Giles Monnickendam, Carlo de Asmundis

Why the distribution matters: using discrete event simulation to demonstrate the impact of the distribution of procedure times on hospital operating room utilisation and average procedure cost

Article (Accepted version) (Refereed)

Original citation: Monnickendam, Giles and de Asmundis, Carlo (2018) *Why the distribution matters: using discrete event simulation to demonstrate the impact of the distribution of procedure times on hospital operating room utilisation and average procedure cost.* Operations Research for Health Care, 16. pp. 20-28. ISSN 2211-6923
DOI: 10.1016/j.orhc.2017.12.001

Accepted Manuscript

Why the distribution matters: Using discrete event simulation to demonstrate the impact of the distribution of procedure times on hospital operating room utilisation and average procedure cost

Giles Monnickendam, Carlo de Asmundis

PII: S2211-6923(17)30045-0
DOI: https://doi.org/10.1016/j.orhc.2017.12.001
Reference: ORHC 145

To appear in: *Operations Research for Health Care*

Received date : 24 April 2017
Accepted date : 13 December 2017

Please cite this article as: G. Monnickendam, C. de Asmundis, Why the distribution matters: Using discrete event simulation to demonstrate the impact of the distribution of procedure times on hospital operating room utilisation and average procedure cost, *Operations Research for Health Care* (2017), https://doi.org/10.1016/j.orhc.2017.12.001

Why the distribution matters: using discrete event simulation to demonstrate the impact of the distribution of procedure times on hospital operating room utilisation and average procedure cost.

Giles Monnickendam[1], BSc, MSc, Carlo de Asmundis[2], MD, PhD

[1] London School of Economics and Political Science, London, UK

[2] Universitair Ziekenhuis Brussels, Vrije Universiteit Brussels, Brussels, Belgium

Corresponding author:
Giles Monnickendam
24 Ancastle Green
Henley on Thames
RG9 1TR
United Kingdom
giles.monnickendam@gmail.com
+44 7484 700 275
ORCID iD: 0000-0001-6630-8970

Funding:
No funding was received to support this manuscript.

Conflicts of Interest:
Carlo De Asmundis receives financial compensation from Medtronic for teaching and proctoring in the field of cardiac electrophysiology.

Key Words:
economic evaluation, capacity utilisation, discrete event simulation, operating room scheduling, micro-costing, atrial fibrillation.

Running Title:
Why the distribution matters

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AD	Anatomically designed
AF	Atrial fibrillation
AIC	Akaike information criteria
AOT	Actual occupancy time
CEA	Cost effectiveness analysis
DES	Discrete event simulation
DRG	Diagnostic related group
EOT	Effective occupancy time
HTA	Health technology assessment
OR	Operating room
PBP	Point-by-point

HIGHLIGHTS

- Surgical procedures that are long and unpredictable in duration are more difficult to schedule efficiently and reliably, compared with shorter and more predictable procedures.
- The impact of variability of procedure duration on OR capacity utilisation is well understood, but the implications for estimates of average costs of health treatments in economic evaluations have not received any attention.
- A DES model, calibrated with real-world data, is used to compare simulated resource consumption for two alternative procedures with significantly different distributions of duration.
- We demonstrate that methods of estimating resource consumption that do not consider the full distribution of duration are likely to produce relative cost estimates biased against technologies with shorter and more predictable procedure times.
- Comparing the full distributions of procedure duration and understanding the implications for OR capacity utilisation will lead to estimates of average costs more representative of the true impact of implementing health technologies.

ABSTRACT

Background: Economic evaluations include estimates of surgical procedures costs that have usually been derived by allocating operating room (OR) costs in proportion to the average duration of different procedure types. However, ORs run with average utilisation below 100%, due to idle time between procedures and at the end of the day. Longer and less predictable procedures generate greater OR idle time, for a given tolerance of schedule over-runs. Estimates of surgical procedure costs that are based on average procedure duration alone as a measure of OR resource consumption will not capture the impact of the length and variability of procedure duration on OR idle time and capacity utilisation.

Objective: To demonstrate how real-world OR scheduling practices lead to different levels of resource consumption than predicted by simple micro-costing approaches based on average procedure duration, and how those differences can vary between procedures with significantly different distributions of duration.

Methods: We use a discrete event simulation model, calibrated with real-world data from a single surgical centre in Belgium, to compare simulated resource consumption, including idle time, for two alternative surgical procedures for ablation for atrial fibrillation.

Results: We demonstrate that simple micro-costing approaches can under-estimate effective resource consumption between 31% and 48% for a procedure with long and unpredictable duration. For a shorter and more predictable procedure the under-estimate is only 15%.

Conclusion: Simple approaches to estimating procedure costs may under-estimate resource consumption and do so in a way that is biased against technologies with shorter and more predictable procedure duration. For health technology decisions where a substantial part of costs are OR resources, a more sophisticated approach, taking account of the real-world implications of the distribution of procedure durations, should be used to avoid potential bias.

1. INTRODUCTION

1.1. Overview

The costs of surgical procedure can be included within economic evaluations in health technology appraisals (HTAs) using either gross-costing or micro-costing approaches. Gross-costing methods are more frequent and typically use diagnostic related group (DRG) tariffs as an estimate of overall costs of spells of inpatient care. Micro-costing methods involve estimating the costs of subordinate components of treatment and are necessary for new therapies without an existing DRG or comparisons of procedures within the same DRG. Micro-costing approaches usually estimate average procedure cost by taking average procedure duration from the literature or calculating from direct measurement at a patient level, and multiplying by unit costs for operating room (OR) resources (staff and facility). Implicitly, costs are assumed to be linear and continuous with procedure duration. However, longer average procedure times make resource planning more problematic and disproportionally increase resource consumption. Variability in procedure times also imposes costs, with increased variability leading to less efficient resource planning and greater absorption of fixed capacity, all other things remaining equal. Consequently, actual resource use can differ significantly from results predicted in these simple costing models. In particular, a procedure with double the average duration of another may consume more than double the resources. And two procedures with identical average duration but significantly different variability in procedure time will absorb different amounts of fixed resource capacity [1]. In this study we focus on this issue within economic evaluations using micro-costing approaches, where it c an most clearly be identified and demonstrated. However, to the extent that the procedure cost components within DRG tariffs are themselves calculated using micro-costing methods with similar approaches to cost allocation, HTAs using gross-costing methods will also be affected.

