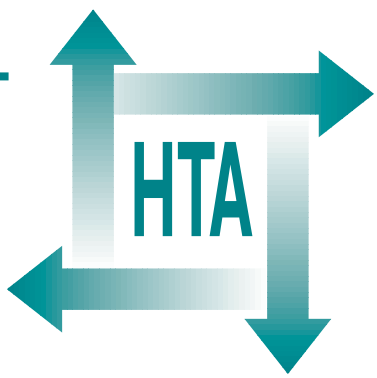
Health Technology Assessment 2001; Vol. 5: No. 12

Methodology

Statistical assessment of the learning curves of health technologies

CR Ramsay
AM Grant
SA Wallace
PH Garthwaite
AF Monk
IT Russell

Health Technology Assessment
NHS R&D HTA Programme



How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Statistical assessment of the learning curves of health technologies

CR Ramsay[1]
AM Grant[1*]
SA Wallace[1]
PH Garthwaite[2]
AF Monk[3]
IT Russell[3]

[1] Health Services Research Unit, University of Aberdeen, Aberdeen, UK
[2] Open University, Milton Keynes, UK
[3] University of York, York, UK

[*] Corresponding author

Competing interests: none declared

Published April 2001

This report should be referenced as follows:

Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, Russell IT. Statistical assessment of the learning curves of health technologies. *Health Technol Assess* 2001;**5**(12).

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE. Copies of the Executive Summaries are available from the NCCHTA website (see opposite).

The study described in this report has also been published as:

Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, Russell IT. Assessment of the learning curves in health technologies: a systematic review. *Int J Technol Assess Healthcare* 2000;**16**:1095–108.

NHS R&D HTA Programme

The NHS R&D Health Technology Assessment (HTA) Programme was set up in 1993 to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS.

Initially, six HTA panels (pharmaceuticals, acute sector, primary and community care, diagnostics and imaging, population screening, methodology) helped to set the research priorities for the HTA Programme. However, during the past few years there have been a number of changes in and around NHS R&D, such as the establishment of the National Institute for Clinical Excellence (NICE) and the creation of three new research programmes: Service Delivery and Organisation (SDO); New and Emerging Applications of Technology (NEAT); and the Methodology Programme.

Although the National Coordinating Centre for Health Technology Assessment (NCCHTA) commissions research on behalf of the Methodology Programme, it is the Methodology Group that now considers and advises the Methodology Programme Director on the best research projects to pursue.

The research reported in this monograph was funded as project number 96/25/02.

The views expressed in this publication are those of the authors and not necessarily those of the Methodology Programme, HTA Programme or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for any recommendations made by the authors.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The editors and publisher have tried to ensure the accuracy of this report but do not accept liability for damages or losses arising from material published in this report. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

Published by Core Research, Alton, on behalf of the NCCHTA.
Printed on acid-free paper in the UK by The Basingstoke Press, Basingstoke.

M

Contents

List of abbreviations

AIC	Akaike information criteria
ANOVA	analysis of variance
ARIMA	autoregressive integrated moving averages
CI	confidence interval
cusum	cumulative sum
GORD	gastro-oesophageal reflux disease
LNSEQ	logarithm of operation sequence
MeSH	medical subject headings
R^2	coefficient of determination
RCT	randomised controlled trial
RSS	residual sum of squares
SD	standard deviation

For details of electronic databases mentioned in the text, see appendix 3.

Executive summary

Objectives

- To describe systematically studies that directly assessed the learning curve effect of health technologies.
- Systematically to identify 'novel' statistical techniques applied to learning curve data in other fields, such as psychology and manufacturing.
- To test these statistical techniques in data sets from studies of varying designs to assess health technologies in which learning curve effects are known to exist.

Methods

Study selection

Health technology assessment literature review

For a study to be included, it had to include a formal analysis of the learning curve of a health technology using a graphical, tabular or statistical technique.

Non-health technology assessment literature search

For a study to be included, it had to include a formal assessment of a learning curve using a statistical technique that had not been identified in the previous search.

Data sources

Six clinical and 16 non-clinical biomedical databases were searched. A limited amount of hand-searching and scanning of reference lists was also undertaken.

Data extraction

Health technology assessment literature review

A number of study characteristics were abstracted from the papers such as study design, study size, number of operators and the statistical method used.

Non-health technology assessment literature search

The new statistical techniques identified were categorised into four subgroups of increasing complexity: exploratory data analysis; simple series data analysis; complex data structure analysis, generic techniques.

Testing of statistical methods

Some of the statistical methods identified in the systematic searches for single (simple) operator series data and for multiple (complex) operator series data were illustrated and explored using three data sets. The first was a case series of 190 consecutive laparoscopic fundoplication procedures performed by a single surgeon; the second was a case series of consecutive laparoscopic cholecystectomy procedures performed by ten surgeons; the third was randomised trial data derived from the laparoscopic procedure arm of a multicentre trial of groin hernia repair, supplemented by data from non-randomised operations performed during the trial.

Results

Health technology assessment literature review

Of 4571 abstracts identified, 272 (6%) were later included in the study after review of the full paper. Some 51% of studies assessed a surgical minimal access technique and 95% were case series. The statistical method used most often (60%) was splitting the data into consecutive parts (such as halves or thirds), with only 14% attempting a more formal statistical analysis. The reporting of the studies was poor, with 31% giving no details of data collection methods.

Non-health technology assessment literature search

Of 9431 abstracts assessed, 115 (1%) were deemed appropriate for further investigation and, of these, 18 were included in the study. All of the methods for complex data sets were identified in the non-clinical literature. These were discriminant analysis, two-stage estimation of learning rates, generalised estimating equations, multilevel models, latent curve models, time series models and stochastic parameter models. In addition, eight new shapes of learning curves were identified.

Testing of statistical methods

No one particular shape of learning curve performed significantly better than another. The performance of 'operation time' as a proxy for learning differed between the three procedures.

Multilevel modelling using the laparoscopic cholecystectomy data demonstrated and measured surgeon-specific and confounding effects. The inclusion of non-randomised cases, despite the possible limitations of the method, enhanced the interpretation of learning effects.

Conclusions

Health technology assessment literature review

The statistical methods used for assessing learning effects in health technology assessment have been crude and the reporting of studies poor.

Non-health technology assessment literature search

A number of statistical methods for assessing learning effects were identified that had not hitherto been used in health technology assessment. There was a hierarchy of methods for the identification and measurement of learning, and the more sophisticated methods for both have had little if any use in health technology assessment. This demonstrated the value of considering fields outside clinical research when addressing methodological issues in health technology assessment.

Testing of statistical methods

It has been demonstrated that the portfolio of techniques identified can enhance investigations of learning curve effects.

Implications and recommendations

For health technology assessment

- A change over time in the performance of a technology because of learning complicates evaluation and impedes rigorous evaluation.
- Useful parameters for describing learning in health technology assessment are the rate and length of learning and the final skill level.
- Reliable assessment of learning effects is most likely to come from prospectively collected data on multiple operators or institutions.
- The experience of the operator should be described each time the procedure is performed. This is particularly important in circumstances, such as randomised trials, in which the technology may have parallel use outside the trial.
- Collection of non-randomised data alongside a randomised controlled trial may, despite possible limitations, aid the interpretation of learning effects.
- Reports of studies of learning should, as a minimum, describe the number and experience of the operators, the data source, the proportion of procedures performed by individual operators and the level of care.
- Proxy measures of learning have advantages and limitations, and finding a suitable measure can be difficult.
- Investigators should consider and adjust for any confounding factors.
- The simplest methods within the hierarchies described in this report should be used in a parsimonious way.
- When there are multiple operators, a method should be used which takes into account the hierarchical nature of the data.

For further research

- Further empirical testing of the techniques identified is required. The generalisability of the various shapes and methods that were identified needs to be assessed for a variety of health technologies.
- Methods for estimation of the time taken to reach an asymptote should be explored further.
- Variables that are good proxies for learning need to be identified.
- Relatively rare, dichotomous outcomes are often the best measures of performance but are currently the least tractable to analysis. Further methodological research is needed to address this issue.
- Further empirical work is required to identify the optimal method for assessing learning curves within randomised controlled trials.
- The impact of learning curve factors on economic evaluations should be explored.
- Appropriate prospective data collection should be built into future evaluations.
- A theory-based approach to learning should be investigated in the context of health technology assessment.
- Parallels between learning curve and quality assurance issues should be explored.

Chapter 1
Introduction

The performance of many repeated tasks changes with experience over time. Improvement tends to be most rapid at first and then to tail off over time until a steady state is reached. The term 'learning curve' is often used as shorthand to describe this phenomenon. Learning curves have long been recognised outside health technology assessment, in psychology, manufacturing and aviation. The term was first used in healthcare in the 1970s, and came to greater prominence in the 1980s in the context of minimal access surgery.

Changes in performance due to learning present particular difficulties in health technology assessment.[1–6] Many innovators and early enthusiasts are reluctant to apply rigorous evaluation to a new technology whose performance may change with learning. They argue, often vociferously, that early assessment gives a distorted picture which is biased against the new technology.[7,8] The argument is illustrated in *Figure 1*. If assessment is carried out at time point A, then the conventional treatment is preferred to the new treatment; if assessment is carried out at time point B further along the learning curve, then the new treatment performs better than the standard. Often, however, by the time that the technology has clearly stabilised (time point B, say), these same people have been convinced of the worth of the technology on the basis of poor quality evidence. They then argue that rigorous evaluation is now unethical if it involves withholding the technique or procedure from potential patients.[9] This situation has been encapsulated in what has become known as Buxton's law – "It is always too early [for rigorous evaluation] until, unfortunately, it's suddenly too late".[10]

In principle, statistical description of a learning curve and subsequent adjustment of an evaluation to take account of learning effects should solve this problem. Identifying the point at which the curve flattens (the asymptote) would allow subsequent evaluation free of any learning effect; description of the learning curve could allow adjustment of earlier evaluation during the learning period, and characterisation of user performance could guide training and subsequent implementation policies. While the advantages of such formal statistical analyses have been suggested previously,[7,11–14] exactly how they should be undertaken has been unclear.

One reason for this is that changes in performance with learning are more complex than they may at first seem. This can be illustrated by minimal access surgery. Learning is likely to depend on whether the operator is using minimal access techniques for the first time or whether he or she already has experience of another different minimal access operation. The learning curve will be influenced by the experience of the supporting surgical team. Performance is likely to be less good if everyone in a team is learning their role in the operation (so-called 'institutional learning'), rather than just the

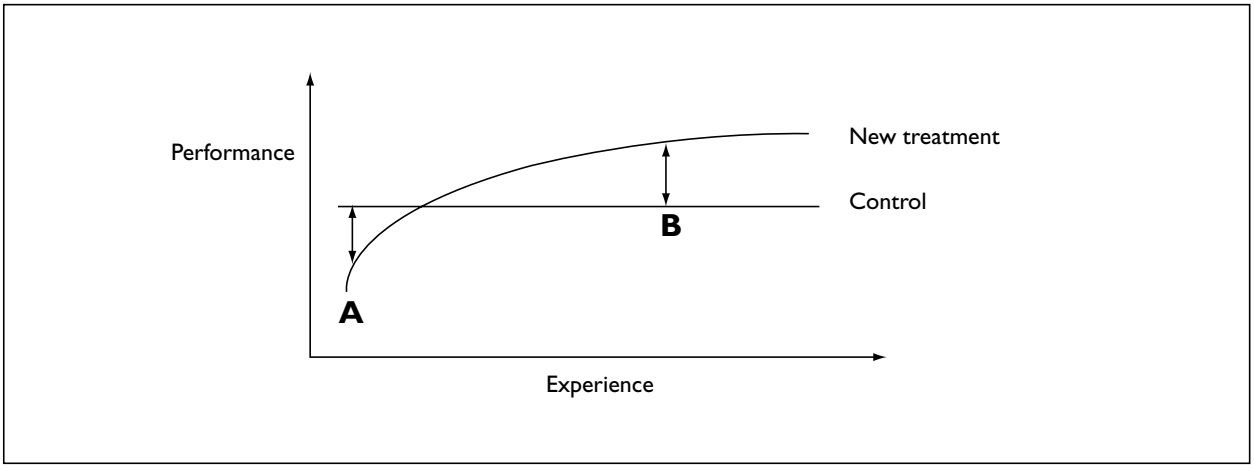


FIGURE 1 *Outcome of evaluation at two points on the learning curve*

operator.[15] The broader clinical community also influences individual learning through policy decisions, such as the number of operations that must be performed before the operator is considered 'experienced' or the type of equipment to be used for an operation.

Changes in case-mix complicate the assessment of learning curves. Sometimes, as an operator becomes more experienced, the cases that are attempted become more difficult. As a consequence, outcome may appear to deteriorate.[4,9,11–13,16–18]

Another difficulty is the identification of reliable measures of learning. Again using the example of minimal access techniques, the most obviously relevant measures, such as complications or conversions, may be too infrequent or unreliable[15] for reliable analysis. Other proxies for learning, such as duration of an operation, may be more tractable to statistical analysis but relatively poor measures of performance. (Performing an operation more quickly does not necessarily indicate better performance; it may, for example, increase the risk of complications.)

It was against this background that the authors were commissioned by the NHS R&D HTA Programme to undertake this study. The aim was to identify statistical methods for characterising learning that would allow reliable description of learning curves of both individuals and institutions. It was important that the methods would be applicable to the various study designs (randomised controlled trials (RCTs)), observational case series and cohort studies) that can be used to assess health technologies whose performance might be influenced by learning.

The project had three components.

1. A systematic description of studies that directly assessed the learning curve effect of health technologies, describing the strengths and weaknesses of the study designs and statistical methods that were employed (reported in chapter 2).
2. A systematic identification of 'novel' statistical techniques applied to learning curve data in other fields, such as psychology and manufacturing (described in chapter 3).
3. Empirical testing of these statistical techniques in sets of data describing a variety of health technologies for which learning curve effects are known to exist: single operator case series (chapter 4); multiple operator case series (chapter 5); multiple operator RCTs with and without non-randomised data (chapter 6).

The main conclusions and implications of the project, including recommendations for future research, are presented in chapter 7.

Chapter 2

Assessment of the learning curve in health technologies: a systematic review

Introduction

The first phase of the project was a systematic description of studies that directly assessed the learning curves of health technologies. The aim was to characterise the various methods identified, to quantify the extent of their use, and to explore their strengths and weaknesses.

Methods

Search strategy for the identification of studies

The search strategy was first developed using MEDLINE by a statistician and a researcher who were experienced in literature searching. Search terms were developed from Medical Subject Headings (MeSH) terms, using the MeSH tree with scope notes and permuted index. Textword searching, that is, searching for terms in titles and abstracts, was also employed, using truncation and adjacency where appropriate.

The numbers of abstracts retrieved for each term were noted and the first 50 were assessed for relevance to learning curves. To optimise the return on the resources available, a focused search strategy was then developed. The most specific search term was chosen for use in MEDLINE (see appendix 1). Other less specific terms were tested but rejected because too many irrelevant studies were retrieved (appendix 1). The search strategy imposed no language or other similar limitations.

The search strategy was modified for other databases. The syntax was changed to suit that of the relevant search software and the thesaurus or indices of each database were used to identify equivalents of the MeSH terms used in MEDLINE (see appendix 1). Other terms were also tested in each of the other databases but were found to be less specific and were thus rejected (appendix 1). Search terms describing complex statistical techniques that may have been appropriate for assessment of the learning curve were also tested (appendix 1).

Systematic electronic bibliographic database searching

Eight databases were searched systematically: MEDLINE (1966 to March 1999); HealthSTAR (1975 to November 1998); EMBASE (1980 to February 1999); Science Citation Index (1981 to March 1999); Social Science Citation Index (1981 to March 1999); CINAHL (1982 to December 1998); BIOSIS (1985 to March 1999); the Cochrane Library (1999).

To estimate the number of studies that described the assessment of the learning curve in the body of an article but which would not have been identified by searching the abstract and title only, the full texts of the following databases were searched: MEDLINE Core Biomedical Collection (1993 to August 1998); Biomedical Collection II (1995 to October 1998); Biomedical Collection III (1995 to June 1998) (see appendix 1). This covered 46 journals in total.

Two electronic databases of ongoing studies were searched: British National Research Register (Issue 1, 1998) and Current Controlled Trials (to January 1999). The NHS Economic Evaluation Database (to April 1999) was also searched.

Handsearching of specific journals

No journal was identified for which handsearching was likely to yield a substantial dividend in terms of the identification of extra relevant studies – the relevant literature covered too many fields and journals. However, the *International Journal of Technology Assessment in Health Care* was identified as the place where new techniques to assess learning curves were most likely to be published. Rather than perform a full handsearch of the journal, a handsearch of all abstracts of full papers was undertaken.

Other methods of ascertainment of studies

To identify any other relevant studies, experts in the field were contacted, mainly members of the International Society for Health Technology Assessment and biostatisticians.

Register of possible studies

All possibly relevant reports were electronically imported or manually entered into the software package Reference Manager (v. 9.0 N: ISI ResearchSoft, USA). Details of the source of each article were added. All electronically derived abstracts and study titles were read by one statistician, in order to assess subject relevance. Those that were deemed to be of possible relevance, because they described a health technology assessment and also referred to a learning curve, were assigned keywords in Reference Manager and the full published paper was obtained. The exception was the searching of the full text version of MEDLINE, for which the full published paper was assessed for relevance to learning curves.

Full copies of study reports were assessed by the statistician, using a standard form, for subject relevance, eligibility and methodological quality. The assessor was not blinded to author, institution or journal.

Inclusion criteria

To be included in the review, a study had to analyse the learning curve formally using a graph, table or statistical technique. The methods of analysis were categorised as follows.

Descriptive

No statistical testing was performed but results were tabulated by experience or shown graphically. The graphical method required one axis to be the case sequence (or grouped case sequence).

Split group

The data were split by experience, and univariate testing of the discrete groups (generally halves or thirds) was performed. The statistical methods used included *t*-test, chi-squared test, Mann–Whitney *U*-test and simple analysis of variance (ANOVA). Also included in this category were reports that compared experienced with inexperienced surgeons.

Univariate (trend)

These tested for some form of trend by experience in the data. Methods included curve fitting, chi-squared test for trend or repeated measures ANOVA. If the data were split into categories, at least three categories were required, with the ordering formally taken into account.

Multivariate (split)

The data were split by experience as in the split group and multivariate testing of the groups was performed to adjust for other variables. For example, a study was included in this category if the experience variable had been dichotomised into the first 50 and the second 50 patients, and then included as a potential confounding variable in a regression analysis along with confounders such as the age and sex of patients.

Multivariate (trend)

Trend by experience was tested for in the data but adjustment for possible confounding variables was also included. These methods included logistic regression and multiple regression, with the experience variable treated as either continuous or ordinal.

Cumulative sum

Trend in experience was measured using the cumulative sum (cusum) procedure.[19] This is a graphical method for identifying trends in data.

Data abstraction and analysis

A single statistician abstracted the following: study design; study size; type of technology (minimal access, other surgical, or diagnostic); type of patient; level of learning assessed (operator or institution); number of operators; proportion of operators performing half of the procedures (to see whether one or a few operators dominated the series); type of institution; data source; prior knowledge of outcome before inclusion of patient; type of outcome used to assess learning; statistical method used (as categorised above).

A random 10% sample of possible studies was independently assessed by another statistician, with double abstraction of data from those studies meeting the inclusion criteria. A kappa statistic was calculated to measure agreement between assessors. Any differences of opinion were resolved by discussion.

Results

Literature search

Of the 4571 abstracts assessed, 559 (12%) were deemed appropriate for further investigation and 272 were later judged, on review of the full paper, to have included a formal assessment of the learning curve (*Table 1*). A complete list of all 272 studies is included in the references (see page 71). Of the 272 studies, 202 (74%) were identified from MEDLINE, with the next largest number being identified from EMBASE; however, this was at least in part a function of the order of the searching rather than the coverage of each

TABLE 1 *Summary of the bibliographic searches*

Source[†]	Number of abstracts assessed	Number of full papers assessed	Papers included in the review
MEDLINE	736	363	203
EMBASE	588	64	30
Science Citation Index	1235	43	25
MEDLINE (full text)	66	66	7
BIOSIS	629	5	5
CINAHL	28	2	1
HealthSTAR	21	3	1
Cochrane Library	54	4[*]	0
International Journal of Technology Assessment in Health Care	862	10	1
Social Science Citation Index	352	0	0
Total	**4571**	**560**	**273**

[†] *Details of electronic databases are presented in appendix 3*
[*] *These were four systematic reviews of technologies that mentioned learning curve effects*

database. Of the included papers, 39 (15%) were published in non-English language journals. In a further 24 (4%) of the 560 assessed full papers, the authors mentioned that learning did (or did not) take place in their study but gave no indication as to how it had been assessed; hence, these papers were excluded. The numbers of studies published per year are displayed in *Figure 2*; a progressive rise is seen in number of relevant studies, particularly during the 1990s.

An additional seven papers were identified in the MEDLINE (full text) database that assessed learning which would not have been identified by

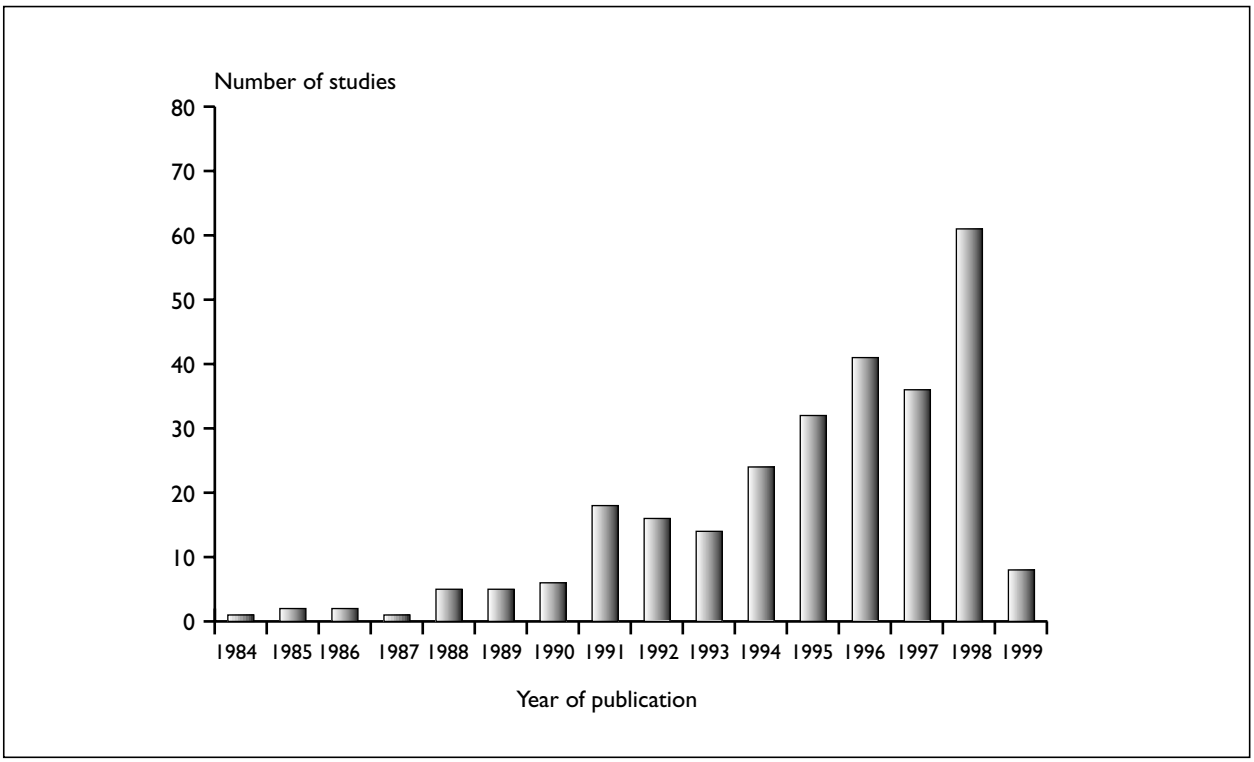


FIGURE 2 *Year of publication of included studies*

TABLE 2 *Study characteristics*

Study characteristic	Number (%)
Type of procedure	
Laparoscopic	140 (51)
Other surgical	110 (41)
Diagnostic	22 (8)
Assessed on:	
Humans	262 (96)
Animals	6 (2)
Machines	4 (2)
Design	
Case series	259 (95)
Controlled, non-randomised	7 (3)
Randomised, controlled	6 (2)
Sample size	
0–50	58 (22)
51–200	114 (41)
201–800	59 (22)
> 800	41 (15)
Number of operators or institutions	
1 only	174 (64)
2–5	43 (16)
6–20	25 (9)
Over 20	30 (11)
Level of assessment	
Operators	128 (47)
Institutions	140 (52)
Operators and institutions	3 (1)
Not operators or institutions	1 (< 1)
Proportion of surgeons performing 50% of procedures	
All	141 (52)
1–10%	7 (3)
11–30%	16 (6)
31–50%	28 (10)
Unclear	80 (29)
Level of care of institutions	
Tertiary	98 (36)
Secondary	66 (25)
Mixed (tertiary and secondary care)	23 (8)
Unclear	85 (31)
Data source	
Prospective	108 (40)
Retrospective	71 (26)
Registry	10 (4)
Unclear	83 (31)
Prior knowledge of outcome	
Prior knowledge	71 (26)
No prior knowledge	101 (37)
Unclear	100 (37)
Type of outcome used to assess learning	
Intra-operative – continuous	122 (45)
Intra-operative – dichotomous (not rare)	138 (51)
Intra-operative – dichotomous (rare)	84 (31)
Intra-operative – categorical	2 (1)
Postoperative – continuous	38 (14)
Postoperative – dichotomous (not rare)	22 (8)
Postoperative – dichotomous (rare)	15 (6)
Postoperative – categorical	1 (< 1)

searching the abstract and title only (*Table 1*). However, these papers did not contain any further novel techniques. Of the 46 questionnaires sent to experts in the health technology assessment field, 35 (76%) were returned. No extra studies were identified but additional methods were suggested.

