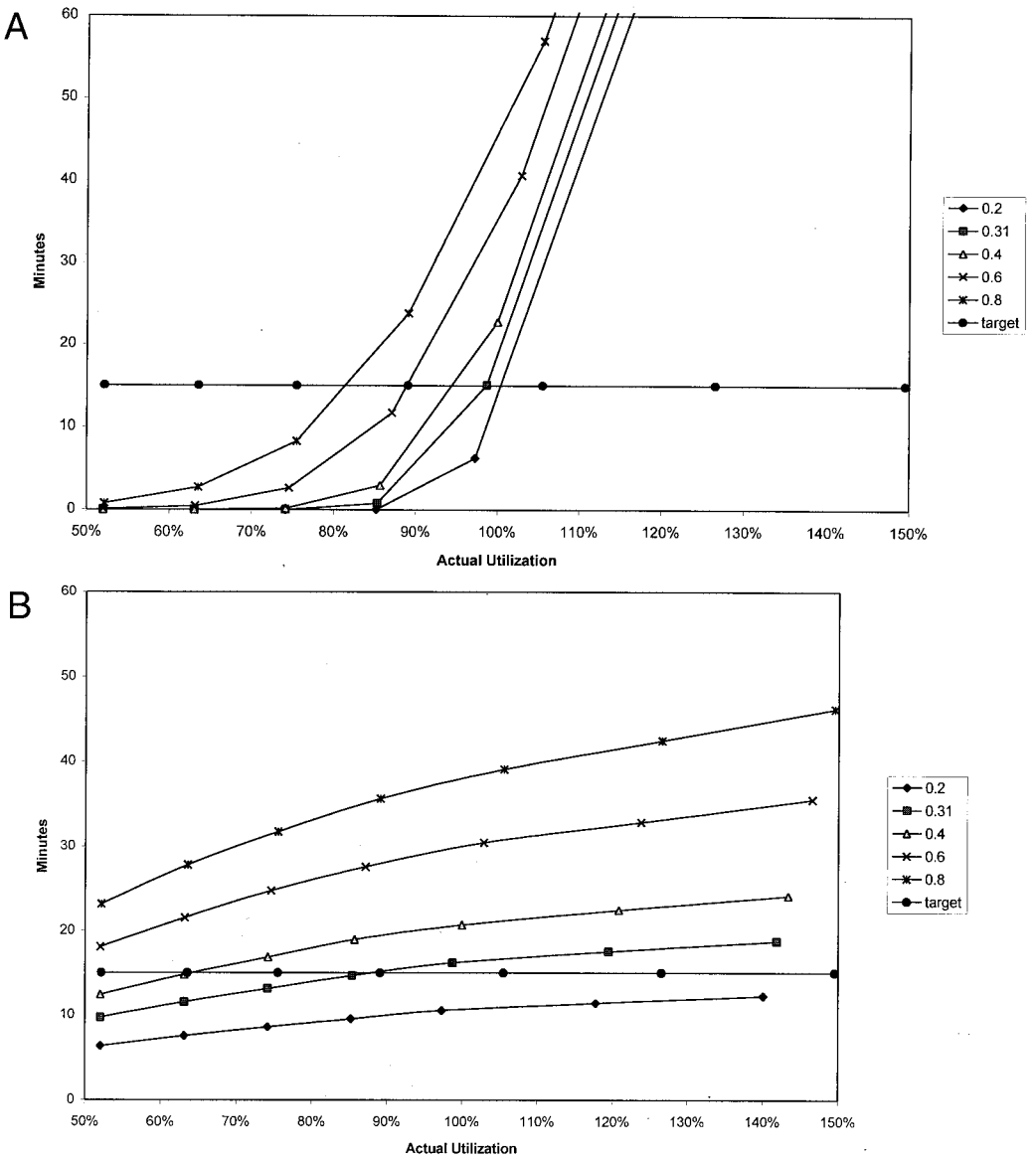

Figure 3. A, Late minutes with different coefficients of variation of case duration. B, Average minutes of delay for individual patients with changing coefficients of variation of case duration.

of the distribution used is important. We used actual times from a large number of cases to determine the distribution. As in previous studies (5,6), the distribution was log normal. This sort of distribution does not have negative numbers and is skewed toward larger values, a situation that makes sense for an OR.

In most ORs, cases are scheduled on the basis of mean case times. One might expect the shorter and longer case times to average out, but this is not the case. Because patients are assigned an arrival time and are not available before their scheduled start time, if a case finishes early, the subsequent patient may not be available, and the case cannot start.