To demonstrate the real-world complexities of OR planning and the potential impact on average procedure costs, we use a discrete event simulation (DES) model, calibrated with real-world data from a single centre in Belgium, to compare two alternative procedures for ablation for atrial fibrillation (AF). AF is the most common form of cardiac arrhythmia, it is a major cause of stroke and is associated with increased mortality and reduced quality of life. Catheter-delivered ablation of pulmonary veins within the heart is a minimally invasive procedure indicated for patients with symptomatic AF refractory or intolerant to at least one antiarrhythmic drug. Procedures are usually conducted in a specialised OR (catheter or electrophysiology lab). "Point-by-point" (PBP) technologies use repeated spot application of energy in a circular pattern around the pulmonary vein. "Anatomically-designed" (AD) technologies use a pre-shaped tool to create circular scarring by a single application of energy. The different surgical techniques generate similar clinical outcomes, but real-world data shows different distributions of procedure time (Figure 1). PBP procedure times are longer on average and also more variable. We use the DES model to show how effective resource utilisation for each type of procedure can differ substantially from simple estimates based on measures of average procedure duration.

1.2. The procedure scheduling problem

The cost of running surgical operating facilities are substantial for most hospitals and optimising the use of this fixed resource is a priority to achieve efficiency in provision of health services [2]. Utilisation is defined as the total amount of time patients are in the OR, plus the minimum turnover time required between procedures, divided by the scheduled period of time that the OR is available and staffed. Hospitals face the challenge of how to plan resources and schedule elective procedures to maximise utilisation, while also ensuring procedures run to time within given tolerances. Schedules with considerable safety margin included to cope with variability of procedure duration give a high probability that all procedures will be completed on time, but lead to significant periods when the OR is left empty (under-utilisation). Schedules with no safety margin eliminate idle time, but lead to frequent over-runs, incurring staff over-time costs and negative impacts on the efficient co-ordination with connected processes within the hospital (over-utilisation). There is an extensive literature on planning and scheduling of hospital ORs, documented in a recent literature review by Cardoen et al [3]. More specifically, Tyler et al demonstrate how procedures that are both long and unpredictable in duration result in lower maximum OR utilisation rates than shorter and more predictable procedures [1]. Unpredictable procedure times are disadvantageous since, if OR schedules are to be met within a given tolerance, a greater safety margin needs to be built in. Longer average procedure times are disadvantageous due to the "bin packing" problem. With a fixed period of availability of the OR, it is easier to pack in many short operations, than a smaller number of long operations.

1.3. Real-world OR scheduling practices

ORs dedicated to elective surgery can either operate with a mix of cases/surgical teams scheduled to use the OR through the day, known as open scheduling, or use block scheduling to allocate a large portion of OR time to a single specialty or surgical team to complete a number of cases in series.

With open scheduling, surgeons are assigned a fixed slot for a single procedure within the OR schedule and are expected to run to time. Since surgical teams are given an arrival time at the OR and are not "queued", a procedure finishing earlier than scheduled does not produce a time gain in the OR schedule, but procedures that over-run delay the start of the following procedures and lead to over-time costs at the end of the OR day. This asymmetry means that the resource impact of under and over-runs do not cancel out.

Block scheduling typically allocates OR time on a 4 or 8-hour basis, in line with a half or full staff shift. The specialty or surgeon allocated the block has some discretion over how to plan/use the OR time, but is responsible for ensuring that procedures do not exceed the end time of the block. Frequent over-runs can incur over-time costs, cause morale problems with staff and create friction with other surgical teams scheduled later in the OR. Cancelations of scheduled cases can be used to prevent excessive over-runs, but also impose significant costs.

Where procedure volumes allow, block scheduling is usually favoured, since set-up and idle time are reduced. Economies of scale are generated, since the OR only needs to be configured once for the surgical team at the beginning of the block, rather than before every procedure. The queuing of cases within the block also allows idle time to be saved when procedures finish early. Finally, block scheduling improves incentives for individual surgeons to manage OR time effectively, by internalising some of the impact of poor time management and eliminating competition/gaming between surgeons to secure earlier or longer slots in the OR schedule.

When scheduling procedures, either using open or block scheduling, hospital managers seek to maximise the utilisation of OR facilities, subject to a constraint of a maximum tolerance of procedure over-runs. This results in idle time built into the procedure schedule, as managers build in buffers to accommodate potential over-runs. This idle time is significant - for ORs dedicated to elective surgery and using scheduling best-practices, utilisation rates in excess of 85% would be unusual [3].

1.4. Micro-costing and capacity utilisation in economic evaluation

The negative impact of variability of procedure duration on OR capacity utilisation has been covered extensively within the operational research literature, but the potential impact on average costs of health treatments has not received any attention within health economics. Raikou and McGuire note that much less attention has been given by health economists to the measurement of costs compared to health outcomes. Information is limited and well-established methodologies are lacking in particular for the measurement of direct costs [4]. Published guidance recommends that quantities of resources used and their unit costs should be identified separately to produce cost estimates [5, 6, 7, 8, 9]. However, Adams et al identify lack of detail within guidelines on how to comply with this recommendation in practice as a key source of variability in costing methodologies employed in cost effectiveness analysis (CEA). Capacity utilisation is singled out as an area of particular concern. Guidelines do not provide clear, practical recommendations and applied studies use a wide variety of approaches and rarely report assumptions made for capacity utilisation [10].

Tan et al categorise methodologies for estimating costs of hospital services by the level of detail at which resource components are identified (gross-costing versus micro-costing) and their valuation is conducted (top-down versus bottom-up) [11]. Bottom-up micro-costing provides the most detailed estimates of costs for hospital services, but is time-consuming and challenging to implement. Hence it is usually considered appropriate only for cost components which make up a large proportion of costs under consideration and with potential to differ significantly across the alternatives under evaluation [12].

Gross-costing approaches are most commonly used in therapy-level CEA. Costs are included at a highly aggregated level, often using DRG tariffs to estimate total inpatient costs associated with a procedure. Differences in calculation methodology for DRG tarrifs have been observed across countries [13] and significant inconsistencies documented in implementation [14] within countries. In particular, variability in calculated DRG tarrifs can be driven by differences in the assumptions employed for the allocation of overheads costs across activities. Consequently, generalising broadly for DRG tariffs requires caution, but to the extent that DRG costs capture OR idle time, they are likely to do so by allocating a total across many different procedure types, in proportion to average procedure duration. This approach will not differentiate between procedures that generate proportionally very little OR idle time (due to being shorter and more predictable and hence easier to schedule) and those which lead to much greater idle time.

Micro-costing approaches are more common in hospital costing studies, where there is a need to distinguish costs of procedures within a DRG or evaluate alternatives where the differences in impact on resource use will not be picked up at a procedure code level [8, 15]. In bottom-up micro-costing approaches for surgical procedures, patient-level procedure duration is used to determine OR resource use for cost estimates and OR idle time is not allocated [16, 17], hence no account is taken of the increased resource usage that results with longer and more variable procedures. Top-down micro-costing approaches may be used in therapy-level CEAs or comparisons of technologies within a therapy. Some top-down micro-costing approaches may implicitly include idle time, through assumptions on OR capacity utilisation [18], but as with DRGs, this effectively allocates based on average procedure duration and so does not differentiate between procedure types that generate different amounts of idle time.