The double assessment of the 10% sample of possible studies showed perfect agreement on the inclusion of papers (kappa = 1) and very good agreement on the methods used (kappa = 0.81). All disagreements were due to a descriptive method being omitted when split group methods were also used. This high rate of agreement convinced us that double assessment of all papers was not justified. Handsearching of abstracts from the *International Journal of Technology Assessment in Health Care* identified one additional study but no new methods.

Included studies

Of the 272 included studies, 140 (51%) assessed a surgical minimal access technique such as laparoscopic cholecystectomy, hernia repair or fundoplication (*Table 2*). Other treatment procedures, such as heart transplantation, were assessed in 41% of studies and diagnostic technologies, such as interpretation of MRI scans, in 8%. Most of the techniques (96%) were performed on humans.

The majority of the studies (95%) were case series. Only 2% used data collected from an RCT. Study sizes varied considerably but about 40% were less than 100 (*Figure 3*). In 64%, the study addressed the learning curve for a single operator or a single institution only. Approximately half of the studies assessed learning only by individual operators. The remainder assessed learning at the level of the institution (or both).

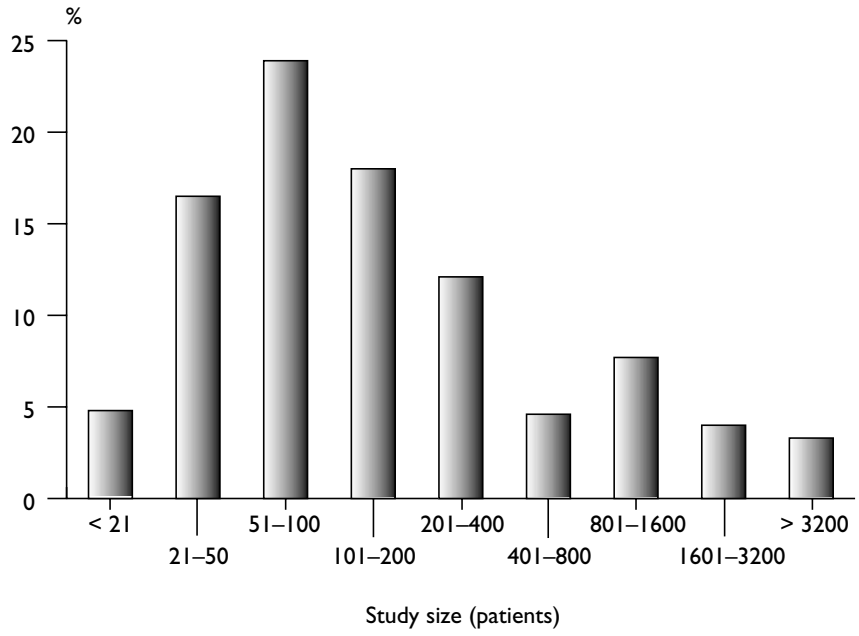


FIGURE 3 *Study sizes of included studies*

A few studies with more than one operator were dominated by a small number of operators; for example, fewer than 10% of the operators may have performed 50% of the procedures. However, this was unclear in 29% of the studies.

Most studies were performed in either tertiary or secondary care centres but the level of care was unclear for nearly a third. No studies were identified in primary care.

The data were collected prospectively in 40% of studies and retrospectively in 26% but in the remaining 34% this was unclear. The outcome was known before the analysis of learning began in 26% of the studies, was not known in 37%, and was unclear in 37%. In some studies, data were collected prospectively but only submitted to a registry or study after the outcome was known.

The types of outcome used to assess learning were mainly intra-operative continuous process variables (45%), such as operation time, and intra-operative dichotomous outcome variables (51%), such as complications. Rare intra-operative events were mentioned in 31% of studies and rare postoperative events in 6%.

The interrelationship between some of the study characteristics and the type of technology was also examined (*Table 3*). Minimal access studies were more likely to have fewer than 50 patients (29%) than studies of other surgical procedures (16%). The type of variables used to assess learning differed between minimal access and other surgical procedures. Minimal access studies more commonly used continuous outcomes than dichotomous outcomes (63% vs. 40%). This pattern was reversed in the other surgical studies in which continuous outcomes were less commonly used (30% vs. 59%). The majority of diagnostic studies used dichotomous outcomes.

Statistical methods used

The statistical methods for assessing learning in the different groups of technologies are described in *Table 4*.

Descriptive

In 120 (44%) studies, the data were displayed graphically as a plot of outcome against experience or as a table reporting when results or complications occurred within the series (*Table 4*). In all studies in this group, this was done without statistical analysis.

Split groups

The most common statistical method, used in 165 (60%) studies, involved splitting the data into groups by 'experience'. This was usually performed

TABLE 3 *Selected study characteristics by type of technology*

Study characteristic	Minimal access: number (%)	Other surgical: number (%)	Diagnostic: number (%)
Sample size			
0–50	40 (29)	16 (16)	2 (9)
51–200	46 (33)	57 (51)	11 (50)
201–800	32 (23)	21 (19)	6 (28)
> 800	22 (16)	16 (14)	3 (13)
Design			
Case series	131 (94)	106 (96)	22 (100)
Controlled	5 (3)	2 (2)	0 (0)
Randomised controlled	4 (3)	2 (2)	0 (0)
Type of outcome used to assess learning			
Intra-operative – continuous	88 (63)	33 (30)	1 (4)
Intra-operative – dichotomous (not rare)	56 (40)	65 (59)	17 (77)
Intra-operative – dichotomous (rare)	50 (36)	34 (31)	0 (0)
Intra-operative – categorical	1 (1)	1 (1)	0 (0)
Postoperative – continuous	21 (15)	17 (15)	0 (0)
Postoperative – dichotomous (not rare)	4 (3)	13 (12)	5 (23)
Postoperative – dichotomous (rare)	6 (4)	7 (6)	2 (9)
Postoperative – categorical	0 (0)	1 (1)	0 (0)
Statistical method			
Descriptive	74 (53)	40 (37)	5 (23)
Split group	71 (51)	74 (67)	20 (91)
Univariate (trend)	19 (14)	14 (13)	0 (0)
Multivariate (split)	3 (2)	1 (1)	0 (0)
Multivariate (trend)	4 (3)	2 (2)	0 (0)
Cusum	5 (4)	1 (1)	0 (0)

TABLE 4 *Statistical methods used in included studies*

Statistical method	Number (%)*
Descriptive	120 (44)
Split groups (no test for trend)	165 (60)
Univariate (trend)	33 (12)
Curve fitting	25 (9)
χ^2 test for trend	2 (1)
Pearson correlation	2 (1)
Repeated measures ANOVA	3 (1)
Komolgorov–Smirnoff	1 (< 1)
Multivariate (split – experience dichotomised)	4 (1)
Logistic regression	3 (1)
Cox's regression	1 (< 1)
Multivariate (trend – experience continuous)	6 (2)
Multiple regression	2 (1)
Logistic regression	3 (1)
Generalised linear mixed models	1 (< 1)
Cusum	6 (2)

* *Can exceed 100% in total as some studies used more than one method*

by arbitrary splitting of a series of consecutive cases for individual operators into halves or thirds. The means of the two or three groups were then compared by test or ANOVA. If these means differed, the authors assumed that learning had taken place. Alternatively, a chi-squared test was used for dichotomous outcomes such as complication rates. Eight (5%) of the 'splitting' studies compared mean operation time between 'experienced' operators and 'inexperienced' operators, to test whether the extent of learning differed between the groups.

Split group methods were used less in minimal access studies (51%) than in other surgical studies (67%) (*Table 3*). Similarly, descriptive methods were used more in minimal access studies (53%) than in other surgical studies (37%). The split group method was generally used for assessing the learning curve in diagnostic studies.

Univariate trend

A more sophisticated approach, used in 25 (9%) studies, was that of fitting a line to the data by least squares regression (*Table 4*). A linear relationship between experience and outcome was most commonly described.[20–31] A variety of other curves were used to describe the learning relationship: logarithmic,[32–35] negative exponential,[36,37] double-negative exponential,[38] power form,[39] reciprocal,[40] quadratic[41,42] and cubic.[43] In addition to the 24 papers using least squares regression, Monte Carlo simulation was used in another paper to estimate the shape of the learning curve.[44]

The correlation between experience and outcome was tested by Spearman's correlation coefficient,[45,46] chi-squared test for trend,[47,48] or a Komolgorov–Smirnoff test.[49] In three studies an attempt was made to model the relationship between experience and outcome using repeated measures ANOVA.[50–52]

Multivariate (split group)

A number of multivariate techniques were used. Logistic regression was used in three studies to test whether there was a relationship between a dichotomous outcome and experience.[53–55] In another study, Cox's regression was used to look for a learning effect in time-to-event data.[56] In all these studies, adjustments were made for other confounding factors such as age or sex but the experience variable was split arbitrarily into equal categories.

Multivariate (trend)

The experience variable was continuous in the remaining papers. Logistic regression was used in three papers[57–59] and multiple regression in two[52,60] to adjust for confounding factors before testing for a relationship between experience and operation time. Generalised linear mixed models were used once.[61] From 1996 onwards, multivariate techniques of both types were increasingly reported.

Cusum

A cusum technique was used in six studies.[40,44,62–65]

Discussion

Systematic search strategy

The aim was to describe the 'epidemiology' of the statistical methods that investigators have used to assess learning in health technology assessment. To avoid bias, a systematic approach was used to identify relevant studies and to extract data. To make the task manageable, the search strategy was kept sufficiently specific to avoid highlighting a large number of irrelevant papers. Even limiting this strategy to searching for 'learning curve' produced nearly 5000 abstracts requiring assessment. Exploratory searching using other search terms made it clear that the dividend was not worth the resources required. Full text searching in MEDLINE allowed the assessment of how many relevant articles might have been missed as a consequence of basing the search on titles and abstracts only. In the event, this did identify seven (3%) studies that would otherwise have been missed; however, these studies did not use any statistical method that had not been identified elsewhere.

Further searches for other statistical techniques were undertaken after performing the review of the included studies. After identifying additional statistical methods known to us that could have been used to assess learning in the clinical field, a new strategy was created to search for these and the abstracts generated were assessed for relevance to learning curves. No evidence was found indicating that any of these had been used for this purpose, so it is considered unlikely that an important technique has been missed.

Approximately three-quarters of the included studies were identified on MEDLINE. This reflects the ordering of searches as MEDLINE was searched first. As 70 studies were identified only in databases other than MEDLINE, this confirmed the importance of broader searching.

Use of the term 'learning curve' increased over time. During the early 1980s, the term was rare

and mainly related to organ transplantation. The increase since the late 1980s coincided with the introduction of minimally invasive procedures, especially laparoscopic cholecystectomy. However, some 40% of included studies related to other surgical procedures. Only 8% of included studies related to diagnostic methods.

Proxies for learning

Two types of variable were used to assess learning in the 272 studies – measures of patient outcome or quality assurance and measures of clinical process or task efficiency.[18] Unfortunately, the patient outcomes used, although acceptable proxies for the goals of healthcare, tend to be dichotomous rare events like complications or survival, which are relatively intractable to statistical analysis. This may be why, in most studies, continuous process measures were chosen; typical examples in surgery were the time to complete an operation and the period of hospitalisation. In minimal access surgery, operation time was more commonly used to assess learning than in other surgical procedures. Although operation time is relatively easy to collect, it is only a weak proxy for learning and does not necessarily relate to proficiency.[63,66] As Darzi and colleagues[66] pointed out, "measuring competence merely by setting time targets for certain procedures is crude and probably unacceptable". Other proxies have been suggested, such as movement of instruments[66] or 'near misses',[67] but these too are probably weak proxies for patient outcome and, hence, learning.

Statistical methods used

This review confirmed that the statistical methods used to assess learning in health technology assessment have almost always been crude. A substantial number of studies (44%) have relied upon descriptive data to claim learning without any formal statistical testing.

The most common formal approach was the split group method. Often papers gave no rationale for the cut points, raising concerns about bias caused by data-dependent splitting. Arbitrarily splitting the data into halves was not uncommon. Yet it takes a minimum of three points before a trend can be characterised. Even when splitting suggests that learning has occurred, it is not possible to describe the underlying curve or to identify where on that curve particular operators lie.

A univariate test for trend using curve-fitting procedures was the most commonly used of the more advanced techniques. Papers used a variety of different shapes but rarely gave a rationale for that selected. A linear relationship was often described but this could reflect the fact that the series was too short and the operators had not yet reached their final asymptote or plateau.

Multivariate techniques that adjust for a drift in case-mix are more robust and potentially useful for investigating trends over time. Unfortunately, in the studies identified the potential of these methods had not been maximised. Firstly, some studies dichotomised the experience variable and, thus, have limitations similar to the split group studies. Secondly, few studies have attempted to model inter-operator differences.

The cusum technique has been advocated as a method for monitoring surgical performance.[67,68] The technique can be useful for identifying when an operator begins to perform poorly but it is not so effective for describing inter-operator differences. The appropriate use of this method within a health technology assessment based on an RCT remains unclear but is useful for exploratory analysis.

The assessment of learning curves in diagnostic technologies was not the primary aim of our study but operators have been compared through receiver-operator characteristic curves.[69]

Individual or institutional learning

The included studies generally considered learning only within an individual operator or institution. While this approach is useful in looking for learning curves, it suffers from three inherent weaknesses. First, since there is no comparison with other operators or institutions, it is difficult to assess where an operator is on the learning curve. Second, it is difficult to use rare complications to assess whether there is a relationship between experience and complication rate for one operator. Finally, these problems are aggravated by the tendency of single-operator studies to rely on retrospective data collection from medical records, raising concerns about the danger of biased abstraction.

It is therefore desirable to obtain prospective data on many operators or institutions. In particular, the creation of data registries for specific technologies could provide a resource for assessing learning curves. To avoid bias, such registries should be prospective and outcomes should not be collected before a patient is registered. However, the continual updating, disseminating and funding of such registries is difficult.[70]

Learning by individuals and by institutions are inextricably linked. Institutional learning adapts processes like those governing referral, patient selection and aftercare to the circumstances of the new technology. At the same time, individual operators refine their skills in performing the procedure. Any statistical analysis of learning curves should account for this inherent hierarchy.

RCTs

Nearly all the included studies were case series, only five (2%) being RCTs. However, this could reflect the search strategy. Assessing the learning curve seems more likely to be presented as only a small part of the analysis of a clinical trial and, hence, is less likely to be mentioned in the abstract. The Cochrane Register of Controlled Trials was therefore searched in an attempt to find more RCTs that assessed learning; however, none were found.

Initial patients receiving a particular technology tend to be either relatively more fit or relatively more sick than those for whom it is later judged to be appropriate.[7,18] Within a randomised comparison, such a drift in case-mix will apply to both groups equally and can be taken into account during analysis.

There are strong arguments that assessment of non-pharmacological technologies should include a pragmatic randomised trial and that this should start as soon as feasible.[12] Nevertheless, it is recognised that this will not be the only element of assessment. Many assessments will include a pre-randomisation phase of observational data collection as the technique is developed or refined. Thus, methods for evaluating learning in these studies are also needed. Such methods would also help to decide whether and when an operator has reached a particular level of competence, and to monitor subsequent performance.

Conclusions

Implications for designing studies

This review has implications for the design of studies, including RCTs. The experience of operators should be collected during a study. If this is done, the investigators can look for trends over time throughout the study. As it is unlikely that every patient having a new procedure will be included in a randomised trial, it would also be important to record the number of procedures performed between randomised patients.

Implications for reporting primary studies

Completed studies need better reporting of the key factors that may be related to the learning curve. As a minimum standard, the number and experience of the operators, the data source and the level of care should be explicitly mentioned. This review has shown poor reporting of these factors, causing problems with interpretation and generalisation. In particular, an unreported data source implies that the validity of the study should be viewed with caution. The level of care could also affect the learning curve; for example, the reporting of results from a tertiary care institution might not be generalisable to secondary care. Finally, to be confident that the aggregated results of a multi-operator study were not influenced by a single operator performing most of the procedures, the proportion of surgeons performing half of the procedures is important.

The difficulties of assessing health technologies with learning curves could be better addressed if rigorous statistical methods were available for measuring and, hence, adjusting for learning. Randomisation could begin as soon as possible consistent with safety and the completion of basic training,[12,14] and continue until well after the learning curve has stabilised. The subsequent analysis would estimate both the point at which the learning curve stabilised and the level of performance achieved (both to within a confidence interval (CI)). These two estimates would lead to two distinct but complementary evaluations. The first evaluation would focus on the benefits and costs of introducing the new technology; the second on the benefits and costs of the new technology in steady state. While the second evaluation would play the major role in deciding where and when a new technology should be adopted, the first would influence how it should be introduced and what additional training and precautions were needed.

Implications for further research

This first component of this study has shown that currently used statistical methods are not sufficiently rigorous. There is a need for methods that can estimate the rate and length of learning, together with the final skill level. These should also be capable of exploring and estimating differences between individual operators. In the next chapter the search for such techniques is described, notably in fields where learning effects have been of concern, such as psychology and engineering.

Chapter 3

A systematic identification of methods for assessing the learning curve in non-clinical literature

Introduction

The first phase of the project, described in chapter 2, showed that the reporting of studies describing the learning curve effect in health technologies could be improved and that the statistical methods used had been sub-optimal. A clear need for more sophisticated statistical methods for describing learning curves was identified.

Learning curve effects have long been described outside health technology assessment and this is why the search for statistical methods was moved outside the health field. As early as 1936, Wright[71] observed that the manufacture of aircraft followed a learning curve – that is, as more aircraft were built, the amount of human labour (and hence cost) to produce each aircraft reduced in a curve-like relationship. He used this relationship to predict the level of labour requirements in the future. The effect was attributed to the skill of the manufacturers increasing over time. The term 'learning curve' became part of everyday language from the 1960s as many psychological studies of human skilled performance were carried out. Fitts and Posner[72] summarised the large amount of literature characterising quantitative and qualitative changes in skill that occur with practice. Much of this work was in finding the correct slope of the learning curve; for example, a number of studies have suggested that learning follows a power model of learning[71–74] or some variant of it.[75] Commonly, only simple perceptual-motor tasks were studied and subsequent research in applied psychology and human factors has sought to apply it to more complex tasks.[73]

The searches were extended into these other areas (such as psychology and engineering) with the aim of finding 'novel' techniques that had not been used in health technology assessment. These techniques were either those that had not been identified in the search of the clinical literature or those that improved an existing technique. The novel techniques also had to model or summarise data from learning curves (performance changes). The results of this search are described here.

The search was for methods that would allow estimation of rate and length of learning, provide a proxy for final skill level and allow individual operator differences to be explored and estimated. Of particular interest were ways of describing the learning curve effect using binary events (for example, presence or absence of complications), especially for the case in which the events are rare.

Methods

Systematic electronic bibliographic database searching

First, clinical databases were explored for further non-health technology assessment use of methods for assessing learning. The search was then extended into other fields, with advice on the most important databases being taken from experts in these fields. The searches were incremental, in the sense that the product of each search is the extra dividend obtained from that database after excluding duplicates found previously.

To optimise the return on resources available, a specific (i.e. focused) search was developed for each database. The most specific search terms were chosen and are shown in appendix 2. Details of other less specific terms tested, but rejected because they retrieved too many irrelevant studies, are also given in appendix 1. There were no language or other limits applied to the search strategy. Search terms describing complex statistical techniques and terms searching for binary events were also tested and are presented in appendix 2.

Systematic searches were made of 22 electronic databases (see *Table 5*).

TABLE 5 *Summary of the literature searches*

Database[†]	Field	Number of abstracts	Possibly relevant papers[a]	Relevant papers
Clinical databases[b,c]				
MEDLINE	*Index Medicus*	736	5	3
EMBASE	*Excerpta Medica*	588	5	1
CINAHL	Nursing and allied	28	0	0
HealthSTAR	Health research	21	0	0
ISI Science Citation Index	Science	1235	9	0
BIOSIS	Biology[d]	629	0	0
Clinical total		**3237**	**19**	**4**
Non-clinical databases[c]				
RSC	Chemistry	13	0	0
ISI Social Science Citation Index	Social sciences	352	17	4
ISI Arts and Humanities Citation Index	Arts/humanities	10	0	0
PsycLIT	Psychology	242	17	3
IBSS	Economics	26	0	0
ISTP	Scientific conference proceedings	67	0	0
Ei Compendex/Page One	Engineering	346	3	0
SOCIOFILE	Sociology	11	0	0
ABI/INFORM	Business	562	2	0
ECONLIT	Economics	50	0	0
CAB abstracts	Agriculture[e]	17	0	0
INSPEC	Engineering	255	0	0
IngentaJournals Online	Many topics	14	0	0
Index to Theses (GB & Ireland)	Theses	8	0	0
Dissertation Abstracts	Theses	147	0	0
NASA Technical Reports Server	Space/aviation sciences	353	0	0
Non-clinical total		**2473**	**39**	**7**
Other sources				
Other terms tested in electronic databases[f]		3375[g]	32[h]	1[i]
Experts in the field		61[b]	8	4
Reference lists (of relevant papers + other lists)		21	2	2
Citation indices		264[j]	15	0
Total other sources		**3721**	**57**	**7**
Total		**9431**	**115**	**18**

[†] *Details of electronic databases are presented in appendix 3*

[a] *Full papers assessed if technique not previously used in health technology assessment*
[b] *Excludes health technology assessment*
[c] *Search term: learning curve**
[d] *Included clinical and experimental medicine*
[e] *Also includes forestry, animal health and environmental sciences*
[f] *Known methods; binary terms; skill acquisition; learning effect; slips & mistakes; other terms*
[g] *Known methods, 709; binary terms, 813; skill acquisition, 503; learning effect, 266; slips & mistakes, 245; other terms, 839*
[h] *Known methods, 19; skill acquisition, 6; binary terms, 5; other terms, 2*
[i] *From curve analysis*
[j] *Number of key papers used = 13*

- ***Clinical databases*** Many of these also include journals from other fields and specialities. For example, MEDLINE includes a number of psychology journals. Six clinical databases were searched.
- ***Non-clinical databases*** In all, 16 non-clinical databases were searched for a 10-year period from 1989, unless otherwise stated.

A description of each database is given in appendix 3.

Full text electronic databases searched systematically

As in the first phase of the project, the full texts of selected journal articles were electronically searched to estimate the number of articles that

described the assessment of the learning curve in the body of the report – a finding that would not have been identified by searching the abstract and title only. To do this, Ingentajournals database was used; this provides access to full text versions of a range of academic journals produced by BIDS (over 550 journals covering economics, engineering, mathematics, psychology and other subjects) on BIDS online. The search strategy used is presented in appendix 2.

Other methods of ascertainment of articles

The reference lists of included articles and other articles of interest that had used novel statistical techniques were followed-up to identify further relevant reports. Citation indices were used to track the subsequent citation of relevant reports. A number of experts in the field were contacted also and asked to identify any other relevant novel techniques.

Handsearching was considered but no journals were identified in which this was likely to lead to a substantial dividend in terms of extra relevant articles identified. This reflected the spread of the relevant literature across many fields and journals.

Identification of possibly relevant articles

References to all possibly relevant articles were electronically imported or manually entered into the software package Reference Manager (v. 9.0 N: ISI ResearchSoft, USA). Details of the source of articles were added. To assess subject relevance, all electronically derived abstracts and titles were read by one statistician. An abstract was deemed possibly relevant if the study was not a health technology assessment, the body of the abstract referred to the statistical modelling of a learning curve, and the statistical method used had not previously been identified. A subject expert assessed full copies of articles for subject relevance, eligibility and methodological quality. The assessor was not blinded to author, institution or journal.

Inclusion criteria

To be included in the review, a study had to assess a learning curve formally using a novel statistical technique. These were categorised into four subgroups of increasing complexity.