We also examined optimum utilization from an economic perspective by evaluating the relative efficiency of the OR in terms of staffing costs. For this we used the concept of overutilization and underutilization described by Strum et al. (1). We concluded that optimum utilization is between 85% and 90%, depending to a large extent on the variability of case duration. The classic definition of utilization considers the time past the end of the day the same as time during scheduled hours. This is undesirable because there is a cost to going past the end of the day. Overtime costs may be incurred, and repeatedly running late is hard on staff morale and may make recruiting and retaining scarce staff more difficult. Using the concept of overutilization and underutilization allows us to put a cost to running late and to quantitate the quality of OR scheduling.

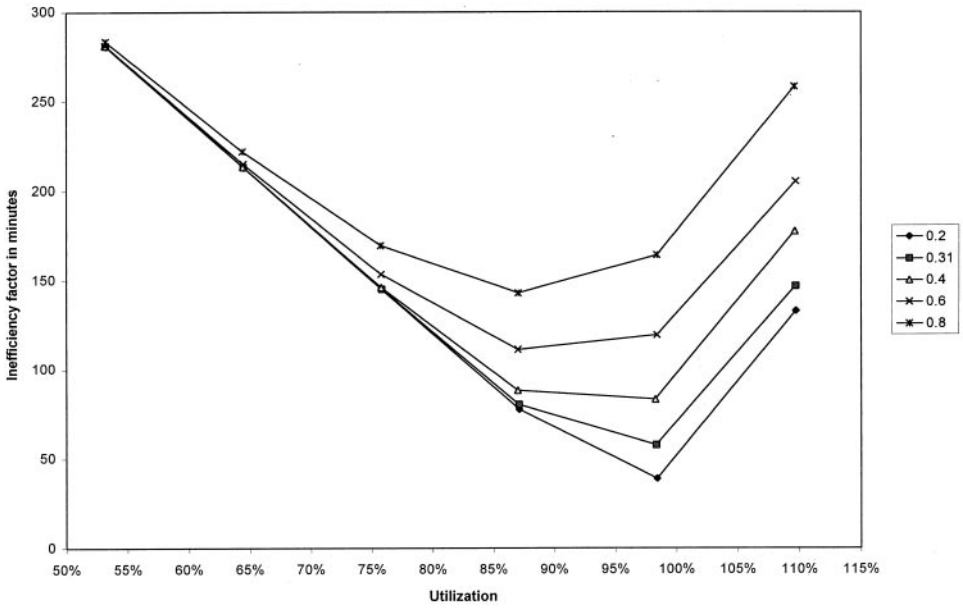

Figure 4. Efficiency of operating room utilization with changing coefficients of variation.

Table 1. Effect of Varying Turnover Time on Late Minutes, Delay Minutes, and Inefficiency

Cases	Turnover time (min)	Actual utilization	Late (min)	Delay (min)	Inefficiency factor (min)
6	10	56%	0.0	11.1	262
7	10	66%	0.0	12.5	204
8	10	76%	0.0	13.8	146
9	10	85%	0.9	15.1	90
10	10	95%	11.6	18.9	74
6	20	64%	0.0	11.1	213
7	20	76%	0.0	12.5	145
8	20	87%	1.2	13.8	80
9	20	98%	20.5	15.0	61
10	20	110%	85.4	19.0	155
6	30	73%	0.0	11.0	162
7	30	86%	0.6	12.5	85
8	30	99%	20.8	13.8	58
9	30	112%	97.3	15.0	171
10	30	125%	177.2	19.1	293
6	20 ± 10	65%	0.0	13.3	213
7	20 ± 10	76%	0.1	15.3	144
8	20 ± 10	87%	2.4	16.8	82
9	20 ± 10	99%	25.9	18.4	73
10	20 ± 10	110%	91.8	23.3	170

Case length was 48 ± 15 min. Turnover time was as indicated except in the last group, where turnover time was a random number selected from a log-normal distribution: 20 ± 10 min.

We were able to increase the number of short cases that could be performed within the guidelines by allowing cases to start up to 30 minutes before the scheduled time. We would argue, however, that to ensure that patients are available up to 30 minutes early, they would have to arrive at the facility sooner than they otherwise would have. This means that to increase the efficiency of the ORs, we must increase patient waiting time.

Because turnover time is included in the calculation of utilization, there was no difference in the utilization that could be achieved within the guidelines when turnover time was lengthened or shortened. When one looks at the number of cases that could be performed within the guidelines, however, it becomes clear that

Table 2. Late and Delay Minutes for Cases Performed at the Main Operating Room (OR) and at an Ambulatory Surgery Unit (ASU)

Facility	Scheduled utilization	Actual utilization	Late (min)	Delay (min)
Combined	89%	89%	2.5	12.2
Main OR	89%	97%	15.3	24.7
ASU	89%	83%	0.1	3.6

Table 3. Data When Cases of One, Two, and Three Hours Are Performed in the Same Room

Scenario	Scheduled utilization	Actual utilization	Late (min)	Delay (min)	Inefficiency factor (min)
Four cases, short ones first	100%	100%	50.0	16.1	125.2
Three cases, omit 1-h case	87%	87%	14.8	16.3	117.4
Three cases, omit 2-h case	77%	77%	4.4	11.8	143.8
Four cases, long ones first	100%	100%	50.8	27.8	127.9

in this situation of many short cases, turnover time is important in the overall efficiency of the OR. When turnover time is shortened, more cases can be performed within the guidelines for late and delay minutes. A variable turnover time appears to reduce the number of cases that can be performed, perhaps by creating significant delays for individual patients.