1.5 Resource-use and CEA studies for AF ablation technologies

A large number of studies have examined procedure duration of PBP and AD technologies. Most comparisons show reduced procedure duration for AD technologies versus PBP [19]. Standard deviations for procedure duration are reported within all studies, but only to determine whether differences in duration are statistically significant. It is rare that any comment is made on the variability of procedure duration as an important characteristic of the technology on its own. Most studies report lower standard deviations for AD procedure times but make no mention of the potential implications for resource utilisation. Only two papers specifically discuss issues of the distribution and predictability of procedure times [20, 21] and do so very briefly.

Within the CEA literature for AF ablation technologies, a range of approaches are used to estimate costs of procedures. The most common are top-down gross-costing [22, 23, 24], using DRG or reimbursement tariffs for total inpatient costs, and top-down micro-costing (25, 26, 27), where procedure costs are included separately but calculated based on high level assumptions, including procedure duration. Often the precise costing methodology is unclear and no paper gives any indication of assumptions made on capacity utilisation or inclusion of idle time.

2. METHODS

2.1. Study setting

We use procedure duration data and the OR scheduling processes from a single, high-volume centre in Belgium, conducting ablation procedures for AF, as the basis for a DES model. The centre is a research hospital with a long experience with both PBP and AD technologies. Over a 7-year period, the two technologies were used "side-by-side". Over time, surgical staff found that routine scheduling of AD procedures was less problematic and they were able to adjust scheduling to achieve higher capacity utilisation and more optimal use of hospital resources than with PBP.

2.2. The DES model

A structured interview was conducted with centre staff to understand the process for admitting and processing patients for AF ablation, and the normal mode for scheduling and completing the elective procedures. Based on this a DES model was built to simulate and compare the scheduling of surgical cases during a day of "PBP only" and "AD only" elective procedures. The model was built within Microsoft Excel, with @Risk as an add-in application to run monte-carlo simulations. The key elements of the model are the estimated distributions of duration for each procedure type and the set of parameters and decision rules that replicate the centre's initial OR scheduling and response to the cumulative variability in procedure times over the course of the day (Table 1).

The outcomes of interest from the DES model are actual OR occupancy time (AOT) and effective occupancy time (EOT). The OR is defined as occupied by a case when either a patient is undergoing a procedure or staff are completing turnover of the OR at the end of the procedure and the total time for this is AOT. Additionally, EOT includes the OR idle time between procedures or at the end of the day, when the OR is available but unoccupied. EOT also includes over-run time weighted at over-time rates. Analysis focuses on comparisons of EOT and AOT, to show the potential difference between simple micro-costing calculations and effective resource consumption that takes account of idle time. The ratio of average EOT to AOT indicates the extent to which simple cost estimates based on mean procedure duration will under-estimate real-world resource consumption.

Three alternative modes of operation were simulated: fully-flexible scheduling; open scheduling; and block scheduling with cancelations (Figure 2). Fully-flexible scheduling is a theoretical scenario where surgical cases are queued and processed with an open-ended and fully-flexible OR schedule. There is no idle time between cases, due to queuing, and none at the end of the day, since the schedule is open-ended. The OR continues until all cases are finished and there are no overtime costs. Nor are there any other costs imposed on connected processes within the hospital from procedures running under or over time. This scenario produces results for resource utilisation that are equivalent to micro-costing approaches that allocate OR cost based on average procedure duration. Costs are entirely proportional to procedure duration, idle time is not included and there is no additional cost imposed by increased variability.

Open scheduling is not used within the study centre, but is simulated within the model using generic assumptions to show the consequences for resource utilisation. For block scheduling scenarios, the actual block scheduling process of the study centre was simulated. The centre operates with an 8-hour block and over-runs are not permitted. 2 PBP or 4 AD elective cases are admitted the day before for pre-op diagnostic tests and queued the morning they are due to take place. A schedule review is conducted part way through the day and if procedures are running significantly behind time, the last case is cancelled. Within the model, the schedule for each procedure type is then optimised (by reducing block time or adding additional procedures, subject to a criteria of a maximum tolerance of days with over-runs, set at 10%) to demonstrate the differences created in resource utilisation for the two procedures when the most efficient configuration is used for block scheduling.

2.3. Procedure duration data

To simulate the impact of variability on procedure scheduling, the model requires a distribution of procedure times to be defined. Patient-level data collected by the centre was used to estimate defined distributions of duration separately for AD and PBP procedures. Procedure duration data were fitted to parametric distributions using the distribution fitting function within @Risk, which uses maximum likelihood estimation to fit a range of distributions. Log-logistic distributions were selected for both PBP and AD procedure duration, with Akaike Information Criteria (AIC) used to determine best fit.

The procedure duration data was sourced from an established registry, covering a period from January 2008 to November 2014 and includes observations for 595 PBP and 570 AD procedures. This dataset has been used and described in a number of published studies [28, 29, 30, 31]. The selection of patients to PBP or AD was not randomised. Patient characteristics were in general well matched across the two treatment groups (see Table A2 within the appendices), but some differences exist which could act as confounding factors for the differences we observe in distributions of procedure duration. Patients with persistent rather than paroxysmal AF were under-represented in the AD group; a greater percentage of repeat ablation procedures were observed in the PBP group; a higher rate of hypertension comorbidity was observed in the PBP group; and the number of PBP and AD procedures were not equally distributed over the 7-year period of the study. PBP and AD procedures were conducted "side by side", but not with equal or even distributions over the years of the study. As the AD technology became a preferred technology for the centre (due to broadening of the indication and scheduling advantages), an increasing number of AD procedures were completed in later years. Procedures were also conducted by a number of different operators over the period, there was some incremental change in the technologies, and methods used evolved over the time period of the study.

To address potential confounding factors and to take account of any potential shifts in the distribution of procedure duration over the period of the study, multivariate regression techniques were employed to generate adjusted distributions of procedure times for AF and PBP procedures. STATA was used to perform parametric survival-time regression (specifying log-logistic distributions) of duration data against potential confounders and a study-year variable. The regression coefficients were then used to generate distributions for a standardised population, estimating the performance of PBP and AD in the final year of the study. Full detail of the analysis and regression methodology is included in sections A4 and A5 of the appendices.

Figure 3 provides a graphic comparison of the results of the regression adjusted approach to unadjusted fitted distributions for the full dataset and a restricted dataset (including only cases of paroxysmal AF). In Table 2, key statistics for each fitted distribution are compared back to the appropriate actual data and with each other. The regression adjusted approach was chosen as the most robust (using the full dataset), conservative (the smallest difference in AD and PBP duration) and most relevant (estimating performance in the final year of the study) approach to derive distributions for the DES modelling. Results for the DES model for the two other approaches are included within the sensitivity analysis, within section A7 of the appendices.