- ***Exploratory data analysis***: techniques that do not estimate statistical parameters or test hypotheses. These included graphical displays of data or the creation of cusums of consecutive cases.[19]
- ***Simple series data analysis***: techniques that use data collected on a single operator or summarised over many operators. For example: (a) a *t*-test that compared the mean operation time for the first 50 procedures with the mean time for the next 50; (b) a study of ten operators performing 20 procedures each, in which the data were analysed as the average performance of the ten operators; (c) fitting the best shape of curve to each individual operator's performance and describing the various shapes that these curves took.
- ***Complex data structure analysis***: techniques that used data collected on many individuals and measured both differences between individuals and the overall pattern of learning.
- ***Generic techniques***: techniques that could be applied to both simple series and complex data structures.

Double assessment

A second statistician independently assessed each included study. Any differences of opinion were resolved by discussion.

Results

Literature search

The numbers of possibly relevant abstracts generated by the systematic searches are shown in *Table 5*. Of 9431 abstracts assessed, 115 (1%) were deemed appropriate for further investigation. Of these, 18 were later judged on review of the full paper to describe a novel technique or to make a significant addition to a previously recognised technique. The dividends from each of the searches are also shown in *Table 5*. Of the 18 included studies, four were identified in Social Science Citation Index and three each from PsycLIT and MEDLINE. A complete list of all 18 studies is given in the references to all included studies (see page 80).

Of the 61 questionnaires sent to experts in their fields, 50 (82%) were returned. Four additional methodologies were suggested. The reference lists of the selected papers produced two other novel techniques.

The most useful single term in the electronic searching was 'learning curve*' across a broad spectrum of databases, identifying 58 possibly relevant articles of which 11 were actually useful. Of the other terms tested, skill* + (acquir* or

acquisit*) appeared the most promising with six potentially relevant papers; however, on assessment of the full papers, none were relevant and this term was therefore abandoned. Other specific terms were also tested: those related to known methods and those describing binary type data in some of the major electronic databases in other fields. These produced very few possibly relevant abstracts of papers and none of these proved relevant; these terms were therefore not pursued further. No studies were identified in which an attempt was made to model rare, binary events in a learning curve.

Included techniques

The 18 included papers were categorised in two ways: shapes of learning curves (*Table 6*) and statistical techniques (*Table 7*). In *Table 6* the curves are presented that were identified from the health technology literature on learning curves (top), together with the additional eight shapes identified from the non-health technology assessment search (bottom). All of the newly identified curves had a similar basic shape that decreased to an asymptote. The additional curves came from psychology,[72,74,76,77] manufacturing[78,79] and aviation.[80] The most widely cited shape of learning curve across all fields was the power law (of practice). All of the shapes presented in *Table 6* may be assessed using the 'curve fitting' techniques that are available in most statistical software packages.

The statistical techniques that were previously identified in the health technology assessment literature and the dividend from searching the non-health technology assessment literature are presented in *Table 7*. A number of techniques were identified for exploratory data analysis and simple series data in the health technology assessment literature but none were identified that analysed learning curve effects using

TABLE 6 *Shapes of learning curves*

Type of curve	Equation[§]
Curves previously used in health technology assessments	
Linear	$y = a + bX$
Quadratic	$y = a + bX + cX^2$
Cubic	$y = a + bX + cX^2 + dX^3$
Power law (asymptote at zero)	$y = bX^{-c}$
Reciprocal	$y = a + \frac{b}{X}$
Exponential	$y = ae^{-bX}$
Double exponential	$y = ae^{-bX} + ce^{-dX}$
Logarithmic	$y = a \ln X + b$
Potentially useful curves identified outside health technology assessments	
Power law (with non zero asymptote)	$y = a + bX^{-c}$
Log-log–linear model	$\ln y = a (\ln (X + 1))^b$
Log–linear curve	$\ln y = aX^b$
Logistic curve	$y = \frac{a}{1 + be^{cX}}$
Weibull curve	$y = a - be^{-cX^d}$
Exponential difference equation	$\xi (X) = a - [a - \xi (X - 1)] e^{-b}$
Cumulative performance curves	$y = \frac{k}{X} \sum_{X=1}^{n} X^{-s}$
Exponential cumulative performance curves	$y = \frac{(a^n - 1)(b - c) + c}{(a - 1) n}$

[§] *y denotes the continuous outcome (for example, 'time') and X denotes the case sequence number (for example, X = 1 is the first procedure, X = 2 the second, and so on)*

TABLE 7 *Techniques used to detect the learning curve*

Techniques previously used in health technology assessment
Exploratory data analysis Graphical Cusum techniques
Techniques for simple series data *t*-test, one-way ANOVA χ^2 test (for trend) Repeated measures ANOVA* Curve fitting Multiple regression* Logistic regression*
Techniques for complex structured data None identified
Techniques that can be applied to both simple and complex data types None identified
Potentially useful techniques identified outside health technology assessment
Exploratory data analysis None identified
Techniques for simple series data Curve fitting
Techniques for complex structured data Discriminant analysis* Two-stage estimation of learning rates Generalised estimating equations Multilevel models Latent curve models Time series models (ARIMA) Stochastic parameter models
Techniques that can be applied to both simple and complex data types Generalised linear models
* *These techniques are special cases of generalised linear models*

complex structured data techniques. In contrast, the non-health technology assessment studies described a number of potentially useful techniques for assessing learning curves using complex structured data that had not been identified in the health technology assessment literature. A brief outline follows of the more complex methods and the studies in which they were used.

Discriminant analysis

One study measured active avoidance learning in 24 rats with streptozotocin diabetes and compared these with 27 control rats over 100 consecutive trials.[81] The average final success rate (asymptote), latency (the period from initial point to 50% success level) and rate (number of trials to reach asymptote) were measured. These measurements were then used in a discriminant analysis to search for subgroups of rats with similar learning characteristics.

Two-stage estimation of learning rates

This study involved 115 students performing a simulated air traffic control task on 18 consecutive trials.[76] The simulation had three task components: accepting aircraft into the airspace, moving aircraft in a three-level hold pattern, and landing aircraft on appropriate runways. The number of correct landings per trial was the outcome of interest and individual learning curves were estimated by two-stage estimation using a negative

exponential curve. This is a relatively simple procedure that can be applied using most standard statistical packages.

Generalised estimating equations

Generalised estimating equations allow the correlation of outcomes within an individual to be estimated and accounted for in any subsequent analyses.[82,83] This procedure is similar to the two-stage estimation procedure above but is strengthened by the fact that the estimates are obtained iteratively, not just in two stages.

Multilevel models

This method has been increasingly used in clinical research[84,85] and involves partitioning the variability between and within hierarchies in the data set (learning curve data have an inherent hierarchy – individuals performing multiple procedures). Multilevel models could have been applied to the air traffic control task in the earlier study.[76] In that case, the students would have been the highest hierarchy (level) and the 18 trials recorded for each student would be the lower level units.

Latent curve/stochastic parameter/ ARIMA models

Both latent curve models and stochastic parameter models are specialised cases of structural equation models.[86] In both cases, factors are calculated for each individual's asymptote, total growth (increase from first to last trial) and rate of growth. The complex structured data techniques identified are completed by time series modelling using autoregressive integrated moving averages (ARIMA) models. A comparison of these three methods was made using data from a study that measured 137 US Air Force personnel performing an air traffic control task on six occasions.[87] Again, the number of correct landings per trial was the outcome of interest.

Generalised linear models

Generalised linear models were included as a method that could be applied to both complex and simple structured data. Multiple and logistic regression are used in generalised linear models that have been used in the clinical literature to assess learning curves of individual operators. They have not, however, been used to assess differences between operators (although they could have been).

Other included papers

In addition to the papers included so far, two other papers were added. The first was a paper by Baloff and Becker[88] that discussed the danger of using an aggregated learning curve. The authors described how the aggregation of curves for many families results in curves that are different from the individual curves being aggregated. For example, the individual curves could be step functions but, when aggregated, they yield a continuous group curve (exponential or sigmoid) that does not describe any of the individual functions. This problem is not restricted to step functions. Hayes[89] demonstrated that aggregating convex exponential functions can result in a concave exponential function and that aggregating exponential functions can yield a sigmoidal function. Sidman[90] has also shown that aggregating linear learning curves with different learning rate parameters can, in theory, yield an exponential function. Estes summed up the problem:

> "…the uncritical use of mean curves even for such purposes as determining the effect of an experimental treatment upon rate of learning … is attended by considerable risk."[91]

The final paper included was by Delaney and colleagues[92] and described a study in which the aggregated operator learning curve was substantially different from the individual operator's learning curves.

Discussion

The primary aim of this phase of the project was to search the non-health technology assessment literature to identify novel statistical techniques that could be used to assess the learning curve effect in health technology assessment. A range of potential sources was searched systematically. An incremental approach was used, based on formal assessment of full papers whose abstracts suggested that the statistical techniques used had not been identified previously.

The resources required for the searches were considerable. Even limiting searches to simple search terms identified approximately 10,000 abstracts that required assessment. The apparently poor cross-fertilisation of techniques between fields meant that there was concern about missing useful methods. An alternative search strategy would have been to select a few key databases and to explore thoroughly the keywording in these . A decision was made against this and in favour of covering a broad range of non-health technology assessment fields. Nevertheless, even

when no relevant papers were identified from a database it may not mean that it contained nothing of relevance, but rather that the searches were too limited. The authors were also aware of other databases that might have been useful but had insufficient resources to assess them fully (for example, ASSIA PLUS, Ergonomic Abstracts and the educational database, ERIC).

In all, 18 papers were identified in which a novel technique was used. Because an incremental approach was used, the order of searching each database influenced the number of relevant papers identified. For example, some of the relevant papers in PsycLIT were also identified in the NASA [National Aeronautics and Space Administration] database. Contacting experts in the field proved one of the most fruitful sources of relevant studies. Four of the 18 relevant papers were identified by this method. It is recommended that reviews of methods used in other fields should include this approach.

The novel techniques that were found were categorised into shapes of learning curves and techniques for modelling learning curves. This reflects an important distinction between methods for **identifying** a learning effect and those for **measuring** (characterising) a learning effect. Fitting a particular shape of learning curve to the data can be used initially to assess if there is a trend over time, whereas the comparison of many different shapes measures the learning effect; for example, best estimates for the asymptote or rate of learning can be obtained. The more advanced modelling techniques then use the 'best' shape of individual curves to explain observed differences between operators.

An extra eight shapes of curves were identified. In contrast to the health technology assessment literature, the non-health technology assessment literature tended to give rational explanations for the shape of curves used; for example, Newell and Rosenbloom[74] give a 'chunking' theory of learning to explain a power-curve type of relationship. The health technology assessment literature tended to describe fitting a curve without explanation for the choice of shape or comparison of different shapes. It is recommended that any analyses of health technologies using simple series data should consider the likely shape of the learning curve *a priori*, bearing in mind that there are a number of different types of learning curves. The type of technology, for example, may influence the shape of the curve.

The ability to describe a learning curve is often limited by the type of data collected. For example, to assess the effect of learning rigorously will often require data from more than one person. That data were often limited to a single operator might have been the explanation for the relatively simple statistical techniques employed in the health technology assessment literature. However, the systematic review of this literature found that 35% of studies had more than one operator but used an analysis of the mean effect. It is well recognised that the shape of an individual learning curve is not necessarily the same as the average shape of many individuals.[88] The use of such averaging should therefore be applied with caution. Another limitation is the availability of software to perform many of the methods. The more sophisticated the method becomes, the more difficult it is to find a standard statistical package that can perform it. For example, multilevel models are available through MLWin software[93] and the SAS PROC MIXED procedure.

Nine novel statistical techniques for measuring learning curves were identified that had not been found in the health technology assessment field. The non-health technology assessment fields used methods that were more sophisticated and this was reflected in the hierarchy of methods that were identified. The techniques used for complex structured data were only identified in the non-health technology assessment literature. There are clear advantages in using these – a measure of how an individual operator performs is obtained, together with a measure of how the operator is performing in relation to other operators in the study. This enables investigators to explore the influence of each operator in a study. These methods also use more data and are statistically more powerful.

Thus, there is a hierarchy of statistical methods that can be used to analyse learning curves. It is not advocated that all learning curve analyses should use the most complex methods, rather that they should employ the simplest method that can answer the questions being posed. The methods used should be parsimonious; that is, they should not use more parameters than are necessary. For example, if a Weibull curve and exponential curve produced similar results, then the exponential curve would be preferred since it only requires two parameters to be estimated rather than the four required by a Weibull curve.

Conclusions

Implications for literature searches addressing methodological problems

This study shows that when investigating a methodological problem in clinical evaluation, searches should not be confined to the clinical field. Many methods begin in fields outside healthcare. The task can expand very quickly, however, so it is suggested that any searches should be kept specific in the first instance. When searching non-clinical fields, it is very helpful to get an expert from that field to identify appropriate sources.

Implications for simple series data analyses

In reality, the shape of an individual operator's learning curve is unlikely to follow precisely the curves that have been identified. There is often a drift in case-mix as the operator becomes more experienced; for example, the operator may perform the procedure on increasingly difficult (or easier) cases. This can exaggerate or conceal the learning curve. Reliable and robust statistical techniques for adjusting clinical outcomes for case-mix effects are required and it is suggested that analyses of single operators adjust for these factors, for example, using logistic or multiple regression. These techniques may not work for binary outcomes if they are rare. No statistical methods were identified that modelled them in learning curves.

Implications for data collection in clinical trials

The reviews of published research have shown that the effects of learning curves on the outcome and interpretation of clinical trials are not clear at present. Statistically adjusting or investigating individual operators' experience in clinical trials would be a first step to better understanding. This would require systematic data collection of factors known to influence the learning curve. Such factors include the number and order of procedures that an operator performs, the number of procedures performed before entering any patients into the trial, and an approximate estimate of the number of procedures performed between randomised patients (as not all procedures are necessarily randomised). Without data on the individual operators during the trial, it is not possible to apply complex structured data techniques.

Implications for conduct of clinical trials

In the past, there has been controversy over the timing of the assessment of those technologies that need some form of learning before they are at their most effective or efficient. Early assessment is open to the criticism that a new technology is at a disadvantage because operators are not fully proficient. Late assessment runs the risk that operators will draw their own conclusions about the advantages and disadvantages of a new technology.[10]

In the authors' view, the portfolio of statistical methods described in this chapter opens the door to another approach. Randomisation could begin as soon as possible, consistent with safety and the completion of basic training, and then continue until well after the learning curve has stabilised. The subsequent analysis would estimate both the point at which the learning curve stabilised and the level of performance achieved (both to within a CI). These two estimates would lead to two distinct but complementary evaluations. The first evaluation would focus on the benefits and costs of introducing the new technology; the second on the benefits and costs of the new technology in steady state. While the second would play the major role in deciding where and when the new technology should be adopted, the first would influence how it should be introduced and what additional training and precautions are needed.

Implications for further research

A number of more sophisticated statistical methods have been found that could be used to model the learning curve effect during health technology assessments. The relative performance of these methods requires assessment before general recommendations can be made. Some of these methods are explored empirically in the next three chapters, using health technology assessment data sets known to exhibit learning effects. All of the data sets used are related to laparoscopic surgery (fundoplication, cholecystectomy and groin hernia repair). This choice was pragmatic, in the sense that there had been a documented learning curve effect for these technologies. However, it should be borne in mind that it is not known whether these results will be transferable to non-laparoscopic technologies.

Chapter 4

Single case series data – a case study of laparoscopic fundoplication

Introduction

Some of the statistical methods identified in the previous chapters for simple series data are illustrated and explored here, using a case series of 190 consecutive laparoscopic fundoplication procedures performed by a single surgeon. In particular, the aims were:

- to compare statistical methods for identifying learning curve effects when using continuous measures of performance
- to compare the various possible shapes of the learning curves identified through the systematic searches described in chapters 2 and 3
- to explore statistical methods for identifying learning curve effects when using dichotomous measures of performance.

Case series designs

The majority (95%) of studies in which the learning curve issue in health technologies were assessed are case series (see chapter 2). The limitations of this design (i.e. the fact that there is no control group or a potentially biased one) are less important in the context of learning curves, where the interest is in changes over time in a technology rather than in comparisons with an alternative form of care. Furthermore, case series do have advantages. Recruitment is relatively easy and all those treated using the new technology can be included in the study. The main concern is that changes over time, in the types of people on whom the procedure is used, may hide or distort any learning effect. However, as discussed further below, changes in case-mix can be monitored and the analyses subsequently adjusted accordingly for any known prognostic variables.

As discussed in the earlier chapters, the statistical methods that can be applied to learning curve data are primarily influenced by three factors. These are: the size of the data set, its complexity (single or multiple operators), and the type of variable (continuous or dichotomous) that is being used as the proxy for learning. In chapter 3, a number of statistical techniques available for investigating learning curves in single series data were described. These ranged from methods for identifying learning curves (such as graphical techniques) to methods for measuring the magnitude of the effect (such as curve fitting), and it was argued that the strengths and weaknesses of the different methods required further investigation.

Laparoscopic fundoplication

From the early 1960s, Nissen fundoplication has been the surgical technique used most frequently in the management of patients with gastro-oesophageal reflux disease (GORD).[94] In this operation, part of the stomach is wrapped around the lower end of the oesophagus. It is a traumatic operation when conventional techniques are used because of the difficulties of surgical access. Developments in 'keyhole' surgical techniques have greatly reduced the surgical trauma associated with the procedure, and there has been a dramatic increase in laparoscopic fundoplication since the procedure was first reported in 1991 by Dallemagne and colleagues.[95] It is widely accepted, however, that the laparoscopic procedure is technically demanding. Reflecting this, studies have shown that performance does change over time: early operation times tend to take longer and technical difficulties are more common.[96–99]

Watson and colleagues,[96] for example, used data describing 280 laparoscopic fundoplications performed by 11 surgeons to show that the rates of complication, reoperation and conversion were higher for the first 20 patients in a series than for later patients. The surgeons were subsequently categorised as 'early' or 'late', depending on whether they were the first to use the procedure in their institution or whether they began doing the operation at least 18 months after the first laparoscopic fundoplication had been performed in their hospital. The results suggested that problems could be reduced by experienced supervision. There was no attempt to measure these relationships statistically, however. Estimates of the number of cases that were considered to constitute the learning phase ranged from 20 to 30 but no explanation was given of how these figures were derived.

Methods

The laparoscopic fundoplication data set

The first 190 patients of one consultant surgeon comprised the case series used in the analyses for this chapter. These patients underwent laparoscopic surgery for GORD in Aberdeen over a 5-year period between August 1993 and August 1998. The indications for surgery included failed medical therapy with either persisting symptoms of pain or volume regurgitation, intolerance of medication, or a patient's wish to avoid long-term medical therapy.

Information was collected prospectively on patient demographic characteristics and operative technical details. These data were supplemented by case-note review to extract information on preoperative symptoms, drug management and investigations. Endoscopic, radiological and physiological data were recorded, although not all patients had all the investigations. Patients were reviewed at 4–6 weeks postoperatively.

Continuous measures of performance

Exploratory techniques

Scatter diagram

The first exploratory method described in chapter 3 for addressing a learning curve in case series data was drawing a scatter diagram of the continuous proxy measure of learning against the sequence number (1, 2, 3, …, n) of the procedure. The sequence number is determined by ordering the cases by date of procedure, with the earliest procedure being allocated the number one, the next earliest number two, and so on until the final procedure is allocated n (the total number of cases in the series). The scatter diagram can be used as a visual aid to look for trends or patterns over time.

Moving averages

In practice, it can often be difficult to observe a pattern in serial data because of variability in the observations. In this situation, a simpler summary measure of the learning variable can be obtained by using moving averages of the series. A moving average of a series of data points is a method for accentuating important trends in the data, while playing down the random fluctuations that are invariably present in this type of serial data.[100] In other words, moving averages attempt to **smooth** the data. This procedure averages a number of successive observations centred on the time point of current interest. The **order** of the moving average is derived from the number of observations that are used to summarise the data from each point. Thus, a moving average of order 3 uses the current point of interest and the point on either side of the current point. This is illustrated in *Figure 4*.

In general, a moving average of n data points $\{y_t: t = 1, \ldots, n\}$ is another series, s_t say, defined by:

$$s_t = \sum_{i=-p}^{p} w_i y_{t+i}: t = p+1, \ldots, n-p \qquad \text{Equation 1}$$

where p is a positive integer and w_i are weights, with $\sum w_i = 1$. The order of the moving average is $2p + 1$. It should be noted that this formula gives no moving averages for observations at either end of the series. For example, in a moving average of order 3, no value is calculated for the first and last observations in the series. There are ways to extend the above formulation to take this into account[100] but, for most practical applications, there are enough data points in the series to make this unnecessary.

Data splitting

A commonly used approach to assessing learning within case series (see chapter 2) is to split the data into consecutive time blocks and calculate the mean and variability of each segment. Differences between groups can be tested using the Scheffe pairwise multiple comparison procedure[101] and a trend across groups can also be tested. As discussed in chapter 2, this approach gives limited information about the nature of the learning curve; these limitations are illustrated later in this chapter.

Simple series data statistical techniques

Curve fitting

A number of possible shapes of learning curves were identified in chapter 3 and these are listed in *Table 8*. It was not possible to compare the models when fitted as shown in the original equation because the measurement scales of the models were different. For example, the power law used the same units as the variable being measured ($y =$), whereas the log–linear model used the logarithm of the variable ($\log_e y =$). Therefore, all the equations were transformed on to the same scale, namely the y scale, and the resultant equations are displayed in the transformed equation column of *Table 8*.

These equations were fitted using weighted non-linear regression techniques that are available in most statistical packages. The weighted part of the regressions used a loss function of the form:

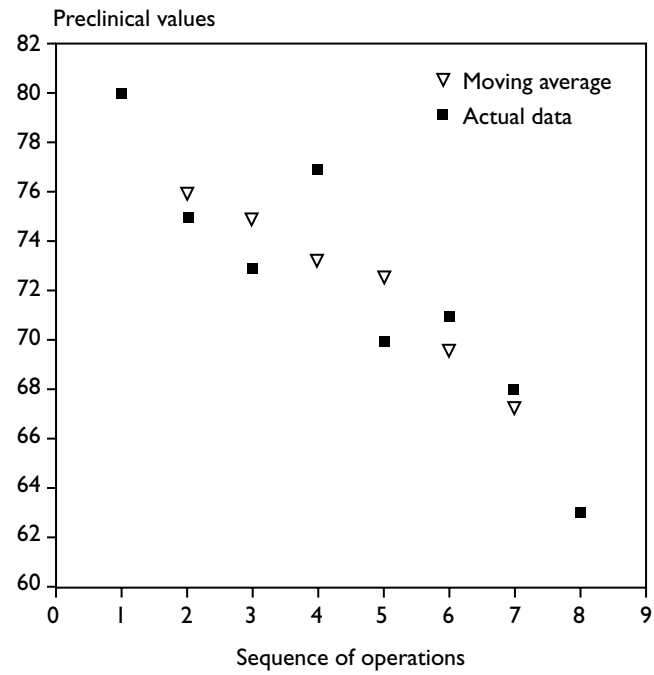


FIGURE 4 *Example of a moving average (order 3)*

TABLE 8 *The learning curve equations*

Type of curve	Original equation	Transformed equation
Linear	$y = a + bX$	$y = a + bX$
Quadratic	$y = a + bX + cX^2$	$y = a + bX + cX^2$
Cubic	$y = a + bX + cX^2 + dX^3$	$y = a + bX + cX^2 + dX^3$
Power law	$y = bX^{-c}$	$y = bX^{-c}$
Inverse	$y = a + \frac{b}{X}$	$y = a + \frac{b}{X}$
Exponential	$y = ae^{-bX}$	$y = ae^{-bX}$
Double exponential	$y = ae^{-bX} + ce^{-dX}$	$y = ae^{-bX} + ce^{-dX}$
Logarithmic	$y = a \ln X + b$	$y = a \ln X + b$
Log–linear	$\ln y = aX^b$	$y = \exp(aX^b)$
Log-log–linear	$\ln y = a(\ln(X + 1))^b$	$y = \exp(a(\ln(X + 1))^b)$
Logistic	$y = \frac{a}{1 + be^{-cX}}$	$y = \frac{a}{1 + be^{-cX}}$
Weibull	$y = a - be^{-cX^d}$	$y = a - be^{-cX^d}$

$$\sum_{i=1}^{n} w_i (y_i - \hat{y})^2$$

where $\hat{y}$ is the predicted value from the model for point i, y_i is the observed value at point i, w_i are the weights at point i, and n is the total number of data points.

Examination of the residuals from the models indicated that error variance was approximately proportional to the square of the predicted values. To permit impartial comparison of the models, the weights were kept constant between models and were denoted by:

$$w_i = \frac{1}{\left(\hat{y}_{i_{power}}\right)^2} \qquad \text{Equation 2}$$

where $\hat{y}_{i_{power}}$ is the predicted value at time i for the power curve model.