For the simulation in which cases were moved to the ASUs, some scheduling questions appeared. When one examines the data, it is clear that after the cases were separated, the cases in the ASU had shorter mean times and that those left at the main hospital had longer times. We ignored those time differences and used the mean case times obtained when the cases were lumped together. Scheduling needed to be done before the data were available; consequently, the best available data were the combined data. Second, the time differences were small and were below the level of precision of most OR scheduling programs, which function with either 10- or 15-minute increments. With longer case times and larger variability of case times, the main OR could not schedule to the same predicted utilization without having excessive late and delay time after the ASU patients were removed. This finding confirms the feelings of the staff that taking simple cases out of an OR schedule increases the inefficiency of the OR suite.

When longer and shorter cases were combined in one room, we found decreases in the utilization that could be achieved within the guidelines. We chose case lengths that would fit within the allotted time, so the 87% utilization that was achieved is perhaps the best possible situation. If predicted case lengths were such that the entire period could not be used, then utilization would be expected to decrease.

The data that we have obtained demonstrate some of the limitations of "classic utilization." Although utilization calculated in this way is a simple calculation that gives some measure of the extent to which an OR is used, it does not give an adequate measure of how efficiently resources, especially staff time, are used. Because staffing costs are the largest component of OR expenses, efficient use of OR time requires that the ORs be used fully when staffed and used as little as possible when staffing relies on overtime or call staff. Classic utilization gives no measure of how well activity is matched to staffing. The concepts of overutilization and underutilization produce a quantifiable measure of how well staffing and use are matched and allow managers to optimize staffing.

One of the clear messages of the attempt to optimize utilization is that a balance needs to be achieved between efficiency from an economic standpoint, and patient satisfaction. It is clear from our data that efficiency of the OR is achieved by increasing patient waiting time. The most efficient OR is one in which patients are waiting when the OR is available. If a case finishes before predicted and the next patient is not available, then the OR sits idle (underutilization) until the patient arrives and is prepared for surgery. However, the most patient-friendly OR is one that has a room ready when the patient arrives. Achieving optimum utilization results when these two factors are balanced.

Our model creates a situation that is unlikely to be achieved in the real world. The cases are short, with a small coefficient of variation. In most of the situations studied, cases are of similar length. There are no patient-related delays and no case cancellations. Thus, the optimum utilization defined in this simulation is more than can reasonably be expected in a real OR. In real life, if these factors are taken into account, a smaller utilization may actually be achieved.

With the assumptions that we have used, actual utilization will almost always be more than scheduled utilization. This observation means that it is difficult to target a specific utilization for a given day and there is variability of the actual utilization achieved for any given scheduled utilization. For the simple situation studied here, a target utilization of 85% would approximate what we would like to achieve in terms of patient delay and overtime. This is perhaps the maximum utilization that can be achieved within the goals that we have set. For more complex OR suites, the optimum utilization will be less. Any change, such as cases of different duration, changes in the variability of case duration, emergencies, cancellations, and so on, will decrease the optimum utilization. The decision about making additional ORs available on the basis of maximum utilization will depend on specific factors in a given suite—factors such as the cost to the organization of asking patients to wait, the cost of

overtime, and the ability of surgeons to take cases to another suite.

References

1. Strum DP, Vargas LG, May JH. Surgical subspecialty block utilization and capacity planning. Anesthesiology 1999;90:1176–85.
2. Donham RT, Mazzei WJ, Jones RL. Glossary of times used for scheduling and monitoring of diagnostic and therapeutic procedures. Am J Anesth 1996;23:1–12.
3. Dexter F, Macario A, Traub RD, et al. An operating room scheduling strategy to maximize the use of operating room block time: computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time. Anesth Analg 1999; 89:7–20.
4. Dexter F, Macario A, Lubarsky DA, et al. Statistical method to evaluate management strategies to decrease variability in operating room utilization. Anesthesiology 1999;91:262–74.
5. Zhou J, Dexter F. Method to assist in the scheduling of add-on surgical cases—upper prediction bounds for surgical case durations based on the log-normal distribution. Anesthesiology 1998; 89:1228–32.
6. Strum DP, May JH, Vargas LG. Modeling the uncertainty of surgical procedure times. Anesthesiology 2000;92:1160–7.