3. RESULTS

3.1. Procedure duration

Figure 1 shows the different distributions of procedure times for the two technologies, along with key statistics. PBP procedure times are longer on average, with a median duration of 140 minutes, compared with 60 minutes for AD procedures. PBP procedure times are also more variable. The difference in variability is moderate in relative terms, as measured by the coefficient of variation, but becoming much more significant in absolute terms, as measured by standard deviation (minutes), due to the longer duration of PBP. However, also notable are the differences in the shapes of the two distributions, with the PBP distribution demonstrating an extended tail of long procedure times. For PBP, 9.1% of procedures exceed 1.5 times the median duration, while for AD the equivalent figure is only 4.4%. In absolute terms, 29.4% of PBP procedures are more than 30 minutes longer than the median duration, while for AD procedures only 10.4% of procedures exceed the median plus 30 minutes.

3.2. Simulation modelling

Results from the DES model are summarised in Table 3 and are all based on the regression-adjusted distributions of procedure times.

"Fully flexible" represents the baseline for comparison for other scenarios and produces AOT and EOT that are identical for each procedure type. Mean procedure time for PBP is 163 minutes, while for AD it is 83 minutes.

"Open scheduling" is representative of an OR where surgical teams are assigned a single slot in the schedule to complete a procedure, rather than having a dedicated period in which to complete procedures in series. Two scenarios are included; one where the allocated slot time is determined by median procedure time and a second where it is set to achieve a maximum of 10% of procedures with over-runs. With slot length based on median procedure time, the mean EOT for PBP is 190 minutes, while for AD it is 90 minutes. However, over-run rates with these slot lengths are very high and would be unlikely to be tolerated. The second, more realistic scenario produces EOT of 226 minutes for PBP and 102 for AD. So with open scheduling, resource use is 39% higher for PBP and 22% higher for AD compared with a simple calculation assuming fully-flexible scheduling.

Under block scheduling, if block length is fixed and cannot be shortened from 8 hours, then optimisation takes place by changing the number of procedures. With only 4 procedures, the AD schedule is significantly under-utilised. The daily schedule can be increased to 5 procedures without incurring over-runs on more than 10% of days. For PBP, however, because of the average procedure time of almost 3 hours, it is not possible to add another procedure without over-running the block almost every day. With 2 and 5 procedures respectively, EOT for PBP is 242 minutes and for AD is 96 minutes. For AD this is only 15% more than the resource utilisation under fully-flexible assumptions, but for PBP it is 48% higher.

Block scheduling may also be optimised by shortening the length of a block - in some hospitals it may be possible for a surgeon to release some of their block time for other use. The time must be of a usable length and reliably available. With 4 AD procedures scheduled, it is possible to shorten the block length by 100 minutes without incurring over-runs more than 10% of the time. For PBP, with 2 procedures scheduled, the block can be shortened by 60 minutes to meet the same criteria. For PBP this delivers an EOT of 213 minutes, 31% more than AOT. For AD the EOT is 96 minutes, 15% above AOT. This is approximately the approach taken by the study centre. With 4 AD procedures scheduled the centre is able to use the available 100 minutes to schedule additional non-complex procedures. However, with PBP, 60 minutes of reliable time at the end of the block is on the margin of being usable. The DES model also predicts this was feasible only in later years of the study, with 40 minutes more representative of the average over the full study period. This fits with the centre's experience that it was more difficult to schedule additional procedures after PBP and the time was often unused. EOT for PBP would then be the same as the non-optimised scenario at 242 minutes.

4. DISCUSSION

4.1. Study findings

Our results illustrate how simple micro-costing approaches can under-estimate effective OR resource usage, as measured by EOT, in anything other than a theoretical, fully-flexible OR configuration. OR scheduling practices generate idle time where expenses are incurred but surgical cases are not present in the OR, and there is a cause-and-effect relationship between the distribution of procedure duration and idle time. Whether micro-costing is top-down or bottom-up, it will reflect average procedure duration which is representative of AOT and does not include idle time. To the extent that resource costs are under-estimated, the under-estimation will be greater for procedures of long and unpredictable duration versus those that are shorter and more predictable. So if two technologies are being directly compared and have differences in distribution of procedure duration, a bias is introduced into the assessment of relative costs. This may result in inefficient decisions in choices of health technologies. We demonstrate that this bias can be significant. In one scenario we show that OR resources consumed would be under-estimated for AD by 15% but the under-estimate increases to 48% for PBP. This difference is attributable to two factors: longer average procedure duration, which makes it more difficult to fit PBP into OR schedules, regardless of variability; and the "fat tail" of the distribution of PBP procedure duration, which means a large buffer needs to be included in the OR schedule to accommodate a small but significant probability of very long procedure times.

Some costing approaches may take account of OR idle time and allocate across all procedures completed. For example, by allocating total OR available time over a period (perhaps a week or a month) across procedure types, based on the proportion of actual occupancy time during the period. An alternative approach is to apply a universal multiplier to up-weight procedure durations to account for OR utilisation of less than 100% (e.g. average procedure duration x 1.25). Implicitly both of these approaches allocate idle time in proportion to average procedure duration. However, as demonstrated, short and predictable procedures generate proportionally less idle time than long and variable procedures, for a given target level of OR over-runs. So although on average, across all procedures, resource consumption will no longer be under-estimated, there is still the same bias introduced to relative costs of technologies with different distributions of procedure duration. Costs of procedures with long and variable duration will be under-estimated while those of procedures with short and predictable duration will be over-estimated.

Only costing approaches which in some way allocate idle time in line with the degree to which a procedure creates the need for it, will produce unbiased relative cost estimates for procedures. This is far from simple to do. The approach used in this study is one potential method, specifically: collect real-world procedure duration data at patient level for alternative technologies considered; analyse and fit the data to a defined distribution; use regression analysis to adjust for the impact of covariates; and assess using a simulation model that seeks to replicate the stochastic elements within the OR schedule and the set of decision rules that govern daily OR operations. To do this accurately is a time consuming exercise. It is also highly setting-specific - conclusions from one hospital may not be appropriate to another, depending on the specific methods of OR planning in place. Simulation models are also a necessary simplification of the indeterminate complexity of human organisations - it is by simplifying that they are able to deliver insights, but they provide only approximations to likely outcomes in the real world. An alternative would be to develop heuristic approaches or "rules of thumb" to convert measures of AOT to estimates of EOT including idle time, based on measures of duration *and* variability. This would reduce the need for simulation modeling, but would still require measurement, analysis and understanding of distributions of procedure duration, as opposed to just mean or median values.