The models were not nested within one another, so the standard statistical method of comparison using the reduction in the residual sum of squares (RSS) was not appropriate. The models were compared using the Akaike information criterion (AIC),[102] which evaluates each model according to its RSS and the number of estimated parameters it contains. It was calculated using the following formula:[103]

$$\text{AIC} = -2\ln(\hat{\theta}) + 2k \qquad \text{Equation 3}$$

where $\ln(\hat{\theta})$ is the natural logarithm of the maximum likelihood of the model and k is the number of estimated parameters in the model. After testing for whether the errors were normally distributed, the maximum likelihood of the models was calculated, thus:

Equation 4

$$\ln(\hat{\theta}) = \frac{n\{\log_e(n) - \log_e(2\pi) - 1 - \log_e(RSS)\}}{2}$$

where n is the total number of observations and RSS the loss function given above.

The possibility of autocorrelation (serial correlation) was investigated using the autocorrelation function and the partial autocorrelation function.

Extension of curve fitting to include covariates

The curve fitting techniques can be extended to include other covariates. The effects of possible confounders Z_i on operation time (y) were explored by extending the linearised power curve model using sequence order X, from:

$$\log_e(y) = \log_e(\beta_0) - \beta_1 \log_e(X)$$

to

$$\log_e(y) = \log_e(\beta_0) - \beta_1 \log_e(X) + \sum \alpha_i Z_i$$

where β_0 and β_1 are the coefficients associated with the power curve and α_i the coefficients associated with the confounding factors.

Dichotomous measures of performance

Measures of performance that involved a dichotomous (yes or no) outcome were investigated using the following methods.

Exploratory techniques

Cusum chart

A cusum chart represents the sequential level of performance of a series of data.[19] For a series $\{(X_1, \ldots, X_i): i = 1, 2, \ldots, n\}$, where X_i are dichotomous (0 = success, 1 = failure) measures of performance, then the cusum series is defined as:

$$s_i = \sum_{j=1}^{i} (X_j - X_0) \qquad \text{Equation 5}$$

where X_0 is a predetermined reference level representing the level of performance that is desirable (90% success, say). A cusum chart is interpreted in three ways:

(i) if the true success rate is X_0, then cusum will be flat
(ii) if the true success rate exceeds X_0, then cusum will decrease
(iii) if the true success rate is less than X_0, then cusum will increase.

The cusum technique was applied to intra-operative complications and postoperative complications of laparoscopic fundoplication.

Simple series data statistical techniques

Data splitting

The variables 'intra-operative complication' and 'postoperative complication' were used as dichotomous proxies for learning to explore data splitting methods. The series were split into thirds and quarters, and trends across groups were tested using a chi-squared test for trend.

Logistic regression

Logistic regressions can be used to determine whether the sequence number of the procedure

and other possible confounding factors predict outcome. This was done by assessing the relationship between age and sex of the patient, and operative and postoperative complications.

Results

Operative details

The mean age of the 190 consecutive patients in the study was 45.8 years (standard deviation (SD) 13.7) and 78 of them (41%) were female. At the time of operation, 130 (68%) patients were reported to have a hiatus hernia (29 large, 32 medium, 69 small). No procedure was converted to an open approach. The average operation time was 87 minutes.

Analysis of the continuous variable 'operation time'

Scatter diagram

A scatter diagram of operation time against operation sequence number is shown in *Figure 5*. The figure demonstrates that the operation times were extremely variable but there is a suggestion that the operation time decreased over time.

Moving average

An order 10 moving average of the operation time is shown in *Figure 6*. This demonstrates a decrease in operation time from approximately 130 minutes in the first ten operations to about 80 minutes from the 60th case onwards.

Data splitting

The operation times were split sequentially into quarters and thirds, and the respective group means of operation time are shown in *Table 9*. When the data had been split into quarters, the mean operation time for the first 50 cases was 104.3 minutes compared with 75.6 minutes for the last 40 cases. There was a highly significant decreasing trend across the four groups ($p < 0.001$) and the Scheffe multiple comparison showed that operation times for the first 50 cases were significantly longer than those for the other three groups. A similar pattern was observed when the data were split into thirds, however, and the mean operation time for the first 70 cases was significantly different from that for the other groups.

Curve fitting

The results of the 12 curve fits are displayed in *Table 10*, with the curves ranked by increasing size

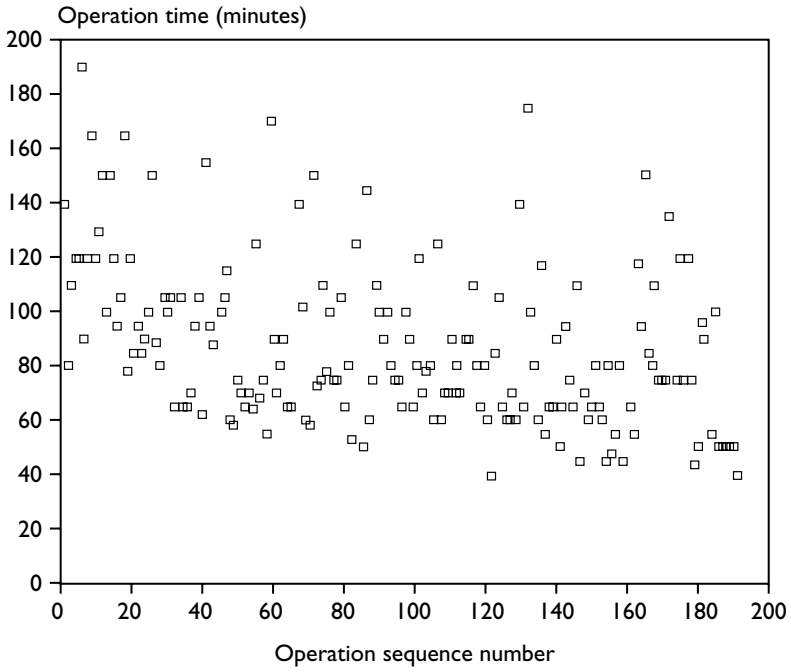


FIGURE 5 *Scatterplot of operation time against operation sequence*

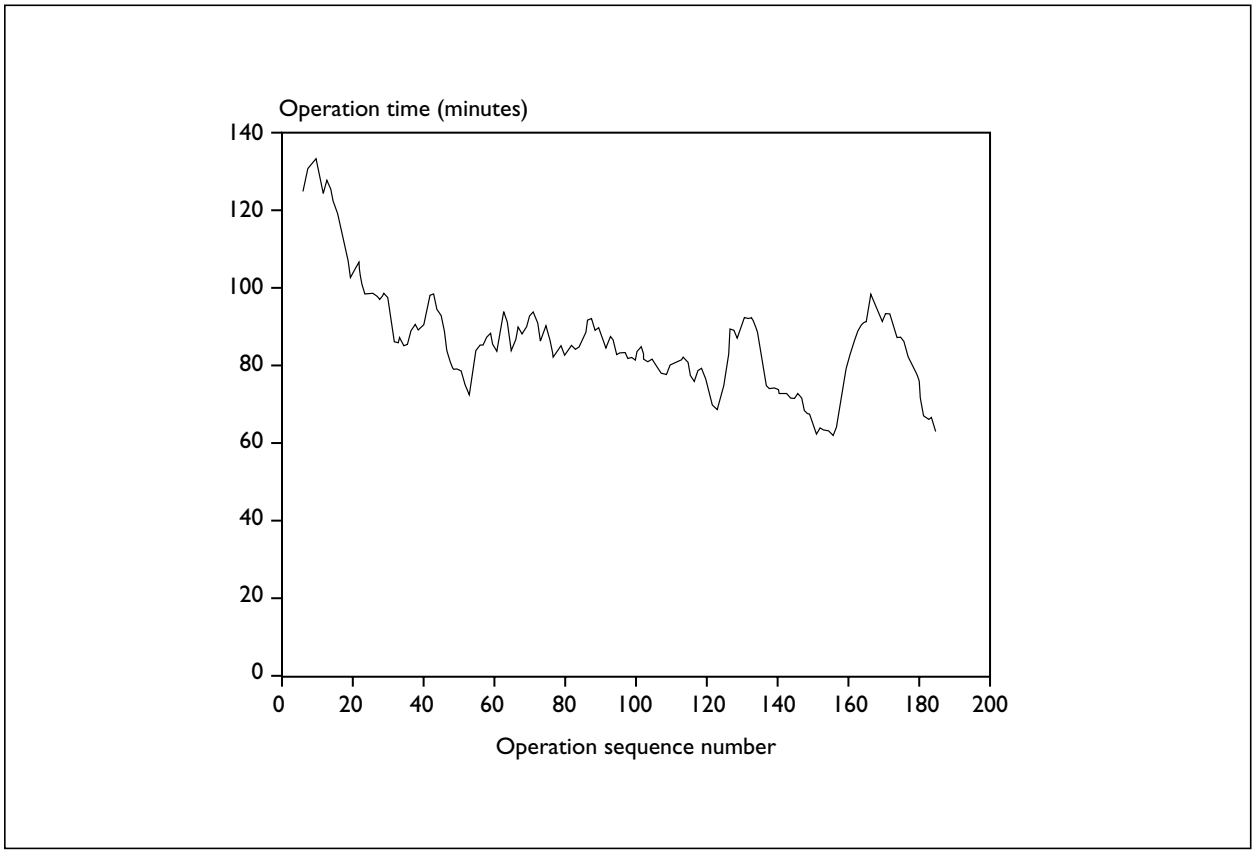


FIGURE 6 *Moving average (order 10) of operation time against operation sequence*

of AIC (low AIC denoted a better fit). The best fitting shape was the logarithmic curve and the worst the inverse curve. There were no marked differences between the fits of the logarithmic, power law and log–linear curves. The predicted operation times for the 5th, 30th, 100th and 190th procedures were similar, to within 2 minutes, across the first three models. All of the predicted values are shown in *Figure 7 (A–L)*. The figure shows clearly that most of the curves were of similar shape except for the linear, quadratic, cubic and inverse curves. However, the predicted operation times for the first procedure ranged from 105 minutes to 205 minutes. Inspection of the autocorrelation function and the partial autocorrelation function suggested that autocorrelation was not present.

Extension of curve fitting

The effects of the logarithm of sequence order, sex of patient, age of patient and size of hernia were explored using multiple regression. The results are shown in *Table 11*. Both sequence of operations and age of patient had a significant impact on operation time – the older the patient, the longer the operation took. For example, if the operation is the 100th that a surgeon performed and the patient was 30 years old, then the predicted operation time was 71 minutes. If the patient was 45 years old, the predicted operation time was 77 minutes. The value of the coefficient of determination, R^2, improved from 19%, with only the logarithm of sequence order in the model, to 24% when age was included.

Analysis of the dichotomous variables 'intra-operative complication' and 'postoperative complication'

There were 20 patients with an intra-operative complication (such as liver injury or bleeding) and 32 patients with a reported postoperative complication (such as difficulty in swallowing).

Cusum charts

Figure 8 is a cusum chart of intra-operative complications assuming a proficiency level of 90%. It shows that performance improved over at least the first 40 procedures, although arguably the improvement was over the first 60 because the cusum chart was increasing over this period. The level of performance achieved was approximately 90% from cases 70 to 100 (cusum chart level), and exceeded 90% for the rest of the series. This is the expected shape of a cusum chart if learning is present.

TABLE 9 *Splitting operation time data into quarters and thirds*

Group	*n*	Mean (SD)	Linear trend across groups (significance)	Scheffe pairwise multiple comparison
Quarters				
1st	50	104.3 (30.0)	$p < 0.001$	1st group significantly different from all other groups ($p < 0.001$). No other groups were different
2nd	50	85.7 (26.4)		
3rd	50	79.1 (25.2)		
4th	40	75.6 (27.1)		
Thirds				
1st	70	98.4 (31.2)	$p < 0.001$	1st group significantly different from all other groups ($p < 0.001$). No other groups were different
2nd	60	83.0 (23.4)		
3rd	60	76.8 (27.8)		

TABLE 10 *Parameter estimates and AIC values of the learning curves*

Type of curve	AIC	Rank	Parameter estimates				Predicted values (minutes)			
			A	B	C	D	5th	30th	100th	190th
Logarithmic	92.126	1	145.27	–13.72			123.2	98.6	82.1	73.3
Power law	92.555	2	158.25	–0.14			125.7	97.2	81.8	74.6
Log–linear	92.876	3	5.08	–0.03			125.8	96.8	81.8	75.0
Logistic	93.480	4	75.76	–0.44	0.02		126.4	99.5	80.3	76.4
Cubic	93.748	5	130.75	–1.36	0.012	–0.00004	124.3	100.0	81.3	65.6
Exponential	93.911	6	77.11	56.53	0.03		125.7	99.8	79.8	77.3
Double exponential	94.183	7	51.19	0.06	91.98	0.001	129.6	97.5	81.9	73.6
Weibull	95.588	8	73.59	71.75	0.11	0.67	126.1	99.2	80.8	75.7
Quadratic	97.427	9	113.68	–0.46	0.0013		111.4	101.1	81.4	76.0
Log-log–linear	97.526	10	5.06	–0.09			121.9	92.5	82.0	77.9
Linear	98.878	11	103.29	–0.18			102.4	97.9	85.2	68.9
Inverse	111.510	12	81.62	121.85			106.0	85.7	82.8	82.3

Figure 9 shows a cusum chart of postoperative complications, assuming a proficiency level of 80%. In contrast with the intra-operative complications chart, the postoperative cusum chart shows that performance consistently exceeded 80% over the first 100 cases but then remained at 80% for the rest of the series.

Data splitting

The complication rates for the series split into quarters and thirds are shown in *Table 12*. The rate of intra-operative complications decreased from 18% in the first quarter to 5% in the final quarter. There was a highly statistically significant decreasing trend across the quarters. A similar pattern was observed when the series was split into thirds. In contrast, there were no significant trends in postoperative complication rates across either set of grouping.

Logistic regression

From *Table 13* it can be seen that the operation sequence number was a statistically significant predictor of an intra-operative complication (the risk decreasing as more procedures were performed) but that no other potentially confounding factors were significant. None of the identified variables predicted a post-operative

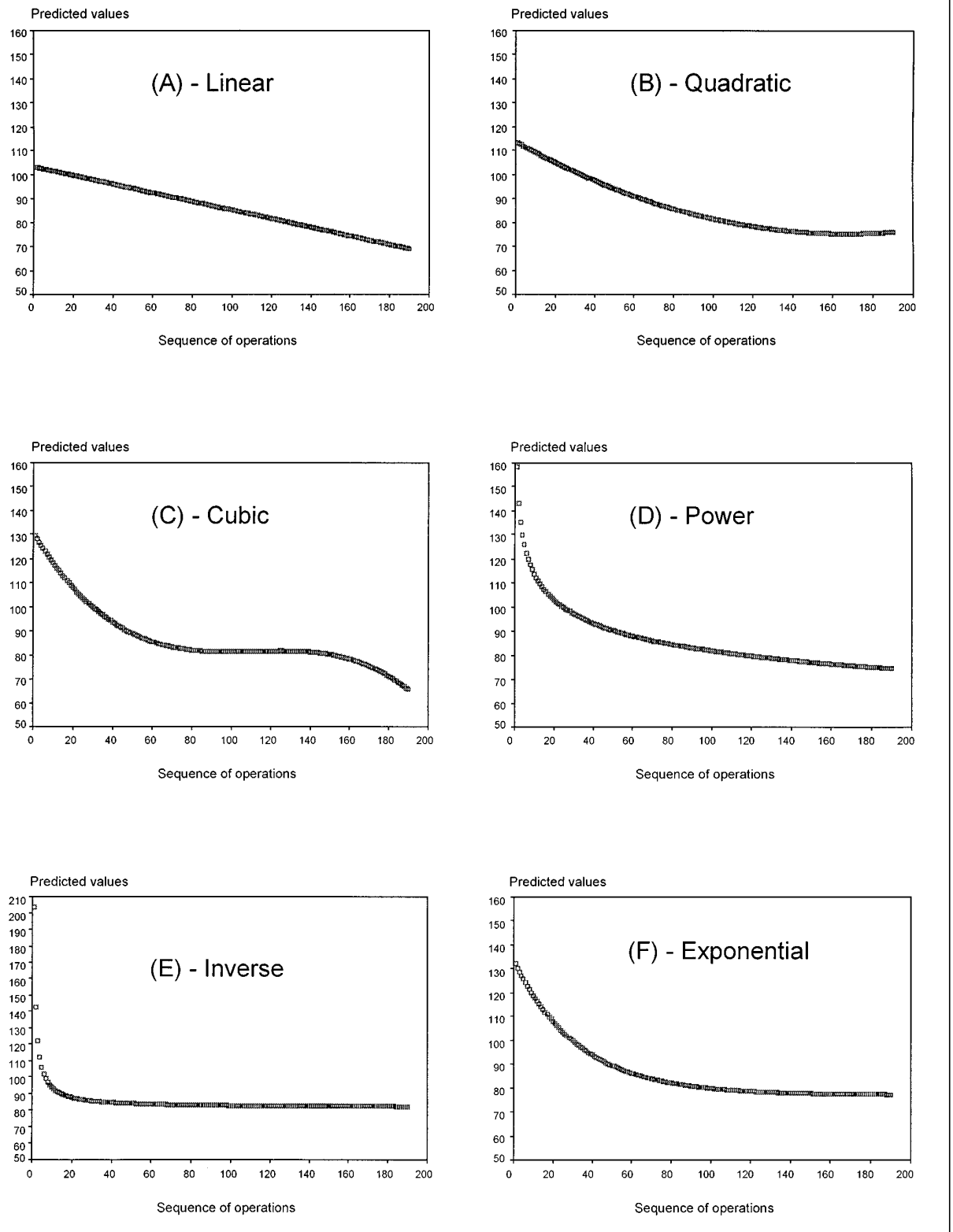
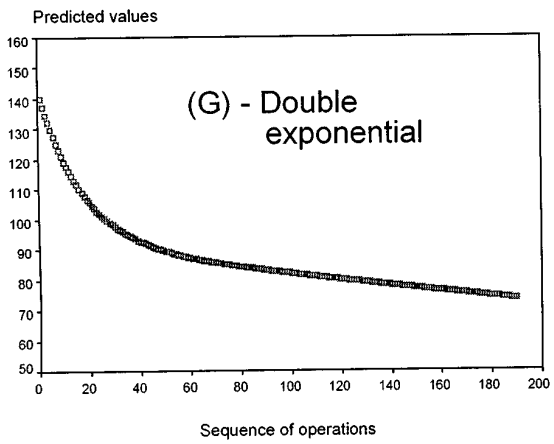
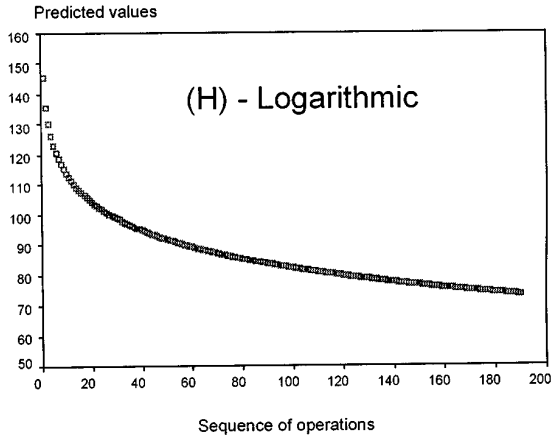
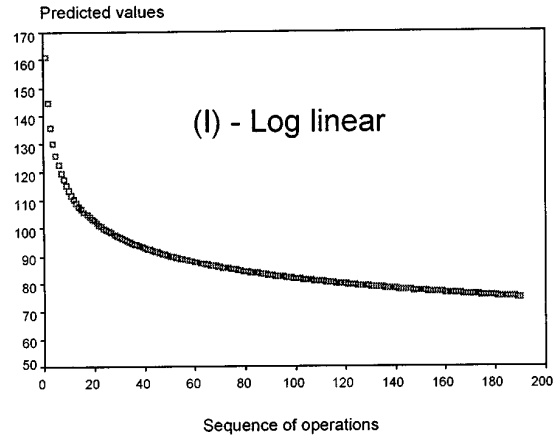
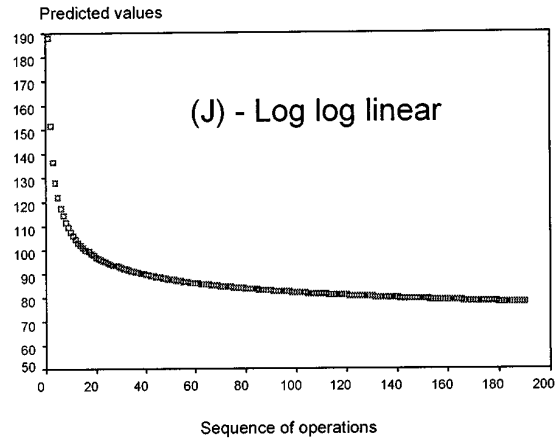


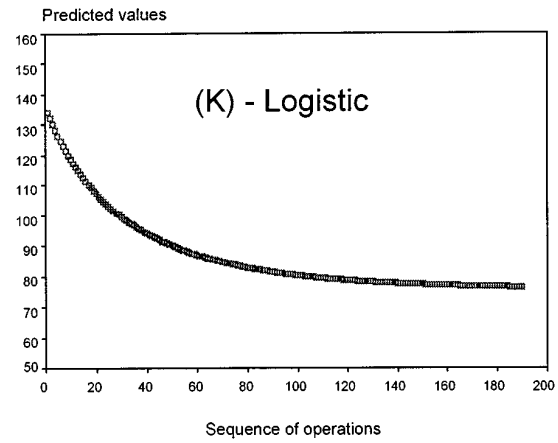
FIGURE 7 *Predicted values of the 12 learning curve models*











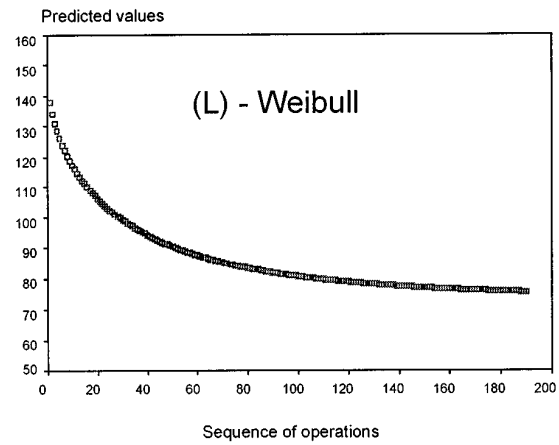


FIGURE 7 contd *Predicted values of the 12 learning curve models*

TABLE 11 *Regression analysis of the logarithm of operation time*

Variable	Coefficient	95% CI	Significance
Included in model			
Constant	4.850	(4.627 to 5.079)	$p < 0.001$**
Logarithm of sequence	–0.159	(–0.202 to –0.115)	$p < 0.001$**
Age	0.005	(0.002 to 0.008)	$p = 0.001$**
Not included			
Male	0.076	(–0.010 to –0.165)	$p = 0.09$
Medium hernia	0.042	(–0.094 to 0.141)	$p = 0.68$
Large hernia	–0.085	(–0.178 to 0.008)	$p = 0.07$

** *Significant at the 0.1% level*

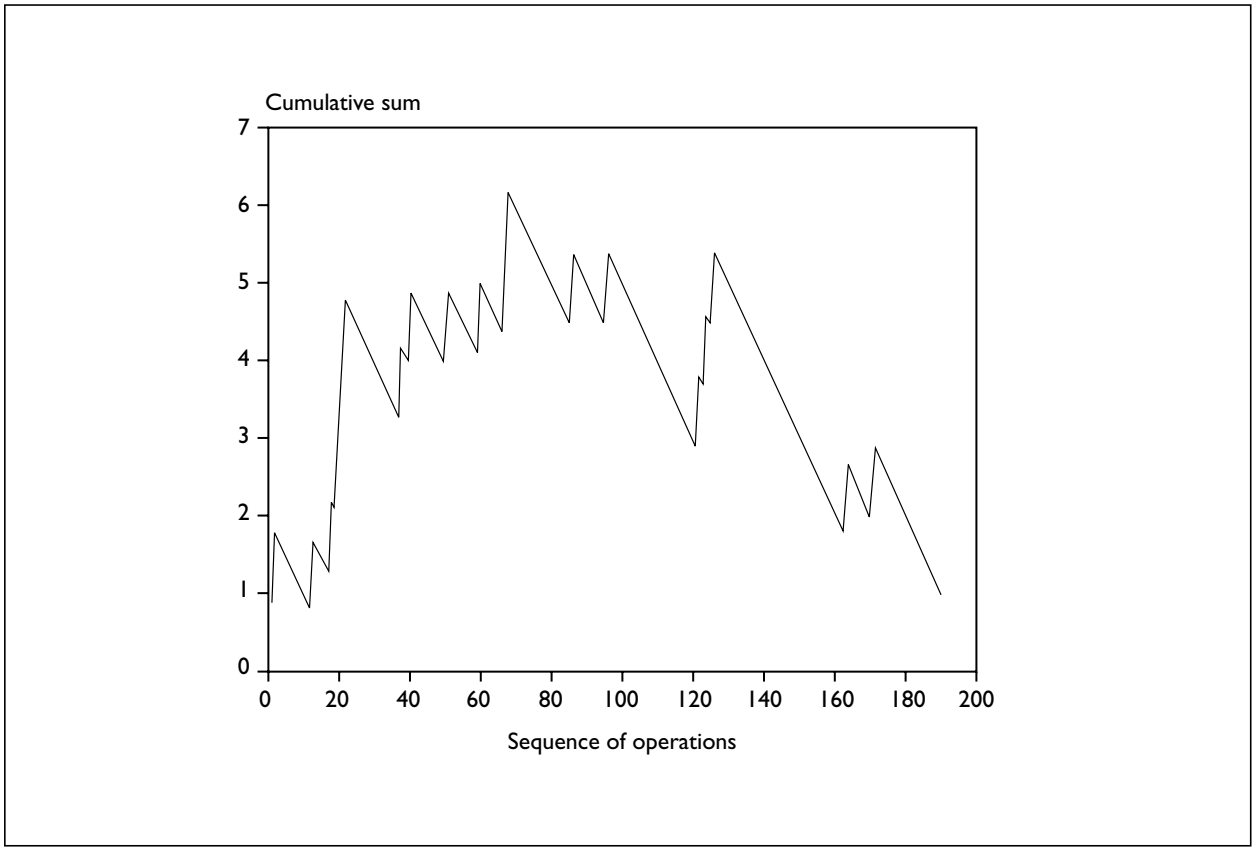


FIGURE 8 *Cusum chart for intra-operative complications at 90% proficiency*

complication and, hence, there was no evidence of a learning curve effect for this outcome.

Discussion

Learning curve effects are commonly observed in case series studies of non-pharmacological technologies such as minimal access surgical procedures. As discussed in chapter 2, the statistical methods that have been used in the past to measure these effects have been sub-optimal. In this chapter, a hierarchy of statistical methods to **identify** and **measure** learning curve effects were illustrated and explored using a case series data set on laparoscopic fundoplication. These methods were identified by literature searches of both clinical and non-clinical fields as described in chapters 2 and 3.

Operation time was used to illustrate statistical methods that are applicable to continuous meas-

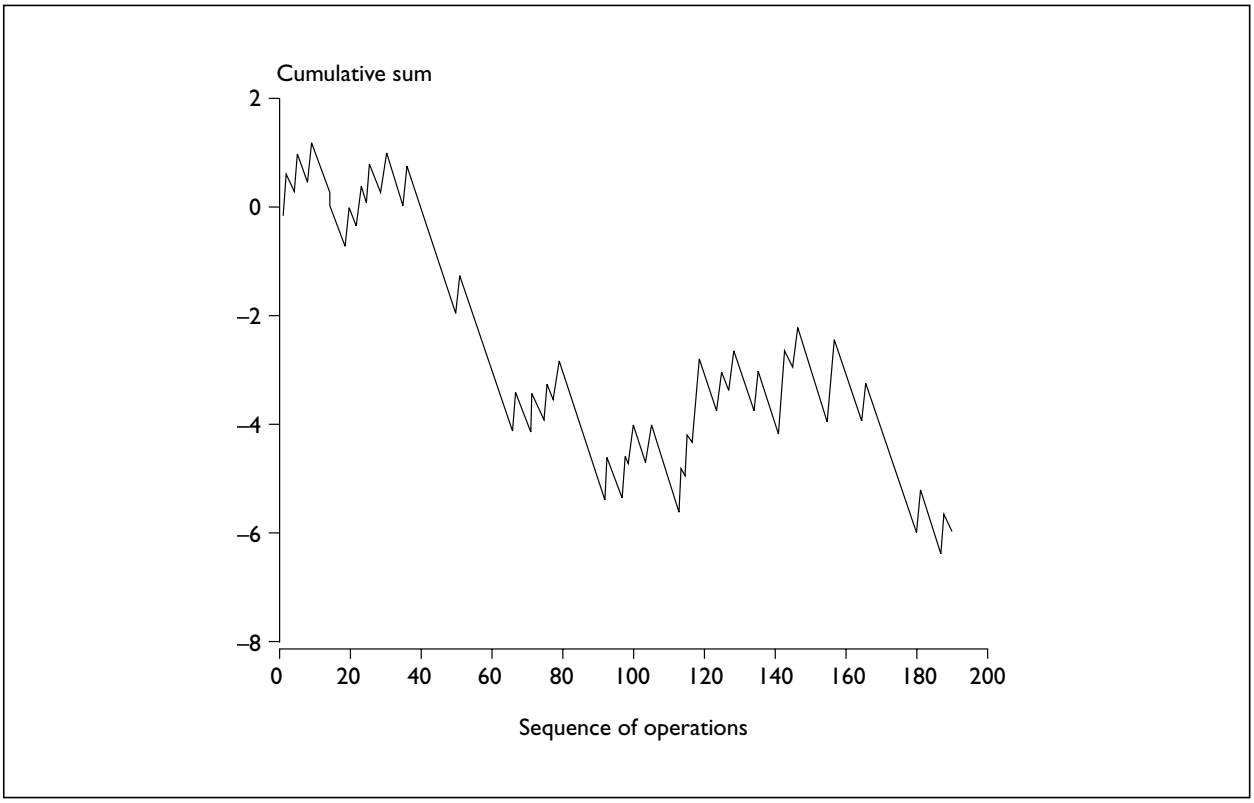


FIGURE 9 *Cusum chart for postoperative complications at 80% proficiency*

TABLE 12 *Splitting complication rate data into quarters and thirds*

Group	Number of patients	Complications	
		Operative *n* (%)	Postoperative *n* (%)
Quarters			
1st	50	9 (18)	8 (16)
2nd	50	6 (12)	8 (16)
3rd	50	3 (6)	11 (22)
4th	40	2 (5)	5 (10)
Linear trend		$p = 0.024$	$p = 0.933$
Thirds			
1st	70	13 (19)	10 (14)
2nd	60	5 (8)	13 (22)
3rd	60	2 (3)	9 (15)
Linear trend		$p = 0.004$	$p = 0.873$

ures of performance. As discussed in chapter 2, there are problems in using operation time as a proxy for performance. Nevertheless, operation time was used often in the literature as evidence of changes in technique over time, and it is certainly one aspect of performance. There were no other continuous measures of performance in the laparoscopic fundoplication data set.

There was great variability in operation times across the study period; this variability was smoothed using a moving average procedure. The smoothed data clearly showed a decrease in operating time over time. However, applying a moving average procedure to data requires a degree of caution. The order of smoothing can affect the observed pattern in the data. If the order is very high it can make the data too smooth, thus disguising real variability in the data. Conversely, a low order may not smooth the data enough and so no pattern is discernible. In practice, it is suggested that investigators try fitting a variety of orders to the data. Also, it should be emphasised that a moving average is an exploratory technique and should not be used as the sole basis for recommendations on the learning curve. The authors also recognise that a moving average is not the only statistical method used to smooth data. Other methods that have been described include kernel smoothing, differencing and polynomial regression.[100]

It has been demonstrated that split groups can certainly be used to identify a change over time when the change is large, but the usefulness of this method in the context of measuring the size of a learning curve effect is questionable. The results of splitting the operation time data into quarters or

TABLE 13 *Logistic regression analysis of intra- and postoperative complications*

Variable	Coefficient (95% CI)	Significance
Intra-operative complications		
Constant	–0.756 (–2.854 to 1.342)	$p = 0.48$
Sequence	–0.013 (–0.023 to –0.003)	$p = 0.007^{**}$
Male	–0.170 (–1.192 to 0.852)	$p = 0.75$
Age	–0.011 (–0.047 to 0.025)	$p = 0.56$
Medium hernia	0.689 (–1.173 to 1.699)	$p = 0.72$
Large hernia	–0.756 (–0.364 to 1.742)	$p = 0.20$
Postoperative complications		
Constant	–1.406 (–3.125 to 0.313)	$p = 0.11$
Sequence	–0.001 (–0.008 to 0.006)	$p = 0.84$
Male	0.100 (–0.716 to 0.916)	$p = 0.81$
Age	–0.001 (–0.030 to 0.028)	$p = 0.97$
Medium hernia	–1.237 (–1.762 to 0.288)	$p = 0.11$
Large hernia	0.029 (–0.784 to 0.842)	$p = 0.94$

** *Significant at the 0.1% level*

thirds led to the similar overall conclusion that early patients had a longer average operation time than later groups. The length of the learning curve is unclear, however, and could be considered to be either 50 cases or 70 cases depending on the type of data splitting performed. The authors recommend that the split group method should only be applied using a linear trend component, and then interpreted as statistical (instead of graphical) evidence of a change in performance over the series.

A range of curves were fitted and statistical methods were demonstrated that can be used to compare these curves using nonlinear regression and the AIC. There are other ways in which this might have been explored, such as the Bayesian information criterion,[104] but these were not investigated further. It is recommended that any analysis of learning curves should compare at least three or four of these curves. Within the laparoscopic fundoplication data, there were only small differences between curves. However, the logarithmic, power and log–linear models provided the best fits, accounting for approximately 20% of the variability. The apparently poor fit to the first four or five data points is not surprising, as it has been observed in many other performance-related data sets.[74] Further research is required to assess the generalisability of the various shapes in other health technologies.

Measurements taken close together in time are often more similar to each other than to those taken further apart in time, irrespective of any other reason for change such as learning. On this basis, it could be hypothesised that the outcome of the first procedure would be more similar to that of the second procedure than it would be to that of the 200th procedure. This type of relationship is termed autocorrelation or serial correlation, the presence of which complicates assessment. The standard error of the estimated parameters is reduced by positive autocorrelation (or increased by negative autocorrelation) and, hence, the apparent significance of each parameter is overestimated (or underestimated). Autocorrelation should be investigated using the appropriate function that is found in most standard statistical packages. Cook and Campbell[105] give a detailed discussion of the interpretation of these techniques.

Changes in patient referral and selection often mean that the type of patient changes as the operator (or institution) performs more procedures. This can be investigated by including potentially confounding factors, such as age or sex, in the curve equations. This is a relatively simple procedure that provides a more informative result. It was shown here that the age of the patient was associated with a significant increase in the operation time for fundoplication. Investigators should consider and adjust for any confounding factors when investigating learning curve effects in case series designs.

The cusum chart is an effective graphical method for looking for trends or changes over time in dichotomous proxies for learning. It was shown that there was almost certainly a decrease in the rate of operative complications during laparoscopic fundoplication, with lower rates from about the 25th procedure onwards, and even lower rates

from about the 125th procedure. This type of information is invaluable for the evaluation of new technologies. If operators (or institutions) are beyond their learning curve, then an RCT can be initiated that is free from this source of bias. Alternatively, if an RCT included operators at various levels on the learning curve, then cusum charts for each operator can be used to investigate potential operator differences. The cusum chart is a graphical method. However, de Leval and colleagues[67] showed that it can be used to monitor surgical performance statistically. It may also be extended to monitor paired binary events simultaneously.[68] Lovegrove and colleagues[106] illustrated the use of the cusum chart for adjusting for case-mix. In all these studies, the cusum chart was used to monitor a system going out of control, whereas learning curve data represent a system coming under control. The cusum method is undoubtedly a powerful tool for observing differences between operators; however, it is not yet clear to the authors how data derived from a cusum chart could be extrapolated to a rigorous assessment by RCT. Further empirical research is needed to investigate the implications of this difference.

Predictors of dichotomous outcomes can be tested using logistic regression. The sequence number was identified as predicting an intra-operative complication – the earlier the procedure, the higher the risk of an operative complication. Nevertheless, dichotomous outcomes can be particularly difficult to analyse statistically if they are rare. No statistical techniques designed for rare events were identified.

Conclusion

It has been shown here how the extension of currently used methods to more appropriate methods can improve the analysis and interpretation of learning curves in single operator case series of new health technologies. In the case series of laparoscopic fundoplication used as an example, the analysis of operation times and intra-operative complications strongly suggested a learning curve for this surgeon. The applicability and extension of these methods to studies that have multiple case series are discussed in the next chapter.

Chapter 5

Multiple operators: a case study of laparoscopic cholecystectomy

Introduction

Some of the statistical methods that can be applied to complex structured data are illustrated and explored in this chapter. For this, a case series of consecutive laparoscopic cholecystectomy procedures performed by ten surgeons has been used. In particular, the aims were:

(i) to illustrate that the statistical techniques for simple series data can be applied to complex structured data for preliminary examination
(ii) to describe the strengths and weaknesses of multilevel modelling of learning curve effects.

Laparoscopic cholecystectomy

Cholecystectomy (removal of the gallbladder) is a long-accepted method of treating patients with symptomatic gallstones. Open cholecystectomy was first performed by Langenbuch in 1882. Open or traditional cholecystectomy involves making a 10–15 cm incision, through which the gall bladder is removed. In contrast, laparoscopic cholecystectomy involves making three or four incisions, varying from 0.5 cm to 1 cm, to provide access for the laparoscopic and surgical equipment and an opening through which the gallbladder is removed.

Laparoscopic cholecystectomy is a minimally invasive procedure that is associated with shorter times for operating, hospitalisation and recovery than open cholecystectomy. Despite this, there remains some uncertainty about the safety of laparoscopic procedures compared with open procedures.[107] In particular, several studies have documented possible learning curve effects with this new procedure.[108–113] These studies suggested that operation times, conversion rates and operative complication rates decreased as the surgeon (or institution) performed more procedures. Estimates of the number of cases that were considered to constitute the learning phase ranged from 8 to 40 but there was no clear justification for the derivation of these figures.

Methods

The laparoscopic cholecystectomy data set

The first groups of patients of ten surgeons were the case series used in these analyses. The total number of patients included in the analyses was 1481. The patients underwent laparoscopic cholecystectomy in Aberdeen over an 8-year period between March 1991 and March 1999. Information was collected prospectively on patient characteristics, operative technical details and hospitalisation. No postoperative details were available.

Preliminary examination of multiple operator data

Some of the techniques for simple series that were described in chapter 4 can be used for initial exploration of complex structured data. These techniques are described below.

Box and whisker plots

A box and whisker plot is a summary plot of the data based on the median, quartiles and extreme values. The box represents the interquartile range, which contains that 50% of values between the 25th and 75th percentiles (*Figure 10*). The line across the box indicates the median. The whiskers are lines that extend from the box to the highest and lowest values, excluding unusually high or low values. An observation is an outlier (denoted by a circle in *Figure 10*) if it lies between 1.5 and 3 box-lengths from the box edge, and an observation is an extreme value if it lies beyond three box-lengths (denoted by an asterisk in *Figure 10*). The usefulness of this plot is illustrated by summarising the operation times of the ten laparoscopic cholecystectomy surgeons (*Figure 11*).

Error bars

If data are normally distributed, then the mean and 95% CI of the mean can be plotted for each operator. These are called error bar plots.

The box and whisker and error bar plots are used to summarise the data for individual operators or institutions; they do not give information about changes over time (i.e. learning effects). Changes

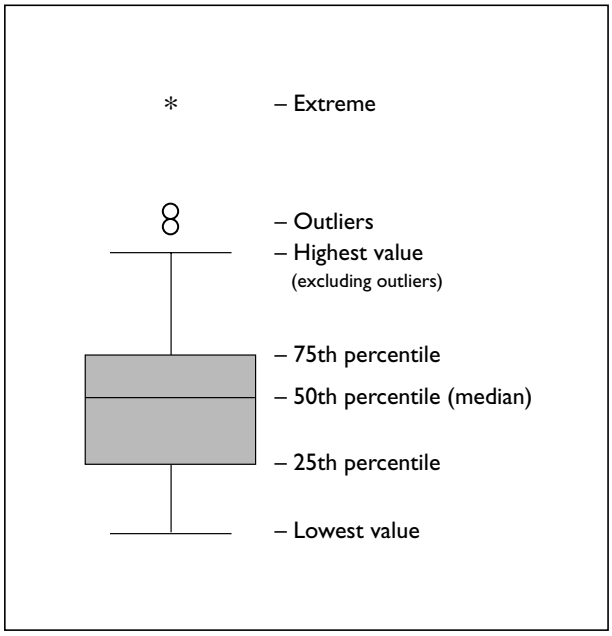


FIGURE 10 *A box and whisker plot*

in individual operators' performance within multiple operator data sets can be investigated using moving averages and curve fitting as described in chapter 4.

Moving averages

The results of the three surgeons who performed the most procedures were selected to illustrate moving averages of operation time within the cholecystectomy data.

Curve fitting

The three best fitting curves were chosen from chapter 4 (logarithmic, power and log–linear) and were applied, using non-linear regression, to the results of the three selected surgeons. Two regressions per surgeon were performed for each shape. The first used all of the operation times for each surgeon and the second used only the first 150 cases per surgeon. The predicted values from these curve fits were used to investigate graphically any intra- and inter-operator differences.

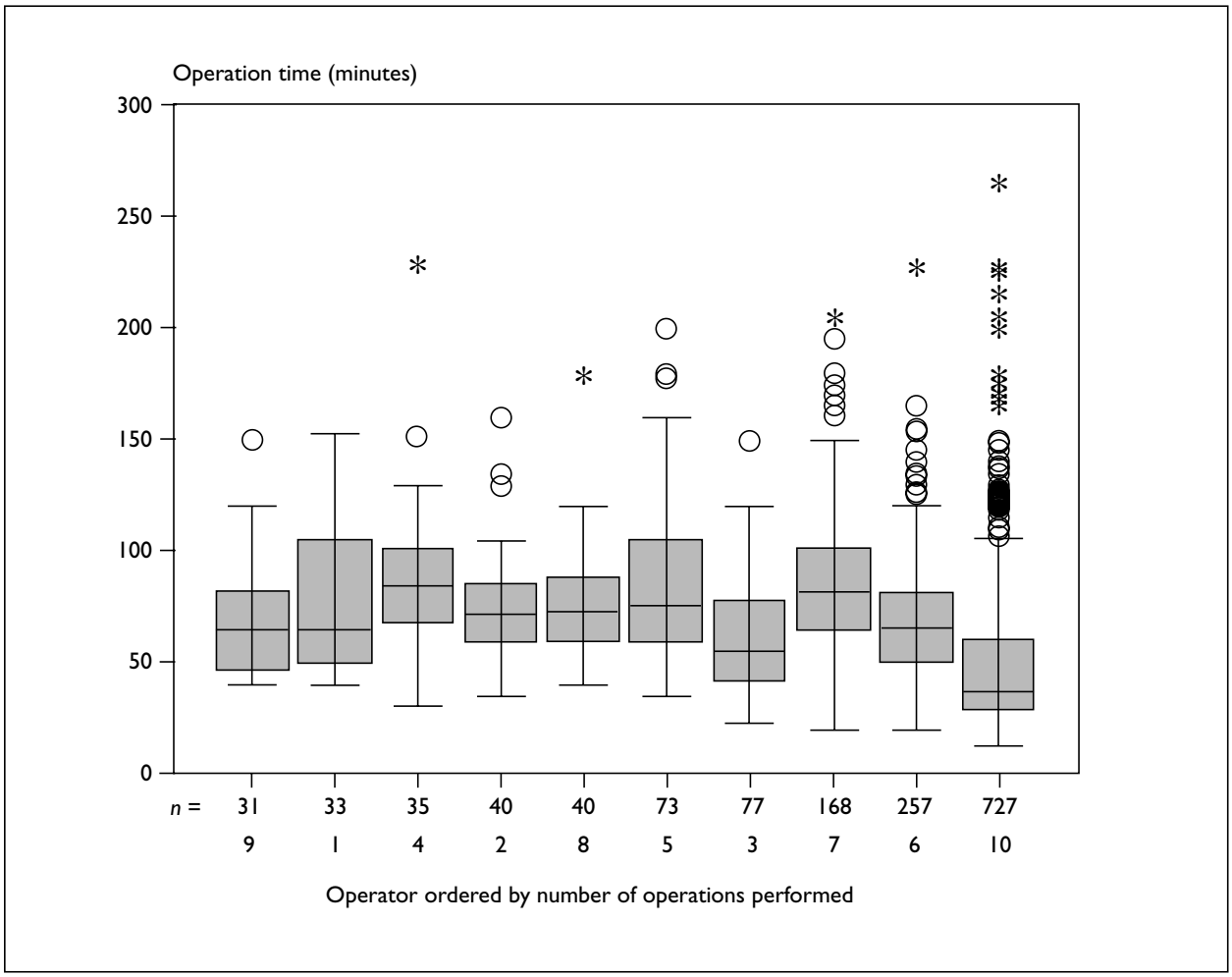


FIGURE 11 *Box and whisker plots of operation time for each surgeon (○ outlier; * extreme)*

Techniques for complex structured data

In chapter 3, a number of statistical techniques were identified that could be applied to learning curve data during a health technology assessment: discriminant analysis, two-stage models, multilevel models, latent curve models, stochastic process models and ARIMA models. Of these, multilevel models were focused on in response not only to time constraints but also to the responses received from the questionnaire sent to experts in various fields. A number of experts have suggested that multi-level models might have been applicable to hierarchical learning curve type data – procedures within operators within institutions.

Multilevel models

The main aim of multilevel analysis is not to describe the individual operators (surgeons) in the sample but to estimate the pattern of variation in the underlying population of operators. In the subsequent discussion of the methodology a two-level model is assumed, namely patients (level 1) nested within operators (level 2). The methodology can be extended easily to more than two levels.

Constant rate of learning and fixed starting levels

It was assumed that, perhaps after an appropriate transformation, the learning curve for each operator (j) could be expressed as a linear regression equation for the i-th patient. Moreover, it was assumed that the regression equations for the different operators were parallel to each other. Thus:

$$y_{ij} = \alpha_j + \beta_0 x_{ij} + e_{ij} \qquad \text{Equation 6}$$

where α_j allows a different starting value for each operator and β_0 represents a common rate of learning. The $\text{var}(e_{ij}) = \sigma_e^2$ represents the variance of the measurements in the patients. The variation in starting level from operator to operator can be modelled simply using standard regression techniques with **fixed effects**. The fixed effect assumes that all individuals in a data set are distinctive. In contrast, the regression model can be generalised by postulating **random effects**.

Constant rate of learning and random starting levels

There is now an assumption that there is some overall starting value (α_0) for the population of operators as a whole and that u_j is the random (rather than distinctive) departure of the j-th operator from this overall value. The departure is assumed to come from a normal distribution with mean zero and variance, σ_u^2. The model can now be written as:

$$y_{ij} = \alpha_0 + u_j + \beta_0 x_{ij} + e_{ij} \qquad \text{Equation 7}$$

where u_j and e_{ij} are random effects at different levels, and it is the existence of these variables that makes it a multilevel model. A multilevel model of this type is known as a variance components model. The intraclass correlation coefficient for this model can be calculated by dividing the variability between operators by the total variability in the model. This is calculated as $\sigma_u^2/(\sigma_u^2 + \sigma_e^2)$ and represents the proportion of variability in the data that is due to differences between operators compared with within-operator variability.

Fixed rate of learning and random starting levels

The above model can be extended to allow the rate of learning to differ between operators. The model can now be written as:

$$y_{ij} = \alpha_0 + u_j + \beta_{0j} x_{ij} + e_{ij} \qquad \text{Equation 8}$$

The fixed effect (β_{0j}) assumes that all rates of learning per operator are distinctive. These are commonly modelled as dummy variables for each operator in a standard regression analysis.[85]

Random rate of learning and random starting levels

In a two-level hierarchy, in addition to a random starting level, it is often desirable to consider if the regression coefficient for the rate of learning varies randomly between operators. The model is a simple extension of the variance components model to:

$$y_{ij} = \alpha_0 + u_j + (\beta_0 + v_j)\, x_{ij} + e_{ij} \qquad \text{Equation 9}$$

where v_j is the departure of the j-th operator from the true population rate of learning (β_1) with mean zero and variance, σ_v^2. The equation for the model is usually written in the following way:

$$y_{ij} = \alpha_{0j} + \beta_{0j} x_{ij} + e_{ij} \qquad \text{Equation 10}$$

with $\alpha_{0j} = \alpha_0 + u_j$ and $\beta_{0j} = \beta_0 + v_j$. To specify the model fully:

$$E(\alpha_{0j}) = \alpha_0;\ E(\beta_{0j}) = \beta_0;\ \text{var}(\alpha_{0j}) = \sigma_u^2;$$
$$\text{var}(\beta_{0j}) = \sigma_v^2;\ \text{cov}(\alpha_{0j}, \beta_{0j}) = \sigma_{uv};$$
$$\text{and } \text{var}(e_{ij}) = \sigma_e^2$$

where α_0 is the overall starting value; β_0 is the overall rate of learning; σ_u^2 is the variance of

starting values between operators; σ_v^2 is the variance of slopes between operators; σ_{uv} is the covariance between the starting values and the rates of learning; and σ_e^2 is the intercept component of level one patient variance.