4.2. Implications of study findings

Accounting for the impact of the distribution of procedure duration on costs within economic evaluations, while also providing results that remain generalisable, is challenging. But we demonstrate that the impact can be significant, with an important bearing on decisions, and so should not be ignored. The awareness of the potential costs of variability of procedure duration amongst operational researchers and hospital OR planning specialists does not appear to extend to those conducting or making decisions based on economic evaluations of relevant technologies. However, if costs of variability are not measured, then they will not be understood, and if they are not understood, they cannot be managed. Managing costs includes the selection of optimal healthcare technologies from alternatives.

HTA guidelines shape the approaches used by decision-makers to assess and prioritise alternative technologies. While it is challenging to specify appropriate techniques for all potential circumstances, outlining only general principles may lead to simple and generic approaches to cost estimation which will fail to capture important differences between technologies. Guidelines should be developed which encourage greater attention on and a more detailed approach to estimating costs for hospital procedures in particular.

It may not be appropriate to use full simulation approaches to estimate costs on a routine basis, but at the very least awareness of the potential value generated by more "OR-friendly" technologies should be increased. A more refined approach to costs is required and in particular evaluation of OR procedures should value length and predictability. This will require a greater focus on and understanding of the distribution of procedure times, how this can impact on value and how changes could be made to technologies and processes to optimise performance. For example, if difficult cardiac anatomy is the cause of "surprise" long procedure times, is there value in conducting diagnostics in advance to identify those abnormal patients who are likely to require a longer time in the OR? If such a diagnostic is available, what would be an acceptable cost to eliminate the uncertainty over OR occupancy time?

Greater understanding of variability of procedure duration will require more real-world studies that generate distributional data *and* publish more than just the mean and standard deviation. Such studies, if they are to capture accurately the impact of tail variability, e.g. the 90th percentile, will require larger sample sizes than would be the case for traditional clinical studies focused only on estimating mean difference.

4.3. Study limitations

A key limitation of the study is the focus on one centre, its specific experience with two procedures and particular mode of OR planning. The results may not be generalisable to other hospitals, particularly those with a much more flexible approach to OR planning. Further research to validate that the same degree of bias may apply within other settings and to other types of procedures would be valuable.

The procedure duration data used within the study is observational data collected over a 7-year period in a real-world setting. Patient characteristics were in general well matched across the two types of procedures, but with some differences that could act as confounders. There were also incremental changes to personnel, technology, and practices over time which mean that observations from earlier in the study period may be less informative than those at the end of the study period. Methods have been employed within the study to control for these issues, but all influence may not have been eliminated from the results.

5. CONCLUSION

For health technology decisions where a substantial part of costs are OR resources, particularly choices between alternative surgical procedures with similar outcomes, understanding the real-world implications of the distribution of procedure durations becomes important. Cost estimates based on simple micro-costing methods, where costs are linear with average procedure duration, are likely to produce relative cost estimates biased against technologies with shorter and more predictable procedure times. Comparing the full distributions of procedure duration and understanding the implications for OR capacity utilisation will lead to estimates of average costs more representative of the true impact of implementing the technologies. Studies that generate and publish distributional data beyond just the mean and standard deviation should be encouraged. Sample sizes above those required to demonstrate mean difference should also be considered, to allow the impact of tail variability to be captured accurately.

REFERENCES

1. Tyler DC, Pasquariello CA, Chen CH. Determining optimum operating room utilization. Anesth Analg. 2003 Apr;96(4):1114-21

2. Dexter F, Macario A, Lubarsky DA, Burns DD. Statistical method to evaluate management strategies to decrease variability in operating room utilization. Anesthesiology 1999; 91:262-74.

3. Cardoen B, Demeulemeester E, Belien J. Operating room planning and scheduling: A literature review. European Journal of Operational Research 201 (201), 921-932.

4. Raikou M, McGuire A. Parametric estimation of medical care costs under conditions of censoring. LSE Health, London School of Economics and Political Science, working paper 13/2009, February 2009.

5. Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada. 3rd edition, 2006.

6. National Institute for Health and Care Excellence, UK. Guides to the methods of technology appraisal 2013. 4 April 2013. http://publications.nice.org.uk/pmg9 .

7. Australian Government: Department of Health, Medical Services Advisory Committee. Technical Guidelines for preparing assessment reports for the Medical Services Advisory Committee, Draft for public consultation, 2015.

8. Haute Autorité de Santé. Choices in methods for economic evaluation: a methodological guide. Department of Economics and Public Health Assessment. October 2012.

9. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ. 1996 Aug 3;313(7052):275-83.

10. Adam T, Koopmanschap MA, Evans DB. Cost-effectiveness analysis: can we reduce variability in costing methods? Int J Technol Assess Health Care. 2003 Spring;19(2):407-20

11. Tan SS, Rutten FF, van Ineveld BM et al. Comparing methodologies for the cost estimation of hospital services. Eur J Health Econ. 2009 Feb;10(1):39-45.

12. Drummond M, Sculpher M, Torrance G et al. Methods for the economic evaluation of health care programmes. Third edition. 2005. Oxford University Publication.

13. Tan S, Serdén L, Geissler A et al. DRGs and cost accounting: which is driving which? In: Busse

R, Geissler A, Quentin W, Wiley M (eds). Diagnosis-Related Groups in Europe. New York: Open University Press. 2011. 59–74.

14. Northcott D, Llewellyn S. Cost Variability in Healthcare. Chartered Institute of Management Accountants, February 2003.

15. Smith M, Barnett PG. Direct measurement of health care costs. Med Care Res Rev. 2003. 60 (3 Suppl):74S-91S.

16. Mercier G, Naro G. Costing hospital surgery services: the method matters. PLoS One. 2014 May 9;9(5):e97290.

17. Whynes DK, Walker AR. On approximations in treatment costing. Health Econ. 1995 Jan-Feb;4(1):31-9.

18. Luce B, Manning W, Siegel J, Lipscomb J. Estimating costs in cost-effectiveness analysis. In Gold M, Siegel J, Russell L, Weinstein M (eds) Cost-effectiveness in health and medicine. 1996. pp: 176-213.

19. Cheng X, Hu Q, Zhou C et al. The long-term efficacy of cryoballoon vs irrigated radiofrequency ablation for the treatment of atrial fibrillation: A meta-analysis. Int J Cardiol. 2015 Feb 15;181:297-302.

20. DeVille JB, Svinarich JT, Dan D, et al. Comparison of resource utilization of pulmonary vein isolation: cryoablation versus RF ablation with three-dimensional mapping in the value PVI study, J. Invasive Cardiol. 26 (2014) 268–272.

21. Klein G, Lickfett L, Schreieck J et al. A Comparison of “anatomically designed” and “point-by-point” catheter ablations for human atrial fibrillation in terms of procedure timing and costs in German hospitals. Europace. 2015 Jul;17(7):1030-7.

22. Eckard N, Davidson T, Walfridsson H, Levin LA. Cost effectiveness of catheter ablation treatment for patients with symptomatic atrial fibrillation. Journal of Atrial Fibrillation. 2009;1(8):461-70.