The rate of learning and starting value was estimated for each operator and CIs were derived from the residuals from the random effects. Differences between models were tested using the reduction in log-likelihood compared to a chi-squared distribution with the appropriate degrees of freedom.

Covariates in multilevel models

It is possible to explain the pattern of variation in terms of characteristics of operators by incorporating further variables into the model. Further mathematical details are not given here but readers are referred to Goldstein's book[114] for further information. The use of additional fixed effects (sex of patient, previous sphincterotomy, operative cholangiogram, and ruptured or inflamed gall bladder) on operation time are illustrated for the cholecystectomy data set.

Multilevel models are fitted using an iterative algorithm, so there can be problems with convergence to a sensible solution. In our data set, the operation times were positively skewed and convergence was not possible using untransformed data. Hence, the linearised form of the power curve was used to test for changes over time. The power learning curve is of the form (see also chapter 4):

$$operation\ time_{ij} = a_j\ (sequence_{ij})^{b_0} \qquad \text{Equation 11}$$

but, if the logarithm of this equation is taken, then the new equation is:

Equation 12

$$\log_e (operation\ time_{ij}) = \log_e (a_j) + b_0 \log_e (sequence_{ij})$$

This equation is a linear regression model as described above and so the multilevel modelling procedure described above can now be applied. All of the subsequent analyses had to be performed on the logarithmic scale for both operation time and experience.

Results

Preliminary examination of data

The ten surgeons performed 1481 laparoscopic cholecystectomies in total, with the number of procedures they each performed ranging from 31 to 727. Approximately 75% of the patients were female. Of all patients, 8% had an operative cholangiogram and 6% had had a previous sphincterotomy. The gall bladder was inflamed in 20% of patients and ruptured in 23% of patients.

Box plots and error bars

Box and whisker plots of operation times for each surgeon are shown in *Figure 11*. These demonstrate how the median operation times varied between surgeons. The surgeons were ordered according to the number of procedures each had performed. There is no evidence from this figure that the variability of operation times decreased as the number of procedures performed by each surgeon increased. However, this observation is based solely upon between-surgeon times rather than within-surgeon times.

Operation times were positively skewed within each surgeon; there were disproportionately more high operation times than expected. Logarithms were taken of the operation times to ameliorate this skewness. It was now possible to calculate the mean and 95% CIs of the logarithm of the operation times and these are illustrated in *Figure 12* as error bars. The error bars indicate that there are differences between the average (geometric) operation times of the surgeons.

Moving averages

A moving average of order 10 for three selected surgeons is shown in *Figure 13 (a–c)*. A small decrease in operation time with experience is shown.

Curve fitting

The parameter estimate and R^2 values for the log, power and log–linear models are shown in *Table 14* for each of the three surgeons. All of the curve fits were statistically significant at the 1% level but the small R^2 values showed that there were large variabilities in the operation times that could not be explained. Within each surgeon there was no apparent difference between the statistical fits of the three models.

The predicted operation times for each of the surgeons are shown in *Figure 14 (a)*, assuming a power curve relationship. The curves suggested that there was a difference between surgeons, with starting values ranging from 103.05 to 121.57 minutes per operation and rates of learning from –0.08 to 0.14 minutes per operation. These differences were not tested for statistical significance.

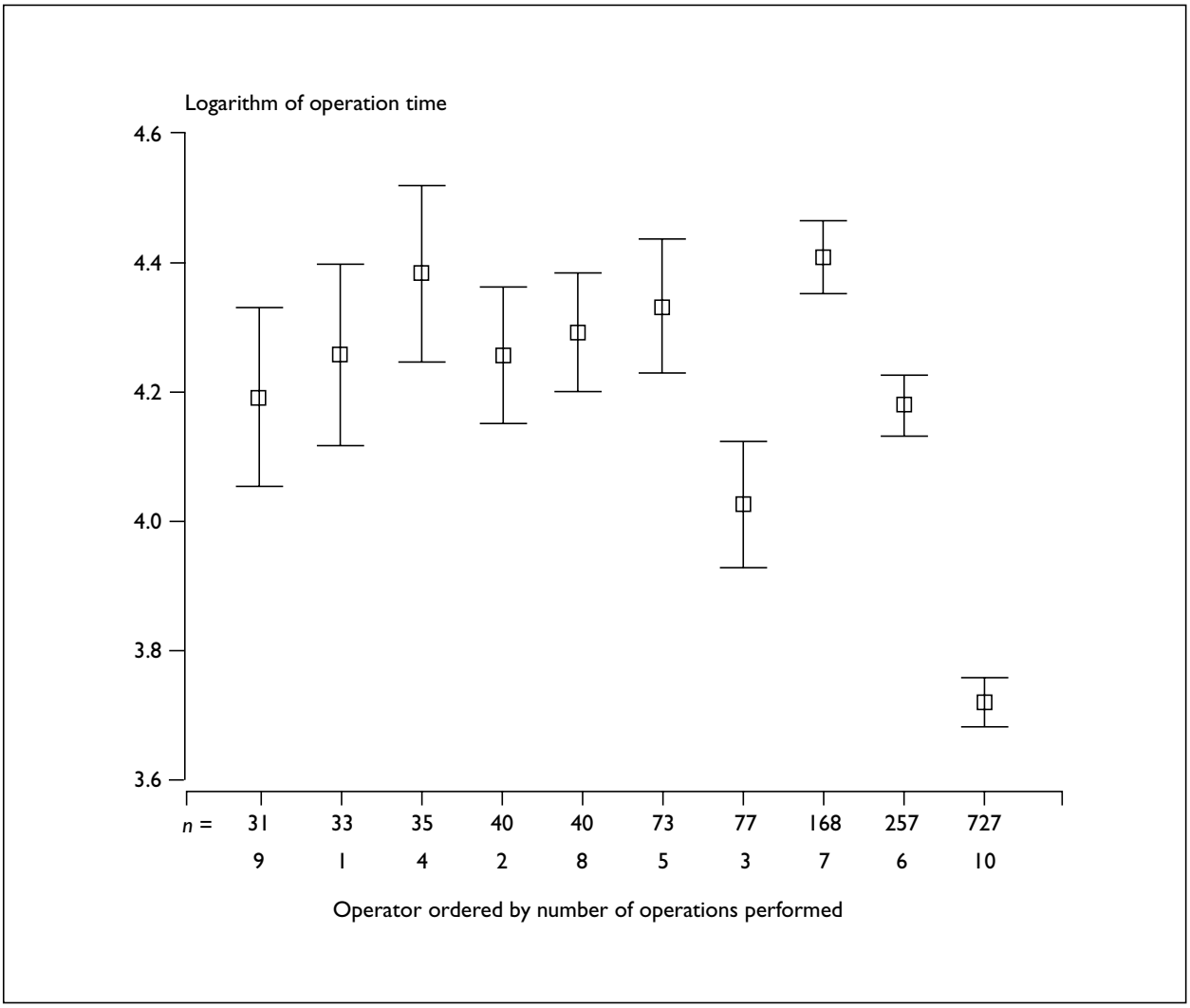


FIGURE 12 *Error bars (95% CI) of logarithms of operation time for each surgeon*

It was possible that these differences may have been an artifact of the varying number of procedures per surgeon. For example, in chapter 4 it was demonstrated that these curves often could not fit the initial values accurately. Hence, the first 150 cases only for each surgeon were selected and these results are also shown in *Table 14* with the predicted curves shown in *Figure 14 (b)*. Again, there was no apparent difference between the different slopes of curves, and the difference in predicted operation times between surgeons persisted. There was a suggestion that as the starting value (parameter 'a' in the power law) decreased, the rate of learning increased. The multilevel modelling procedure was used to investigate these issues further.

Techniques for complex structured data

Multilevel models

Initially, the total variability in the data set was partitioned between and within operators. The intraclass correlation coefficient was 0.15 and, hence, 15% of the total variability in operation times was explained by between-operator differences.

The results of the multilevel modelling, using the logarithm of operation sequence (LNSEQ) as the only covariate, are presented in *Table 15*. The reduction in log-likelihood from a model that did not include LNSEQ to one that included LNSEQ with random intercepts was highly statistically significant. The subsequent reduction in log-likelihood for the random starting levels and random rate of learning model was also significant at the 5% level and, hence, there was evidence that the rate of learning and the starting level were different for each surgeon.

The parameter estimates of the random intercept and slopes model are shown in *Table 15*. The predicted deviations from the overall rate of

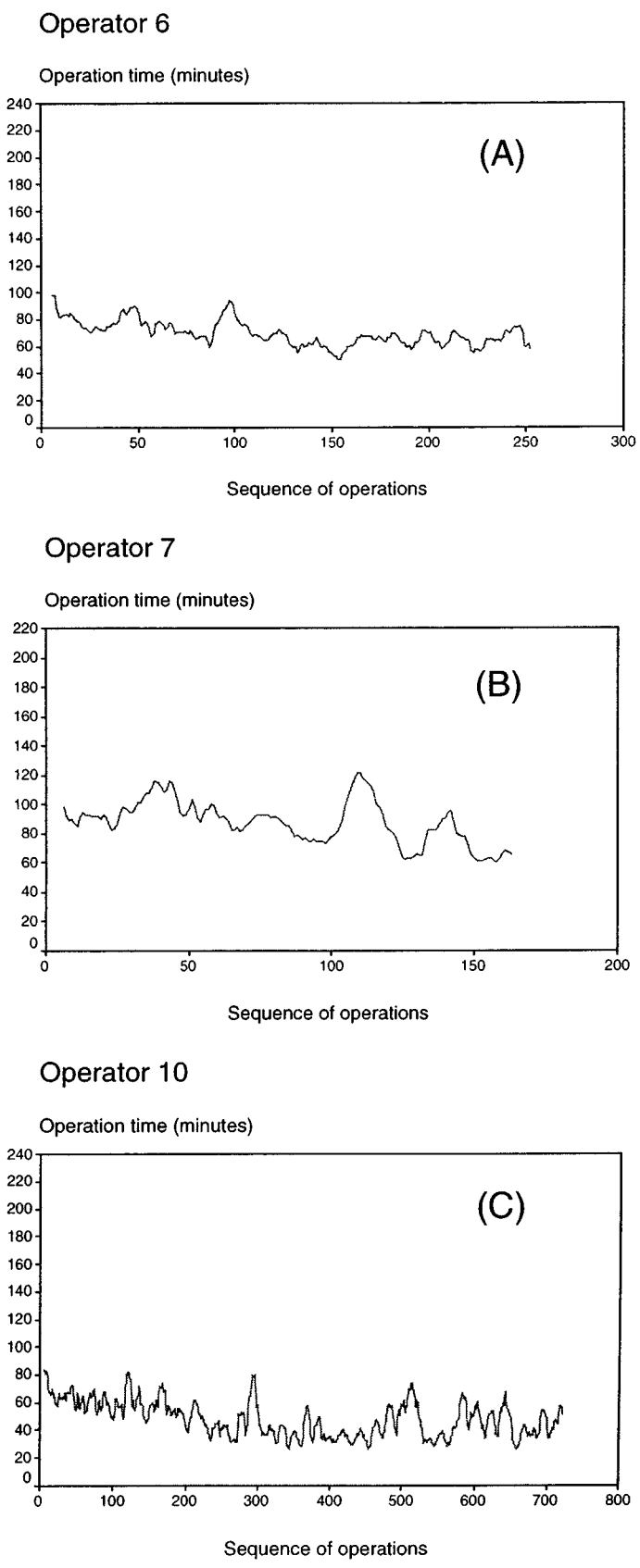


FIGURE 13 *Moving averages (order 10) for three surgeons*

TABLE 14 *Comparison of curve fits for three surgeons*

Operator	Number of cases, *n*	Logarithmic			Power			Log–linear		
		a	b	R^2 (%)	a	b	R^2 (%)	a	b	R^2 (%)
All cases										
6	(257)	108.12	–8.29	7.9	113.95	–0.11	7.9	4.75	–0.02	7.9
7	(168)	120.23	–7.82	4.9	121.57	–0.08	4.7	4.80	–0.02	4.7
10	(727)	92.31	–7.86	5.7	103.05	–0.14	5.5	4.66	–0.03	5.5
First 150										
6	(150)	108.95	–8.48	7.7	112.99	–0.10	7.7	4.74	–0.02	7.7
7	(150)	113.78	–5.79	2.6	114.81	–0.06	2.6	4.75	–0.01	2.5
10	(150)	84.61	–5.48	2.1	86.31	–0.08	2.1	4.46	–0.02	2.1

learning and overall starting values for each surgeon were estimated for this model and the corresponding 95% CIs for these deviations were calculated. These deviations are displayed in *Figure 15*. Differences between the starting levels for each surgeon appear less marked than the differences between the rates of learning, particularly between surgeon 10 and surgeons 5, 6 and 7.

The starting value/rate of learning covariance (σ_{uv}) was positive and demonstrated that, as the starting value increased, the rate of learning decreased (since the higher the negative rate, the faster the surgeon is learning). This relationship was also suggested by the earlier curve fits in the previous section. The statistical evidence for this was weak.

Multilevel models with additional covariates

The random intercepts and slopes model was extended to include a number of covariates. These covariates were included as fixed effects only. The results of this model are shown in *Table 16*. All of the fixed-effect covariates were statistically significant and improved the model fit. For example, an operative cholangiogram on a patient would increase the logarithm of operation time by 0.50. There were no resulting significant changes in the parameter estimates for the random effects and variances/covariances.

This model can be used to predict the mean operation times for a patient with certain characteristics. For example, a male patient who had had a preoperative sphincterotomy, and who was the tenth patient on whom that surgeon had performed the procedure, would have a predicted logarithm of operation time of:

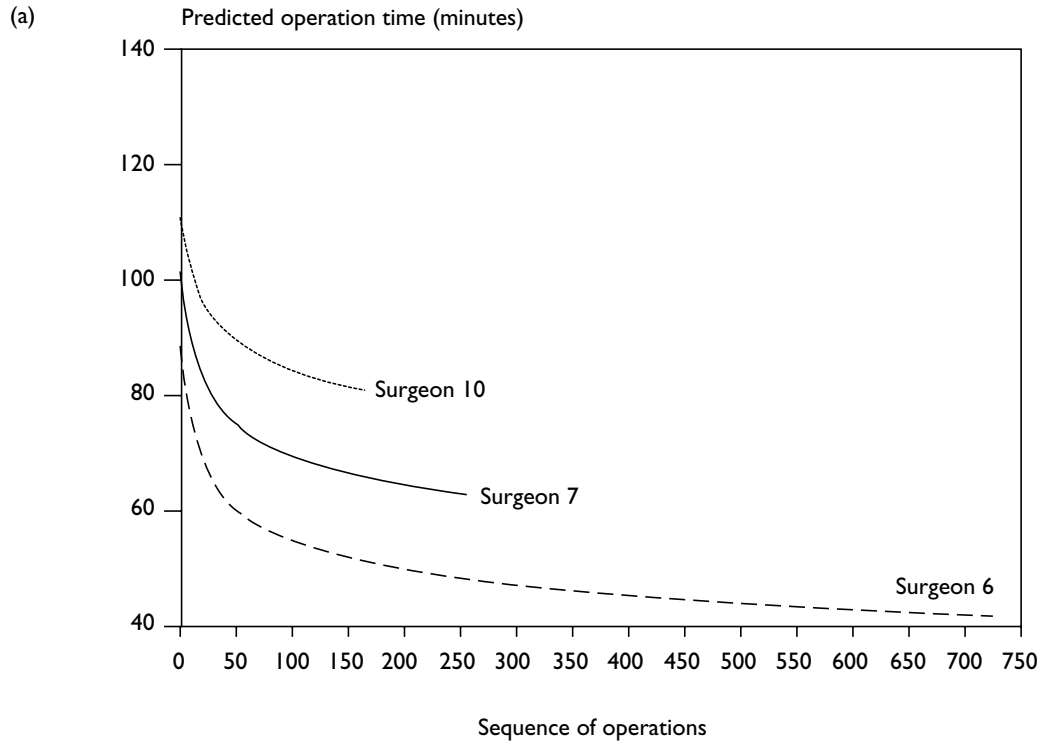
$$4.428 + \log_e(10) \times (-0.101) + 0.107 + 0.164 = 4.466 \text{ (or 87 minutes)}.$$

However, if this was the 200th procedure performed by that surgeon, then the predicted logarithm of operation time is:

$$4.428 + \log_e(200) \times (-0.101) + 0.107 + 0.164 = 4.164 \text{ (or 64 minutes)}.$$

TABLE 15 *Multilevel modelling using logarithm of operation sequence only*

Parameter	Estimate (95% CI)	
α_0, overall starting level	4.587 (4.476 to 4.698)	
β, overall rate of learning	–0.107 (–0.142 to –0.072)	
σ_u^2, variance of starting level between surgeons	0.005 (–0.019 to 0.029)	
σ_v^2, variance of rate of learning between surgeons	0.001 (–0.001 to 0.003)	
σ_e^2, starting level component of patient variance	0.197 (0.182 to 0.211)	
σ_{uv}, starting level/rate of learning covariance at patient level	< 0.001 (–0.006 to 0.006)	
	Reduction in –2 × log (likelihood)	**Significance**
Rate of learning constant, starting level random	–	–
Rate of learning fixed, starting level random	113	$p < 0.001$
Rate of learning and starting level random	6.5	$p < 0.05$



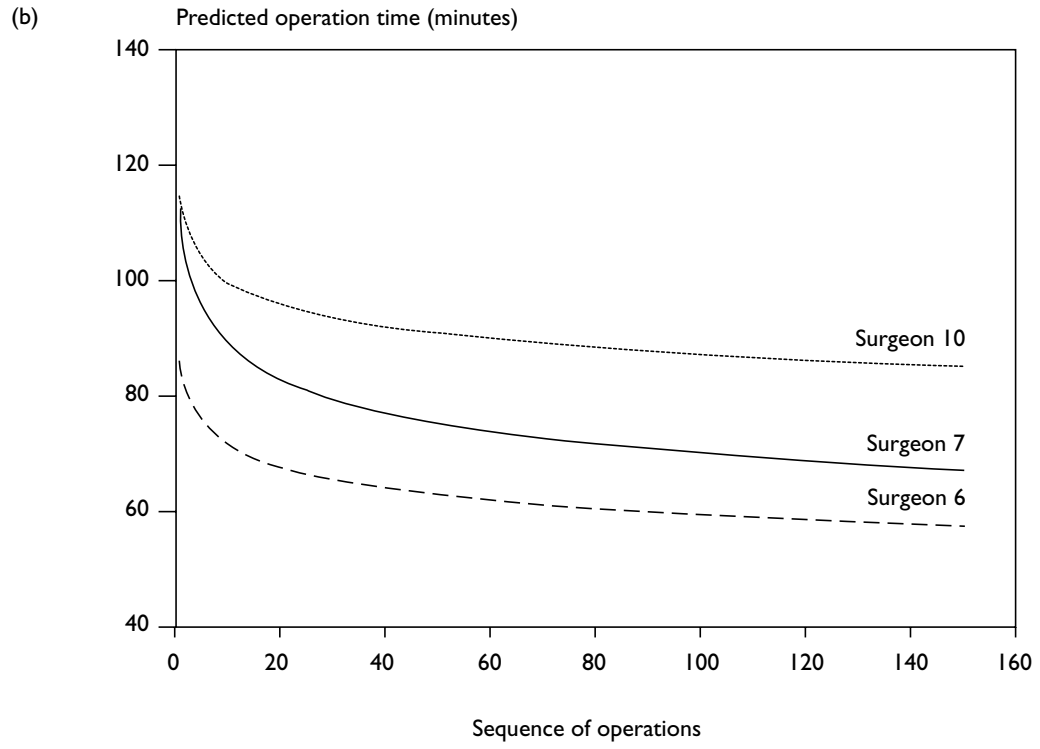


FIGURE 14 *Predicted operation times for three surgeons based on the power curve model*

TABLE 16 *Extension of the single covariate multilevel model to include other fixed effects*

Parameter	Estimate (95% CI)
α_0, overall starting level	4.428 (4.296 to 4.561)
β, overall rate of learning	–0.101 (–0.139 to –0.063)
σ_u^2, variance of starting level between surgeons	0.021 (–0.017 to 0.058)
σ_v^2, variance of rate of learning between surgeons	0.001 (–0.001 to 0.004)
σ_e^2, starting level component of patient level variance	0.150 (0.139 to 0.161)
σ_{uv}, starting level/rate of learning covariance at patient level	–0.003 (–0.012 to 0.006)
Fixed effects	
Male	0.107 (0.060 to 0.154)
Sphincterotomy	0.164 (0.079 to 0.248)
Gall bladder ruptured	0.130 (0.081 to 0.179)
Operative cholangiogram	0.499 (0.425 to 0.572)
Gall bladder inflamed	0.294 (0.242 to 0.345)

Discussion

Statistical methods to identify and measure learning curve effects have been explored using multiple case series of laparoscopic cholecystectomies. It has been demonstrated that some of the methods identified in the systematic searches described in chapters 2 and 3 can be extended to more complex structured data sets such as these.

Some of the simple series techniques can be used as graphical aids to investigate differences between operators and to check for trends over time. It is recommended that box plots, moving averages (continuous data) and cusum charts (dichotomous data) are applied to the data where applicable. These methods are useful for identifying if learning curve effects are present in the data.

For comparative purposes, some of the shapes used in chapter 4 were also fitted. These shapes did not describe the change in operation times for cholecystectomy as well as they did for fundoplication (R^2 values: approximately 7% for cholecystectomy; 20% for fundoplication). This apparent discrepancy illustrates two points. First, there is a difference between different health technologies. Second, the same proxy for learning (operation time in this example) does not necessarily have the same strength of effect in other technologies. Nevertheless, investigators should try fitting some of the learning curves described in chapter 4. The results aid the multilevel modelling. For example, no great difference was observed between the various shapes of curve and so a simple power curve was used in the multilevel modelling.

The multilevel modelling gave an insight into the variability in the cholecystectomy data set that ordinary regression techniques could not. Differences between surgeons' rates of learning and starting levels between surgeons were considered as random effects around the population average. This implied that there was an underlying true population and that surgeons deviated from this true value by some random distribution. From this information it was possible to derive slope and starting value residuals for each surgeon. This illustrated that there were differences between surgeons.

The rate of learning residuals disentangle the proportion of total variability that can be attributed to true variation in rates of learning from surgeon to surgeon from that proportion which can be attributed to random variation between patients. In other words, this technique can separate variation due to the surgeons being different from variation caused by the surgeon having more extreme cases. The differences in these rates of learning residuals should be smaller than those obtained from a standard multiple regression analysis carried out on each surgeon separately, because other sources of variation have been taken into account.[83]

Multilevel modelling can also estimate the correlation between the residuals for rate of learning and starting value. There was a weak positive correlation between these two sets of residuals. This means that as the starting value decreased the rate of learning got faster. This is consistent with surgeons with more natural aptitude for laparoscopic surgery beginning faster and learning faster than those with less natural aptitude. This is one of many possible explanations that require testing in a larger data set.

Multilevel models can easily be extended to include other covariates as fixed or random

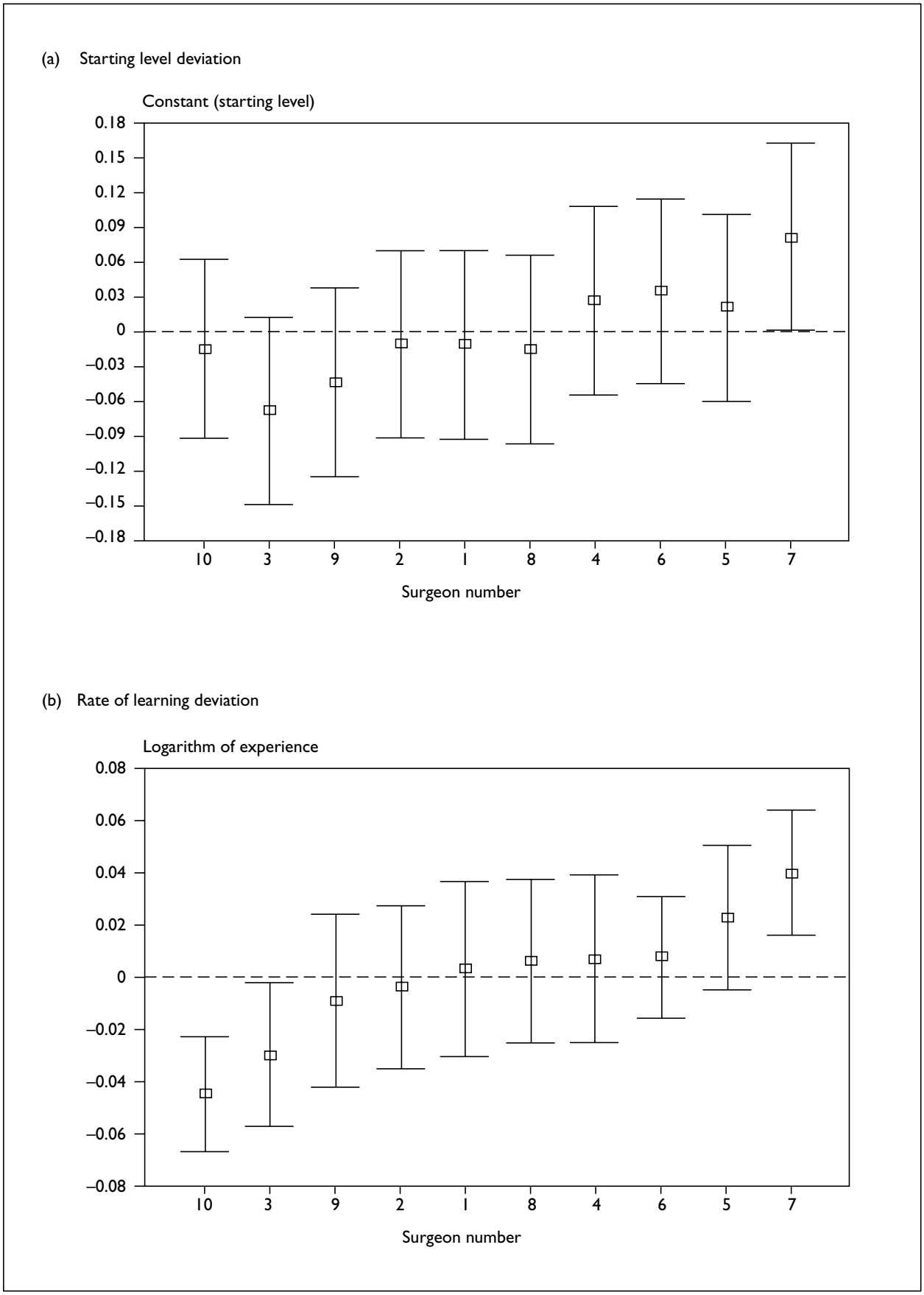


FIGURE 15 *Predicted deviations and 95% CIs from the overall mean intercept (a) and mean rate of learning (b) for all surgeons, based on the random slopes and random intercepts model*

effects. Multilevel models can also test for interaction, explore more complex variation such as autocorrelation, and use other estimation procedures such as Markov chain Monte Carlo methods.[93] The authors did not have time within the constraints of this project to explore these.

If the proxy for learning is a dichotomous outcome, then multilevel models can still be applied to the data. This is the great advantage of this method and one of the stronger arguments for using it. Multilevel models also allow a natural extension to the study of institutional learning effects. If a data set covers many institutions and different operators within each institution, then the interaction between institutions and operators could also be investigated.

The other statistical techniques were not tested for the complex structured data that were identified in chapter 3. Further empirical work is required to assess the relative strengths and weaknesses of these techniques in comparison to multilevel modelling.

Conclusion

This chapter illustrates how relatively simple techniques can be applied to health technology data sets to investigate learning curve effects. These can be used to explore whether differences between operators do exist. The use of multilevel models enabled differences between and within operators to be described more accurately than in the existing literature. In the next chapter consideration is given to data from randomised trials with multiple operators, in which learning may have affected one or both of the interventions being compared.

Chapter 6

Multiple operators in an RCT: a case study of laparoscopic groin hernia repair

Introduction

The RCT is the standard approach to evaluating new technologies. The principle advantage is the avoidance of selection biases such as case-mix differences or changes over time. Investigators sometimes believe that their randomised trial may have been compromised or biased by learning effects in one or both arms of the trial.

The recommended method of analysis of a learning curve effect in a clinical trial is unclear; however, it will be influenced by the question being posed. If the trends over time in one arm of the trial were 'nuisance' parameters, they could be taken into account during an evaluation by fitting curves to the data as described earlier. The treatment effect could then be introduced into the model as a fixed effect. Further empirical work is required on this method of analysis using clinical trial data sets before any strong recommendation can be made. In particular, there are potential biases in trials that do not have information on every procedure that an operator has performed. If, however, the purpose were to describe the learning curve effect in one arm of the trial only, then an analysis similar to that undertaken on the laparoscopic cholecystectomy data would be appropriate.

To begin to address this issue, investigators require to identify and measure any learning effect in that arm (or arms) of the trial. Unless the procedure is limited to the trial, however, the assessment is commonly complicated by incomplete case series for the operators in the trial (not every case is included in a trial). In this chapter the analysis of complex structured data is extended to randomised trial data derived from the laparoscopic procedure arm of a multicentre clinical trial of groin hernia repair.[115] The randomised data were supplemented by information on non-randomised operations performed during the trial. The specific aims were:

(i) to consider methods for coding the experience variable
(ii) to quantify the learning curve effect under the different experience variables
(iii) to discuss the strengths and disadvantages of collection of non-randomised data in relation to learning curve effects in RCTs.

Laparoscopic groin hernia repair

Groin hernia repair is one of the most common procedures in general surgery. Laparoscopic hernia repair has been proposed as an alternative to standard open repair but, unlike laparoscopic cholecystectomy, it has been slow to gain acceptance in the surgical community. There have been reports of rare operative and postoperative complications and, because the procedure is technically demanding, several studies have described a long learning curve that included a high failure rate while surgeons were in this 'learning phase'.[115–120]

Typically, the variables used to assess the learning curve of laparoscopic hernia repair have been operation time, recurrence rate and conversion to open surgery. There have been no firm recommendations on the number of procedures that should be performed before the learning curve is ascended. Liem and colleagues[116] indicated that most of the recurrences in their studies were among patients treated by a surgeon with 'limited experience' but do not go on to say how 'limited' was defined. Champault and colleagues[118] performed 50 procedures before including patients in their study, and the MRC Laparoscopic Groin Hernia Trial Group[115] required surgeons to have performed ten procedures before they could randomise patients. Again, there was no rationale given for the number of procedures chosen as the learning phase.

Methods

The laparoscopic hernia data set

The data set used in the analyses for this chapter had two components. The first was 421 patients from the laparoscopic arm only of a pragmatic multicentre, randomised, controlled comparison of open versus laparoscopic hernia repair. Patients were recruited between January 1994 and March 1997. All of the surgeons had experience of at

least ten procedures and, if they felt they were still learning the technique, they received additional training from an experienced surgeon. The second component of the data set was a prospective clinical audit of all non-randomised laparoscopic cases (281 in total) performed by each surgeon over the same period. Thus, there were details on both randomised and non-randomised cases.

Information was collected prospectively on standard forms to record patient demographic characteristics, operative and postoperative measures. The forms were identical for the randomised and non-randomised cases.

Coding of experience variable

The experience variable was coded in three ways, as illustrated in *Figure 16*. The 'true' experience is denoted by the position within the case series of all cases (randomised and non-randomised). Hence the first case was coded as 1, the second as 2, and so on. If analyses of only the randomised cases were available (denoted by the arrows in *Figure 17*), the coding of experience was performed in two ways. First, the sequence number was denoted by the position of the randomised case within the case series of all cases: for example, the second and fifth laparoscopic procedures were included in the trial, so the trial cases sequence numbers associated with these procedures were 2 and 5. Second, the sequence number was coded as the order that patients were recruited into the trial. Using the example above, the coding would be 1 and 2 instead of 2 and 5. This was the closest proxy for experience when no information on non-randomised cases was available.

Proxies for learning

Two possible proxies for learning were investigated – operation time (continuous) and complications at 1 week (dichotomous). Operation time was the time taken from first incision to last stitch. Complications at 1 week included wound complications such as bruising, scrotal complications such as hydrocele, special complications such as nerve injury, and general complications such as pulmonary embolism.

Statistical analysis

Preliminary investigation of data

Preliminary investigation of the two outcomes involved graphical and tabular techniques as described in chapter 4.

Multilevel modelling

The relative impact of the three ways of coding the experience variable on operation time and complications was investigated using multilevel models (as described in chapter 5). Because of convergence problems, the analyses for operation time and experience had to be performed after logarithmic transformation.

The equation for the fitted model is:

$$y_{ij} = \alpha_{0j} + \beta_{0j} x_{ij} + e_{ij} \qquad \text{Equation 14}$$

where y_{ij} is the logarithm of operation time and x_{ij} is the logarithm of the sequence number; $\alpha_{0j} = \alpha_0 + u_j$ and $\beta_{0j} = \beta_0 + v_j$, where u_j and v_j have mean zero and variance of σ_u^2 and σ_v^2, respectively, and are the departure of the j-th operator from the true population rate of learning β_0. To specify the model fully:

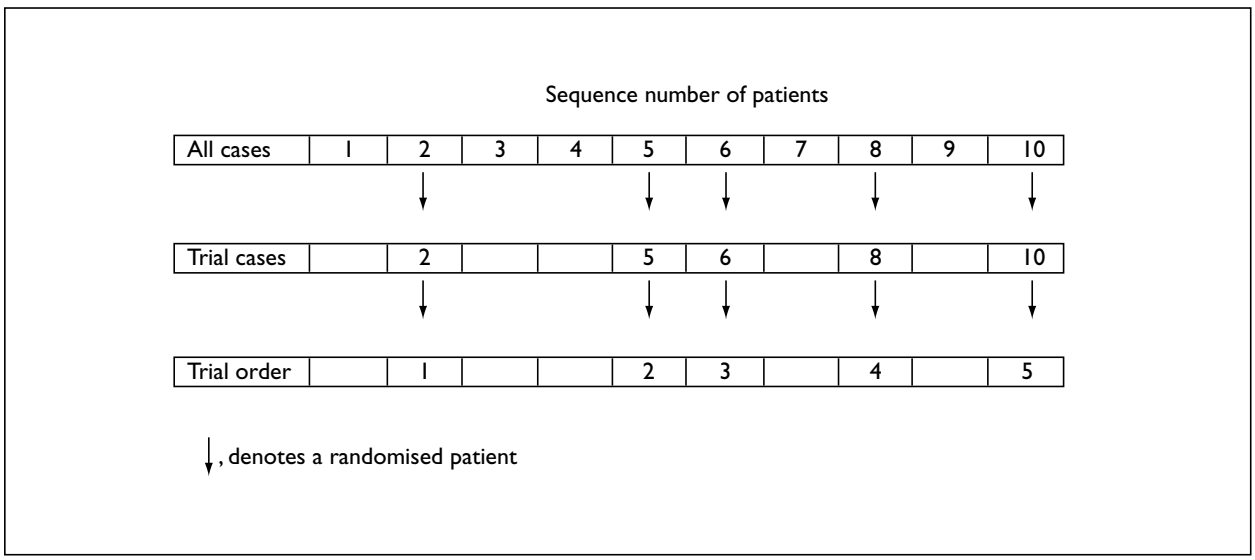


FIGURE 16 *Sequence numbers used in the analyses*

$E(\alpha_{0j}) = \alpha_0$; $E(\beta_{0j}) = \beta_0$; $\text{var}(\alpha_{0j}) = \sigma_u^2$; $\text{var}(\beta_{0j}) = \sigma_v^2$; $\text{cov}(\alpha_{0j}, \beta_{0j}) = \sigma_{uv}$; and $\text{var}(e_{ij}) = \sigma_e^2$.

Consultant surgeons did not undertake all the procedures. In some cases, a senior registrar performed the operation with the consultant in attendance as the assistant. To investigate the possible impact of grade of surgeon upon the outcomes, a fixed effect for grade of operator was entered into the multilevel models.

Results

Operation time

The 27 surgeons performed 421 randomised laparoscopic hernia repairs in total, with the numbers of randomised procedures performed individually ranging from 1 to 149. One surgeon (number 3) performed 35% of the randomised cases. By adding the non-randomised cases to the randomised cases, the total number of laparoscopic procedures increased from 421 to 702. Senior registrars performed approximately 28% of the procedures.

The number of patients for each surgeon and the corresponding mean operation times for laparoscopic repair are shown in *Table 17*. The mean operation times for the randomised cases only and for all cases showed that operation times varied between surgeons. The mean operation time varied within surgeon for some surgeons when comparing the randomised cases to all cases. For example, for surgeon 14 the mean operation time for the randomised cases was 64.9 minutes and for all cases was 74.8 minutes. However, the majority of the operation times varied only marginally between randomised and all cases.

To explore the position of randomised and non-randomised cases within a case series, operation time versus sequence number was plotted for three surgeons (*Figure 17*). For surgeon 7, the non-randomised cases appeared to follow a

TABLE 17 *Comparison of mean operation times for randomised and all patients*

Operator number	Randomised cases		All cases	
	Number of patients	Mean time (minutes)	Number of patients	Mean time (minutes)
1	2	75.0	27	68.4
2	5	66.0	21	72.0
3	149	51.3	153	51.2
4	41	56.2	41	56.2
5	4	87.5	4	87.5
6	2	81.0	26	81.7
7	53	66.9	109	69.9
11	3	86.7	3	86.7
13	15	73.7	36	79.6
14	15	64.9	31	74.8
15	5	72.8	19	69.8
16	18	66.1	85	65.9
21	13	66.2	24	62.3
24	4	74.5	5	71.0
26	4	45.0	6	50.8
29	9	78.3	10	78.5
30	3	48.3	23	38.7
31	7	54.3	7	54.3
33	8	60.0	8	60.0
34	5	57.6	6	58.0
36	15	36.9	15	36.9
38	5	98.0	6	96.7
39	1	–	1	–
43	8	65.5	9	62.7
46	5	54.8	5	54.8
47	17	52.4	17	52.3
48	5	83.0	5	83.0
Total	**421**	**59.2**	**702**	**63.2**

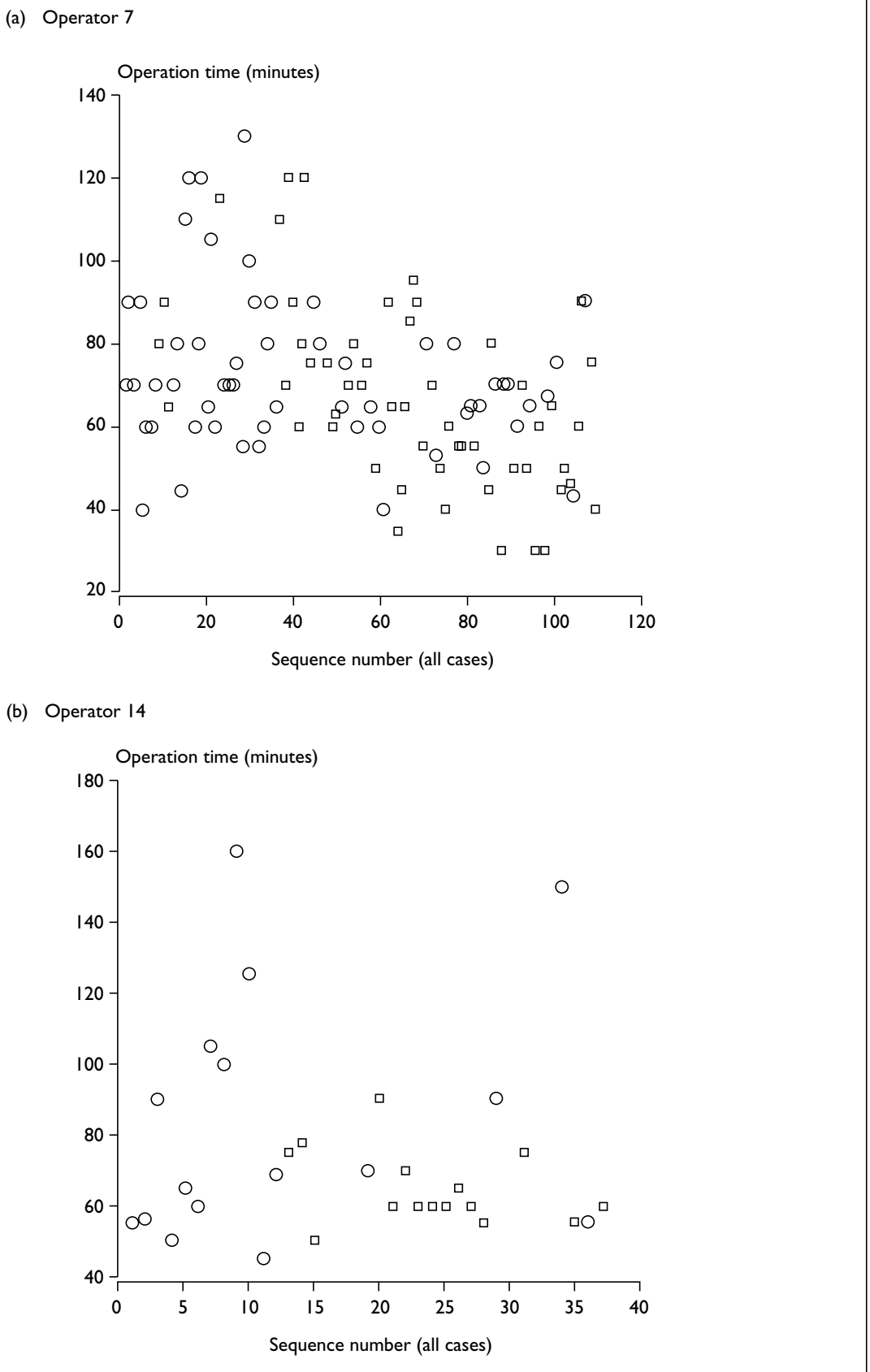


FIGURE 17 *Distribution of operation times for three operators using randomised (□) and non-randomised (○) data*

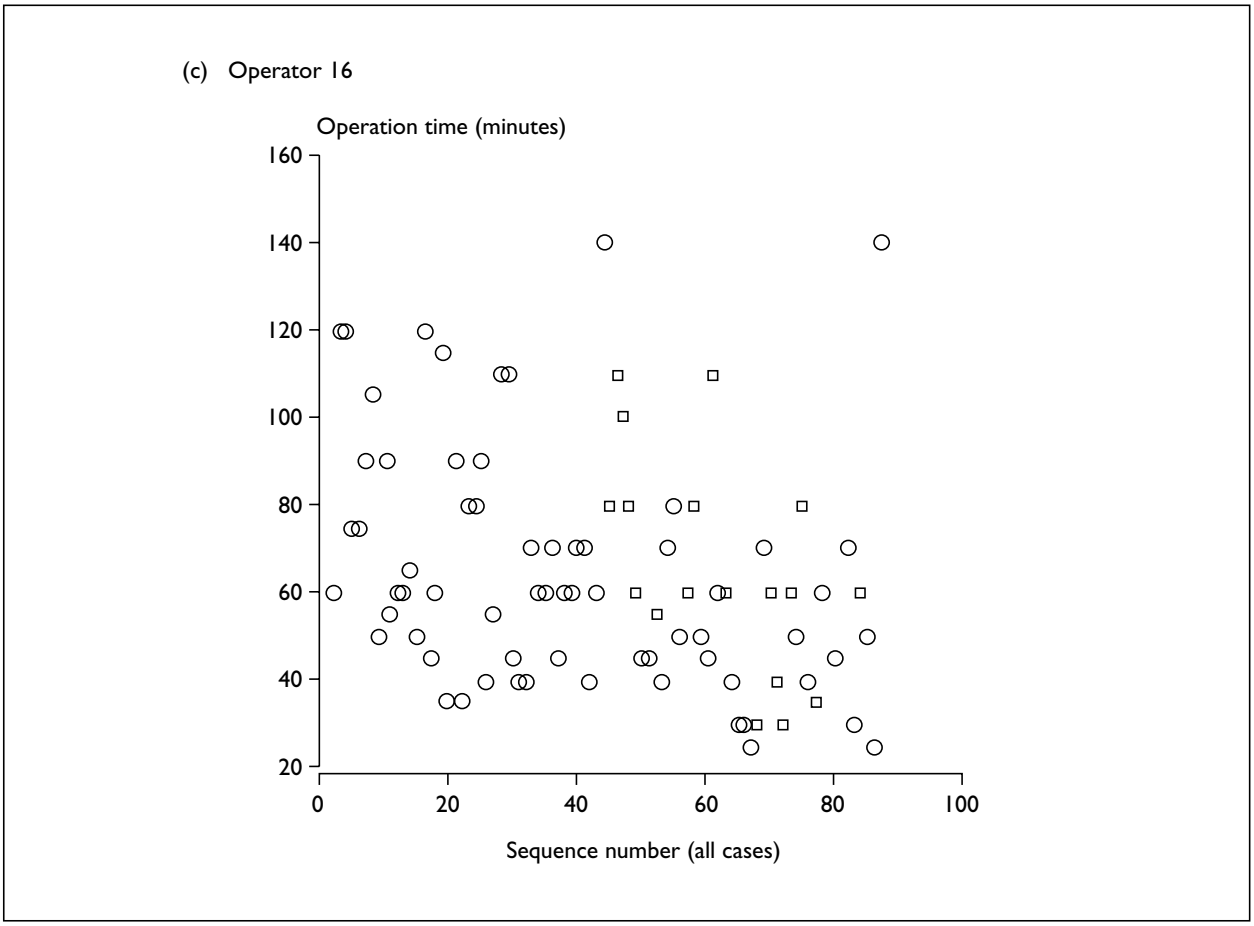


FIGURE 17 contd *Distribution of operation times for three operators using randomised (□) and non-randomised (○) data*

similar pattern to the randomised cases. That is, the operation times appeared to decrease at approximately the same rate for both types of data. For surgeon 14, the variability between non-randomised cases was greater than for the randomised cases. Nevertheless, there was no discernible change in mean operation time with experience for surgeon 14. In contrast, surgeon 16 showed a decrease in operation time with experience but the rate of decrease was greater in the randomised cases compared with non-randomised cases.

The intraclass correlation coefficient was 0.24 for the randomised cases and 0.29 for all cases. Hence, there was considerable variability between surgeons.

The results of the multilevel modelling of logarithm of operation time and logarithm of sequence for each of the three methods of coding sequence are shown in *Table 18*. There was strong evidence ($p < 0.001$) that random starting levels were preferred over fixed starting levels for all three models. However, there was no statistical evidence that random rates of learning were applicable for the all cases coding. There was evidence that a random rate of learning was preferred to a fixed rate for the trial order coding. The multilevel model for a random rate of learning using the trial cases coding would not converge to a solution and, thus, fixed rates are shown in *Table 18*.

The parameter estimates for the overall starting levels and variances were similar for all three models. The estimates of the rates of learning differed between models – the trial order rate of learning was higher than either the trial cases or all cases coding. Based on these estimates, the predicted mean values (transformed back to the original scale using the antilogarithm function) were calculated for different levels of experience; these are displayed in *Table 19*. The predictions for the first case were 70.0, 71.7 and 70.5 minutes, respectively, for the trial order coding, trial cases coding and all cases coding. By the 200th procedure, the estimates were 44.1, 48.7 and 49.1 minutes, respectively. The predicted fall in operation time over the first 50 cases was greatest for the trial order coding (20.2 minutes).

TABLE 18 *Operation time: parameter estimates of the models using the three different sequence numbering*

Parameter	All cases[a] estimate (95% CI)	Trial cases[a] estimate (95% CI)	Trial order[b] estimate (95% CI)
α_0, starting level	4.255 (4.147 to 4.363)	4.273 (4.150 to 4.396)	4.248 (4.132 to 4.364)
β, rate of learning	–0.068 (–0.095 to –0.041)	–0.073 (–0.110 to –0.036)	–0.087 (–0.111 to –0.063)
σ_u^2, variance of starting level between operators	0.047 (0.016 to 0.078)	0.040 (0.011 to 0.069)	0.063 (0.014 to 0.112)
σ_v^2, variance of rate of learning between operators	– (–)	– (–)	0.002 (< 0.001 to 0.004)
σ_e^2, patient level variance	0.110 (0.098 to 0.122)	0.109 (0.093 to 0.125)	0.105 (0.091 to 0.119)
σ_{uv}, covariance between starting level and rate	– (–)	– (–)	–0.012 (–0.022 to –0.002)

[a] *Log sequence modelled as a fixed effect (random effect was non-significant)*
[b] *Log sequence was modelled as a random effect (X^2 difference = 8.5; p < 0.025)*

TABLE 19 *Predicted mean operation times*

	Operating time (minutes)				
Sequence number	**1**	**25**	**50**	**100**	**200**
Trial order[a]	70.0	52.9	49.8	46.9	44.1
Trial cases[b]	71.7	56.7	53.9	51.3	48.7
All cases[c]	70.5	56.6	54.0	51.5	49.1

[a] *Prediction equation was* optime = $e^{4.248}$ *x (order)*$^{-0.087}$
[b] *Prediction equation was* optime = $e^{4.273}$ *x (order)*$^{-0.073}$
[c] *Prediction equation was* optime = $e^{4.255}$ *x (order)*$^{-0.068}$

Surgeon number 3 performed over one-third of all cases; in order, therefore, to test the sensitivity of the results, this surgeon was excluded from the data and the models were refitted. There was no difference in the conclusions derived from these analyses. For example, the overall starting level changed from 4.248 to 4.257 and the overall rate of learning changed from –0.087 to –0.092 for the trial order coding.