23. Blackhouse G, Assasi N, Xie F et al. Cost-effectiveness of catheter ablation for rhythm control of atrial fibrillation. International Journal of Vascular Medicine 2013, Art. ID 262809.

24. Reynolds MR, Lamotte M, Todd D et al. Cost-effectiveness of cryoballoon ablation for the management of paroxysmal atrial fibrillation. Europace. 2014 May;16(5):652-9.

25. McKenna C, Palmer S, Rodgers M et al. Cost-effectiveness of radiofrequency catheter ablation for the treatment of atrial fibrillation in the United Kingdom. Heart. 2009;95(7):542-9.

26. Khaykin Y, Wang X, Natale A et al. Cost comparison of ablation versus antiarrhythmic drugs as first-line therapy for atrial fibrillation: an economic evaluation of the RAAFT pilot study. J Cardiovasc Electrophysiol. 2009 Jan;20(1):7-12.

27. Rodgers M, McKenna C, Palmer S et al. Curative catheter ablation in atrial fibrillation and typical atrial flutter: systematic review and economic evaluation. HTA Technology Assessment Report. NIHR Health Technology Assessment programme; 2008. 12(34).

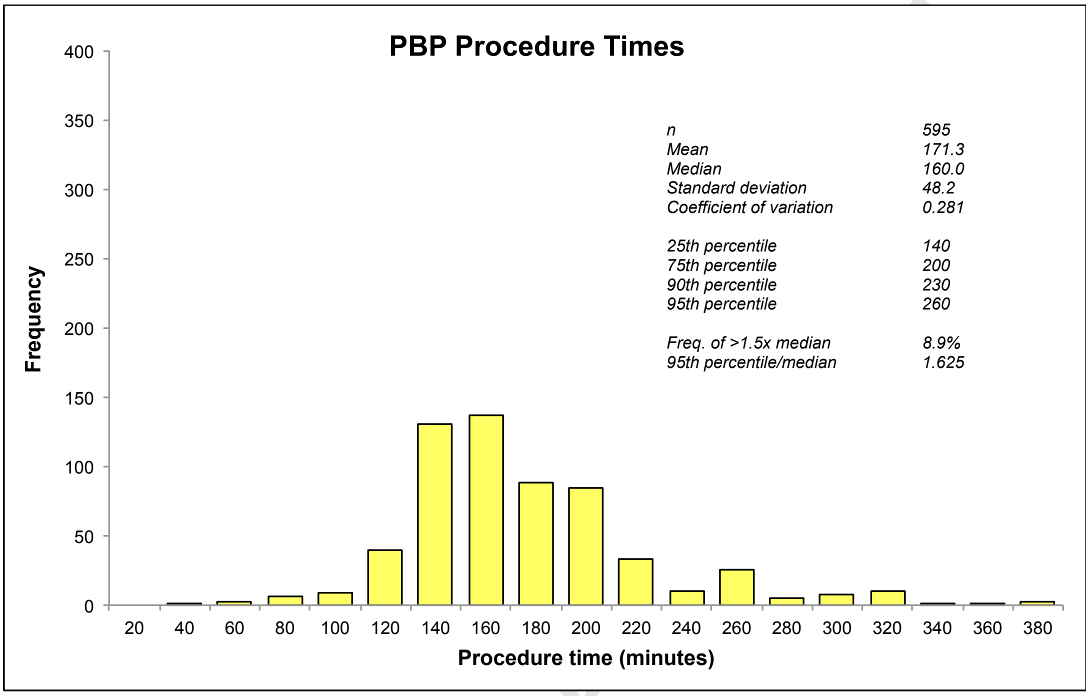
28. Mugnai G, Irfan G, de Asmundis C et al. Complications in the setting of percutaneous atrial fibrillation ablation using radiofrequency and cryoballoon techniques: A single-center study in a large cohort of patients. Int J Cardiol. 2015 Oct 1;196:42-9.

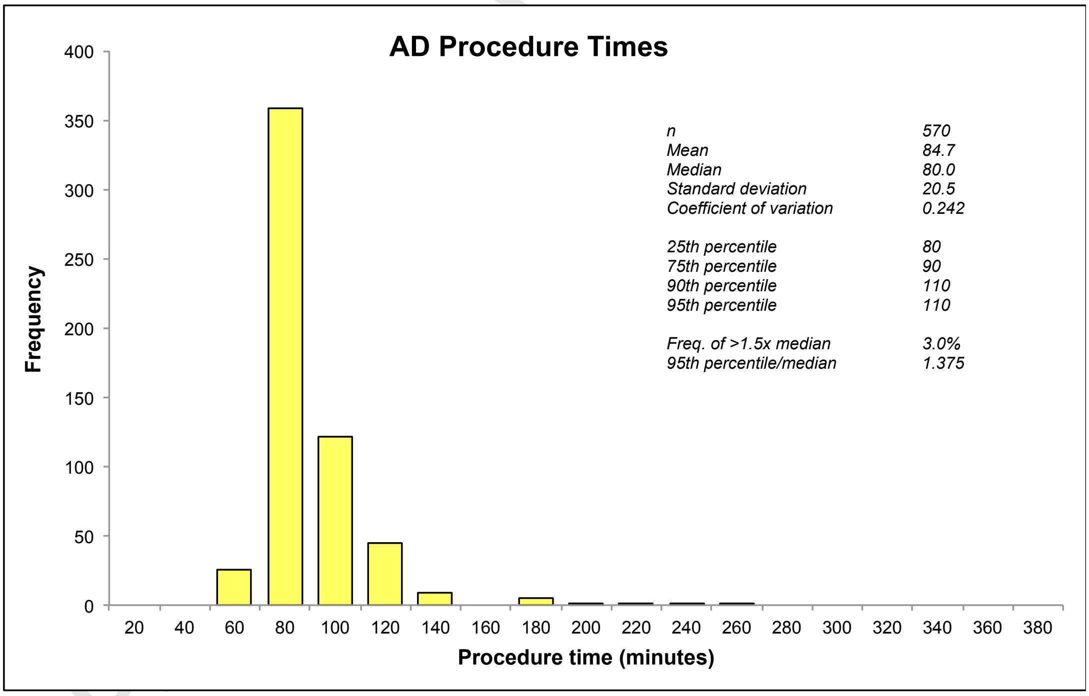
29. Mugnai G, Chierchia GB, de Asmundis C et al. Comparison of pulmonary vein isolation using cryoballoon versus conventional radiofrequency for paroxysmal atrial fibrillation. Am J Cardiol. 2014 May 1;113(9):1509-13.

30. Juliá J, Chierchia GB, de Asmundis C et al. Regular atrial tachycardias following pulmonary vein isolation for paroxysmal atrial fibrillation: a retrospective comparison between the cryoballoon and conventional focal tip radiofrequency techniques. J Interv Card Electrophysiol. 2015 Mar;42(2):161-9.

31. Chierchia GB, Capulzini L, Droogmans S et al. Pericardial effusion in atrial fibrillation ablation: a comparison between cryoballoon and radiofrequency pulmonary vein isolation. Europace. 2010 Mar;12(3):337-41

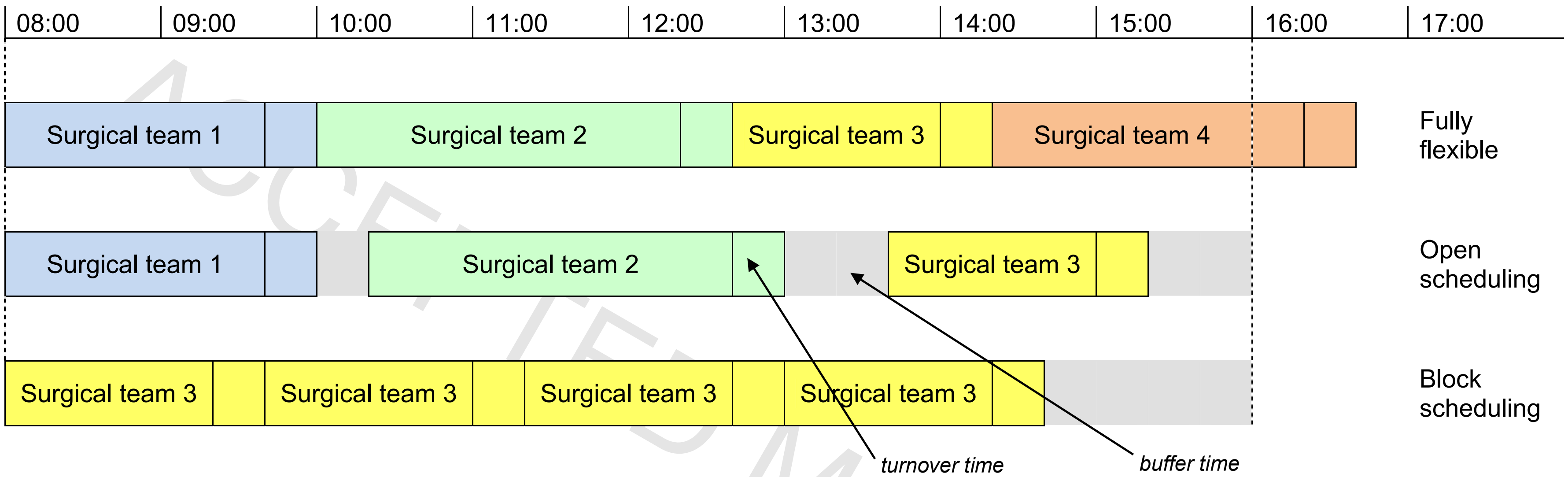
Figure 1: Observed distributions of procedure duration





Procedure duration is recorded "skin-to-skin" time plus fixed estimate of patient OR occupancy before and after (2 x 10 minutes prep & wake time).

Figure 2: OR Scheduling Methods



- OR time is allocated in 20 minute segments.
- Scheduled OR day is 08:00 - 16:00 for open and block scheduling; and 08:00 to open-ended finish for fully-flexible scheduling.
- Fully-flexible scheduling imposes queuing on both surgical cases and surgical teams, other than the first case of the day.
- Open scheduling provides fixed start times for cases and teams, subject to some uncertainty on over-runs. Buffer time is built into the schedule between cases to reduce uncertainty to tolerable levels.
- Block scheduling provides fixed start time for surgical team. Cases are queued and buffer time required only at the end of the day.

- **Fully-flexible scheduling** eliminates OR idle time. But queuing costs are imposed on surgical teams and there is an uncertain end of working day. High set-up/turnover time between surgical teams may also mean that although measured utilisation appears high, throughput may not be maximised.
- **Open scheduling** eliminates queuing of surgical teams, but reduces OR utilisation with buffer time between procedures and at the end of the OR day. High set-up/turnover time between surgical teams remains.
- **Block scheduling** eliminates queuing of surgical teams but queues cases during the block, with buffer time only required at the end of the day. Set-up/turnover time is also reduced. However, block scheduling creates inflexibility of scheduling across surgical teams and requires sufficient volume of procedures to be efficient.

Figure 3. DES Model Fitted Distributions (probability density)

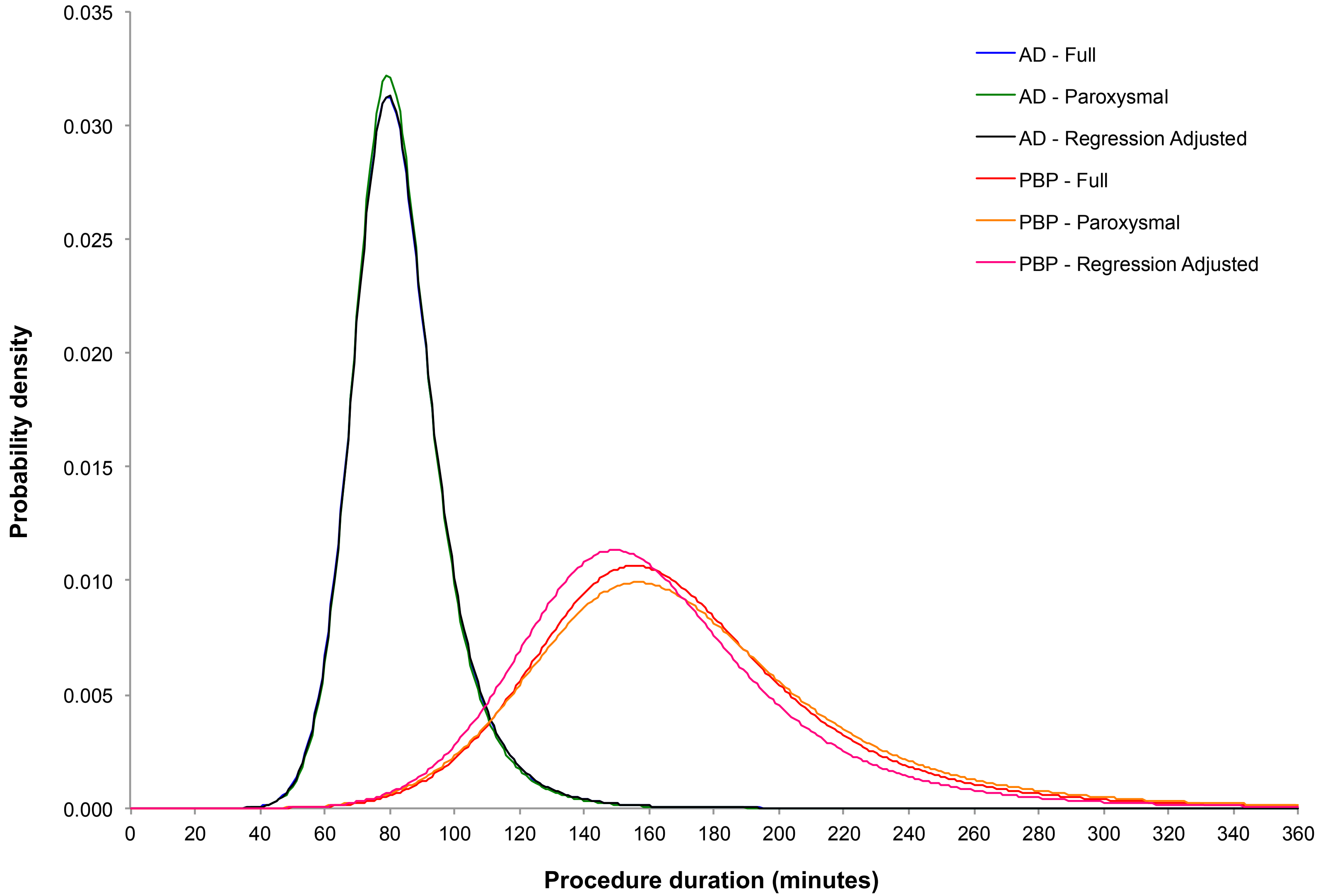


Table 1: DES Model - Key Parameters

	DES Modes			
	Fully Flexible	**Open Scheduling**	**Block Scheduling**	**Source**
Procedure Duration				
PBP "skin-to-skin" distribution	Range of log-logistic distributions applied			Fitted distributions from centre data
AD "skin-to-skin" distribution	Range of log-logistic distributions applied			Fitted distributions from centre data
Preparation time (mins)	10	10	10	Centre estimate
Wake time (mins)	10	10	10	Centre estimate
Turnover time (mins)	0	0	0	Simplifying assumption*
Scheduling				
Over-run tolerance	NA	10%	10%	Dexter et al [29]
OR shift start time	08:00	08:00	08:00	Centre scheduling practice
OR shift end time	Not defined	16:00	16:00	Centre scheduling practice
PBP slot length (mins)	NA	100	NA	90th percentile of procedure duration
AD slot length (mins)	NA	220	NA	90th percentile of procedure duration
Cancellations?	No	No	Yes	
PBP schedule review time	NA	NA	14:00	End of day - median procedure duration + 30 mins
AD schedule review time	NA	NA	15:30	End of day - median procedure duration + 30 mins
Slot increments (mins)	NA	20	20	Centre scheduling practice
Initial PBP procedures scheduled	1	1	2	Centre scheduling practice
Initial AD procedures scheduled	1	1	4	Centre scheduling practice
Overtime				
Overtime applies?	No	Yes	Yes	
Overtime rate	NA	1.5x	1.5x	Tyler et al [1]

** Turnover time will, by definition, be zero for block scheduling, but non-zero for open scheduling and fully flexible modes. We simplify by setting all to zero, to focus results on the impact of the actual procedure duration.*

Table 2: DES Model, Actual and Fitted Distributions

Procedure duration (minutes)	Full Dataset				Paroxysmal Only				Regression Adjusted			
	PBP (n = 595)		**AD** (n = 570)		**PBP** (n = 308))		**AD** (n = 509)		**PBP** (n = 595)		**AD** (n = 570)	
	Actual	**Fitted**	**Actual**	**Fitted**	**Actual**	**Fitted**	**Actual**	**Fitted**	**Actual**	**Fitted**	**Actual**	**Fitted**
Mean	171.3	170.4	84.7	83.4	173.3	173.5	84.1	83.2	173.3	163.0	84.1	83.2
Median	160.0	163.5	80.0	81.6	160.0	165.5	80.0	81.5	160.0	156.7	80.0	81.5
Std. Deviation	48.2	48.5	20.5	15.8	49.9	53.5	18.7	15.2	49.9	45.4	18.7	15.2
Coeff. of Variation	0.281	0.285	0.242	0.189	0.288	0.308	0.222	0.183	0.288	0.279	0.222	0.183
Minimum	34.0	47.1	50.0	38.1	34.0	49.7	50.0	38.8	34	48.3	50	35.6
Maximum	380.0	700.1	245.0	265.8	352.0	941.0	245.0	237.9	352	600.9	245	218.2
25th Percentile	140.0	139.3	80.0	73.3	140.0	139.6	80.0	73.4	140.0	133.9	80.0	73.4
75th Percentile	200.0	192.6	90.0	91.2	200.0	197.1	90.0	90.8	200.0	184.0	90.0	90.8
90th percentile	230.0	227.7	110.0	102.3	230.0	235.5	105.0	101.5	230.0	216.7	105.0	101.5
95th percentile	260.0	255.5	110.0	110.9	278.3	266.2	110.0	109.7	278.3	242.7	110.0	109.7
Frequency of > 1.5x median	8.9%	6.4%	3.0%	2.1%	4.9%	7.4%	2.5%	1.9%	4.9%	6.1%	2.5%	1.9%
95th percentile/median	1.625	1.563	1.375	1.359	1.739	1.608	1.375	1.347	1.739	1.549	1.375	1.347

Statistics for "fitted" data are derived from DES simulation, using the fitted probability distributions within the model.
Actual procedure duration is derived from recorded "skin-to-skin" time, plus 2 x 10 minutes for prep & wake time.
Fitted procedure duration derived from "skin-to-skin" probability distributions, plus 2 x 10 minutes (fixed, no variability) for prep & wake time.
Regression adjusted distributions assume 100% paroxysmal AF cases, 100% first procedures, final year of the study, no hypertension comorbidity.
Regression adjusted distributions are compared to paroxysmal only actual data.

Table 3: DES Model Results

Procedure duration (minutes)	PBP					AD				
	Scheduled Procedures	Block/slot length	AOT per procedure	EOT per procedure	EOT/AOT Ratio	Scheduled Procedures	Block/slot length	AOT per procedure	EOT per procedure	EOT/AOT Ratio
Flexible running	1	N/A	163.0	163.0	1	1	N/A	83.5	83.5	1
Open scheduling										
- median time	1	160	161.4	188.4	1.17	1	80	83.5	90.3	1.08
- 10% over-run tolerance	1	220	161.7	226.3	1.39	1	100	83.5	102.1	1.22
Block scheduling										
- non-optimised	2	480	163.1	241.5	1.48	4	480	83.5	120.1	1.44
- optimised by # procedures	2	480	163.1	241.7	1.48	5	480	83.5	96.3	1.15
- optimised by block length	2	420	163.0	213.5	1.31	4	380	83.5	95.7	1.15

All results using the regression adjusted distributions for the DES model
Slot length is constrained to a multiple of 20 minute increments, in line with OR planning practices