Grade of surgeon (coded 1 for senior registrar, 0 for consultant) was entered into the three models as a fixed effect. There was no significant reduction in the log-likelihood for any of the models when this variable was entered and, hence, there was no evidence that grade of surgeon affected the operation time. For example, for the all cases coding, the point estimate for the effect of a senior registrar performing the procedure was –0.012 (95% CI, –0.078 to 0.054).

Any complication at 1 week

There were some patients for whom no information on complications at 1 week was recorded. Data for only 18 surgeons performing 341 randomised laparoscopic hernia repairs in total were included, and 586 repairs were included when the non-randomised cases were added.

The number of patients for each surgeon and the corresponding proportion of those patients who had a complication are shown in *Table 20*. The complication rates for the randomised cases only and for all cases varied between surgeons but were reasonably consistent within surgeon. The intra-class correlation coefficients were 0.054 and 0.089, respectively, for trial cases and all cases. So the variation between surgeons was about 5–9% of the total variability in the data set. The intraclass correlation coefficient for these binary outcomes were calculated from an ANOVA table under the assumption of normality.

The overall complication rates within the randomised cases only were divided into sequential groups of ten patients. The mean and corresponding 95% CIs were calculated, and are displayed in *Figure 18*. There was no detectable increase or decrease in the overall complication rate with experience. The figure may, however, have hidden

TABLE 20 *Comparison of 1-week complication rates for randomised and all patients*

Operator number	Randomised cases		All cases	
	Number of patients	Percentage with complications	Number of patients	Percentage with complications
1	1	0	13	46
2	5	40	13	31
3	146	47	150	45
4	41	63	41	63
5	4	50	4	50
6	2	50	26	58
7	53	38	109	34
11	3	0	3	0
13	15	40	36	50
14	19	32	33	27
15	2	50	18	33
16	18	61	86	46
21	13	38	24	38
24	3	33	5	20
25	0	0	6	17
26	4	50	6	67
29	9	56	10	50
36	3	67	3	67
Total	**341**	**46**	**586**	**43**

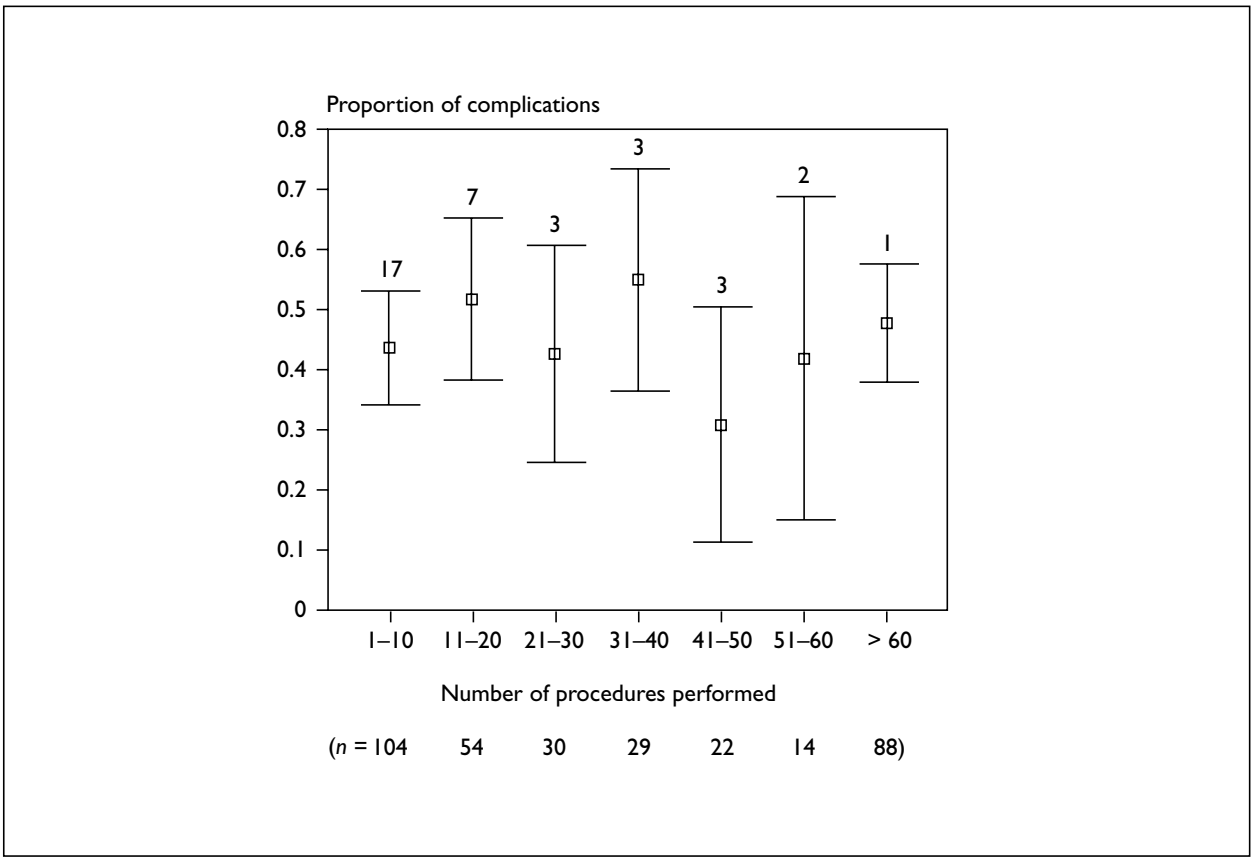


FIGURE 18 *Distribution of complications rate and corresponding 95% CIs for randomised cases only (numbers above the 95% CIs denote the number of operators contributing cases to that group)*

changes in individual performance. Cusum charts for the three operators who contributed most of the randomised cases are shown in *Figure 19 (a–c)*. The cusum charts were random for all three operators. The apparent increase relating to operator 4 means that that operator was not attaining a 55% rate of complication; however, the randomness of the deviations as the cusum increases means that the complication rate did not change significantly with experience.

The results of modelling the complication rate by the three methods of coding experience are shown in *Table 21*. In all three cases, there was no statistically significant relationship between complication rate and experience (although the direction of effect suggested that complication rate **increased** with experience). The overall mean predicted probabilities of a complication were 39%, 45% and 45% for all cases, trial cases and trial order variables, respectively. Again, the effect of inclusion of grade of surgeon as a fixed effect in the models was not significant.

Discussion

The analysis of multiple case series learning curve data using multilevel models was extended to the investigation of a randomised cohort of cases embedded within a case series. It was demonstrated that some parameters of learning curve effects were changed by the inclusion of the non-randomised data.

Coding of the experience variable

One of the critical components when assessing the learning curve effect in a randomised cohort of patients is attributing a level of experience to each of the randomised patients. Three different sequences for experience were applied to the laparoscopic hernia data. The first ignored the fact that there were randomised patients and described the learning curve effects in terms of all patients (randomised and non-randomised). The second sequence used the non-randomised data to estimate the number of procedures performed between randomised cases and modelled only the randomised cases. The third sequence assumed that there was no knowledge of the position of each operator in their complete case series and modelled the order in which patients were entered into the trial. There were a number of assumptions underpinning the appropriateness of the various sequences.

First, non-randomised cases were assumed to be the same as randomised cases. Second, it was assumed that all non-randomised cases contributed to an operator's learning curve. When both these assumptions were made, it was shown, using multilevel modelling, that ignoring the non-randomised cases overestimated the rate of change in operation times in the laparoscopic hernia trial (*Table 19*). It is possible that the non-randomised cases were different from the randomised cases, since the non-randomised cases failed to meet the entrance criteria of the trial. This potential bias was not addressed.

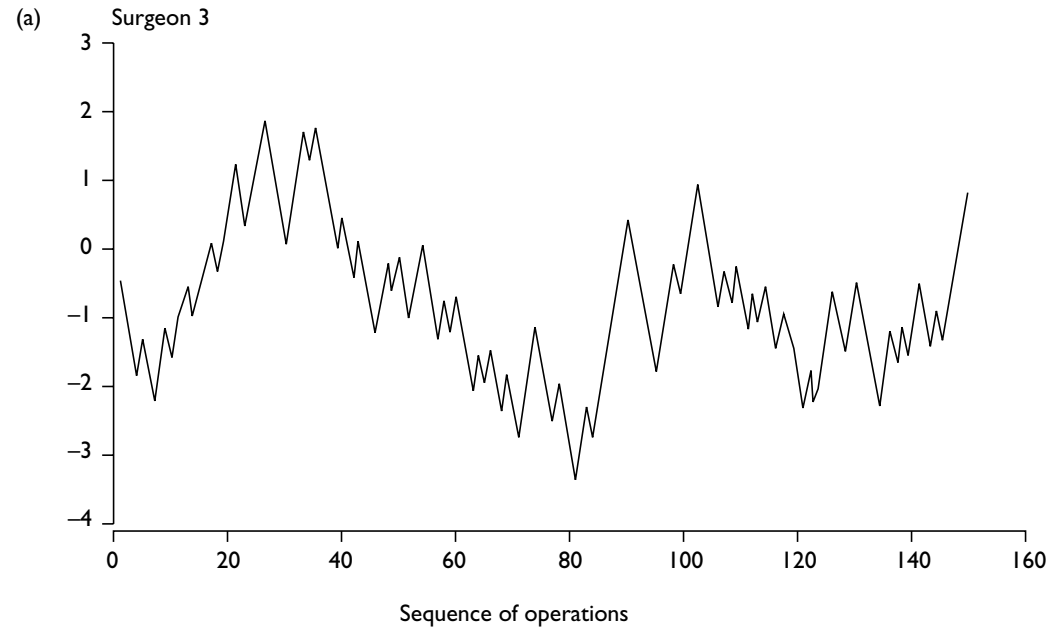
Preliminary investigation of data

The graphical and tabular representations of the randomised and non-randomised data on operation times were useful methods of describing the data. It was demonstrated that there were changes in operation time with experience but that it was not consistent across operators.

Multilevel modelling

The multilevel modelling of operation time demonstrated a statistically significant relationship between experience and operation time for the three different types of coding of experience. Experience was, therefore, potentially a confounding factor in the analysis of the randomised trial. The trial order coding of experience appeared to overestimate the rate of learning in comparison to the other two methods of coding. There was, however, great variability between operation times **within** surgeons and, thus, the 95% CIs for the rate of learning in the three models had substantial overlap. Sources of variability could have included case-mix factors such as age, sex or body mass index of the patient.

There was no evidence of a relationship between experience and 1-week complications. Operative complications may have been more likely to show a relationship with experience but there were only 25 operative complications in the hernia data set, which is too few for statistical analysis. In contrast, there were 156 complications at 1 week (but not necessarily as a direct consequence of the procedure). The intraclass correlation coefficient for the complications was between 5% and 9%. This was substantially smaller than the coefficient for operation time (24–29%). This result implies that variability between operators was less for complications. This is in agreement with research into the size of intraclass correlation coefficients in other areas.[121] Campbell and colleagues[121] demonstrated that the size of an intraclass correlation coefficient on a process of care variable (for example, operation time) was greater than the intraclass correlation coefficient of an outcome variable (for



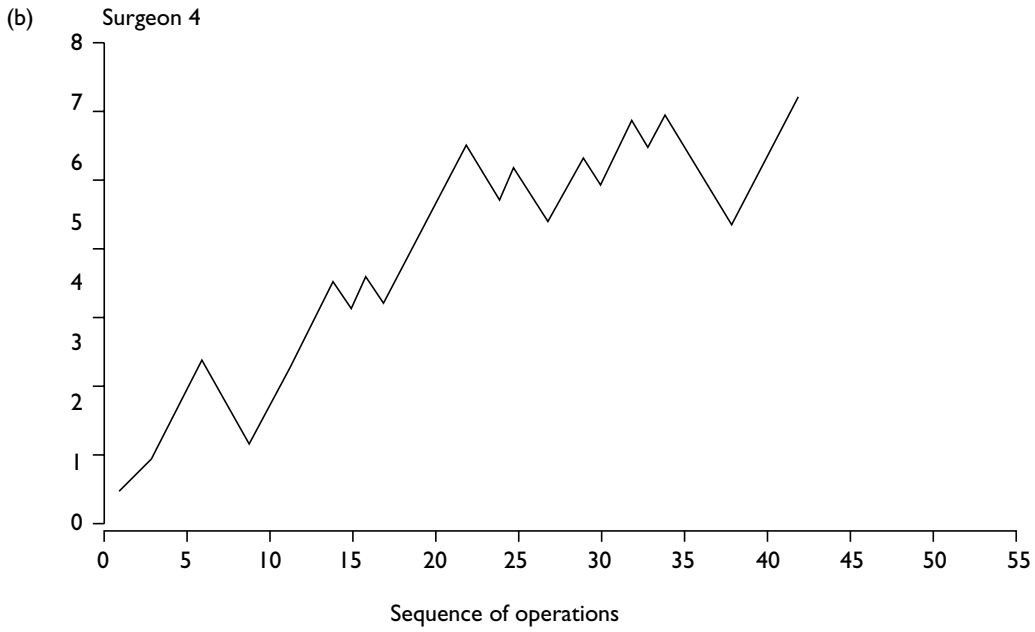


FIGURE 19 *Cusum charts of complication rates at 55% proficiency for three operators*

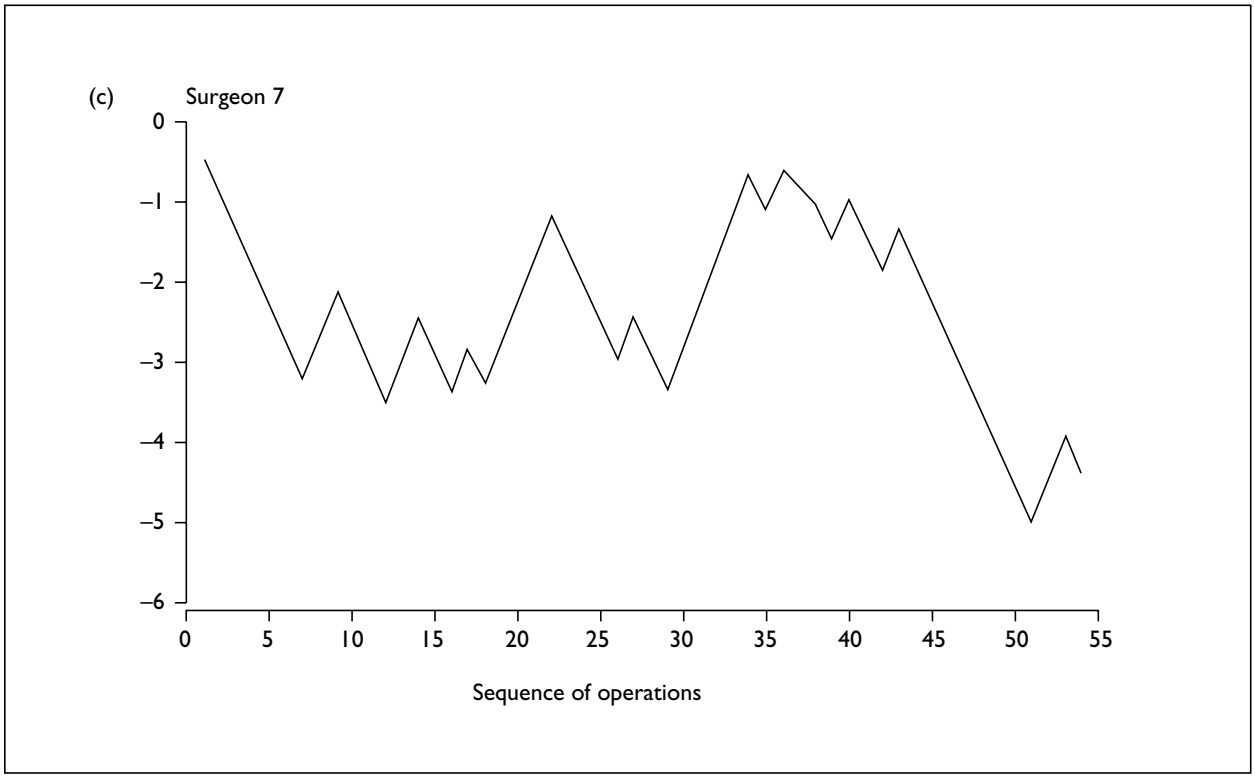


FIGURE 19 contd *Cusum charts of complication rates at 55% proficiency for three operators*

TABLE 21 *Any 1-week complication: parameter estimates of the models using the three different coding sequences*

Parameter	All cases[a] estimate (95% CI)	Trial cases[a] estimate (95% CI)	Trial order[b] estimate (95% CI)
α_0, starting level	–0.433 (–0.751 to –0.115)	–0.215 (–0.609 to 0.179)	–0.209 (–0.559 to 0.141)
β, rate of learning	0.005 (–0.001 to 0.011)	0.001 (–0.005 to 0.007)	0.002 (–0.004 to 0.008)
σ_u^2, variance of starting level between operators	0.126 (–0.066 to 0.318)	0.071 (–0.113 to 0.255)	0.070 (0.114 to 0.254)

[a] *Log-sequence modelled as a fixed effect (random effect was non-significant)*
[b] *Log-sequence modelled as a random effect (X^2 difference = 8.5; $p < 0.025$)*

example, a complication). This observation may be explained by biological variability in measures of patient outcomes compared with measures of operator performance.

A single operator performed about one-third of all cases and might have disproportionately influenced the results. This surgeon had relatively more experience with the procedure and, hence, the learning curve might have reached a plateau. In the event, however, the results of the multilevel modelling did not change substantially when the data for this surgeon were excluded.

The grade of the operator performing each procedure did not have, apparently, a great influence on outcomes. This may be because a surgical procedure such as laparoscopic hernia is a team effort or because the registrars had experience of other laparoscopic procedures. Indeed, although senior registrars performed 28% of the laparoscopic hernia procedures, there was no evidence of differences in operation times or complications between them and consultants.

Limitations of the hernia data set

The analysis of the hernia data presented in this chapter has a number of limitations. First, it is likely that some of the non-randomised cases were not reported to the study; hence the methods of coding experience may be inaccurate for some surgeons. Secondly, no attempt was made to

identify how many procedures each surgeon had performed before recruiting patients; hence, the assumption of a starting experience number of one in the sequences was also potentially misleading.

Comparison with laparoscopic cholecystectomy results

The results of the multilevel modelling of operation time with the hernia repair data can be contrasted with those obtained from the laparoscopic cholecystectomy data (see chapter 5). The estimated operation time for the first procedure was lower for the laparoscopic hernia data (70 minutes) than for the laparoscopic cholecystectomy data (98 minutes); however, the rate of learning was quicker in the laparoscopic cholecystectomy data. Other researchers have suggested that laparoscopic hernia repair typically takes longer than 70 minutes.[120] This may support the hypothesis that these hernia surgeons were further along their learning curve at the beginning than the data suggested. Indeed, each surgeon should have performed at least ten laparoscopic hernia procedures before recruiting patients. The apparently slower rate of learning in the hernia data may therefore have been caused by the surgeons being further along their learning curves, with less improvement available for them to make.

Implications for the design of trials that include possible learning effects

These findings indicate that non-randomised data should be collected in a trial if learning curve effects are to be investigated. Unfortunately, this could potentially make the costs of running the trial prohibitively high or adversely affect trial recruitment. A simpler alternative is to identify the number of procedures between randomised patients (but not collect outcome data for them). This can be done in several ways. First, for each recruited patient the operator could be asked how many previous procedures they have performed. This is certainly one of the easiest methods of collecting previous experience information but has questionable validity because of possible recall bias. Second, the hospital records (or general practioners' notes if the treatments are in primary care) of all patients with the condition can be checked. This is resource intensive. Third, if there are computerised databases available at institutional or national level, they can be a cost-effective source of information provided that the records are up-to-date and accurate.

A further advantage of adding non-randomised data is that they increase the sample size of the study and, hence, may enable the learning effects of relatively rare events to be incorporated into any analysis.

Implications for the analysis of learning in trials

As described in chapter 5, multilevel modelling is a promising approach to addressing the learning curve issue in case series type data. If possible learning effects are to be investigated in future trials, it would be helpful if further comparisons could be made of the three methods of coding of experience illustrated in this chapter. The extension of this type of analysis to other health technology data sets is required before recommendations on addressing learning curves in randomised trials can be made.

It is sometimes not possible to use multilevel modelling. There can be problems with convergence to a meaningful solution. For example, the analyses here used ten operators to investigate between-operator differences but often 20–30 are recommended.[85] The minimum number will depend upon the context. The iterative estimation technique can also be problematic. In the case study in this chapter, the trial cases variable would not converge to a solution as a random factor. Strategies for attempting to solve convergence problems include centring the variable, or changing the algorithm from iterative generalised least squares (IGLS) to restricted iterative generalised least squares (RIGLS).[114]

Conclusions

It was demonstrated that the addition of the non-randomised data changed the interpretation of the learning curve effect in one arm of a trial of groin hernia repair. This finding should be interpreted cautiously, however, because of concerns about the completeness of the non-randomised data and the generalisability of the finding.

A promising approach to assessing learning effects in one arm of an RCT is to use the non-randomised data to gauge the position on the learning curve for each of the randomised cases (trial cases coding), without using the actual data from the non-randomised cases. Empirical testing of this strategy is now required within a variety of health technologies.

Chapter 7

Conclusions and implications

Overall conclusions

It is clear that both individual operators and institutions learn through experience. This change in effectiveness over time could invalidate health technology assessments. The overall aim of this project was to determine whether there were existing methods that had been, or could be, adapted to the purpose of allowing for 'the learning curve' when assessing a new health technology.

The systematic search of the health technology assessment literature revealed that, while many researchers were aware of the problem, there were no satisfactory methods for addressing it. The search of non-health technology assessment literature in areas such as psychology and human factors revealed some additional techniques but, again, none that provided an obvious solution to the problem in hand. Finally, some of the curves and methods identified in the reviews of health technology assessment and non-health technology assessment literature were fitted to some large data sets. These revealed further weaknesses in current data collection systems and statistical methods.

Given the effort put into the search for data and references, our main conclusion is drawn with confidence – new methods will have to be developed and this will require considerable further research. This research will only be possible if health technology assessments record and report much richer data. In particular, information is required about individual operator experience and outcome measures additional to operation time.

Specific conclusions

Learning curves are an important part of health technology assessment and the following conclusions are well-recognised.

1. Changes over time in the performance of a technology because of learning effects complicate evaluation and are an impediment to rigorous assessment.
2. Useful parameters for describing learning and hence exploring or adjusting for its effects are the rate and length of learning and the final skill level.
3. The effects of learning can be complex and an important distinction can be drawn between individual (operator) learning and institutional learning.

Searching for statistical techniques

A solution to the type of statistical analysis required to address learning curve effects in health technology assessment might have existed in the clinical literature but was not widely disseminated. Therefore, a systematic review of studies in the health technology assessment literature was undertaken. The purpose of the systematic review was not to track down every single paper related to learning curves but to describe the 'epidemiology' of studies that have addressed learning curve effects in health technology assessment. It was important to avoid missing a useful paper that described a new statistical method. The advantage of a systematic approach is that the results are transparent and potentially reproducible by other researchers. The review was extremely resource intensive – some 4571 abstracts had to be read and 559 full papers required assessment. Data were abstracted from 272 included papers. In addition, questionnaires were sent out to experts in health technology assessment. The conclusions from the review were as follows.

4. As in other systematic reviews of methodological issues in health technology assessment, designing search strategies for studying learning effects was difficult because there is little relevant indexing or key-wording. Even limiting electronic searches to 'learning curve' generated thousands of abstracts, most of which were not relevant. Contacting experts in the field proved particularly useful in a search for novel methods.
5. Often there was too little information contained in the papers to make a decision on study quality.
6. Systematic searching of published clinical literature showed that the statistical methods used in health technology assessment have

generally been crude, and many reports rely on simple description without any formal statistical analysis.

Learning curve effects exist in other fields, such as psychology, aviation and manufacture, and it was possible that more advanced statistical techniques had been used these fields than in health technology assessment. To keep the task of finding these techniques as unbiased as possible, a systematic search of a number of fields was undertaken. Once again, the resources required were considerable. Nearly 10,000 abstracts were reviewed, 100 full papers assessed and 18 papers were included in the review. Experts in the various fields were also sent questionnaires.

Based upon the results of the searches, the conclusions were as follows.

7. The systematic searching in these fields identified an additional eight possible shapes of curve and nine statistical methods for assessing learning effects which had not hitherto been used in health technology assessment. This demonstrated the value of considering fields outside clinical research when addressing methodological issues in health technology assessment.
8. Statistical methods for exploring learning can be categorised as either for **identifying** learning (Was there a change over time?), or for **measuring** learning (What size was the change?).
9. There is a hierarchy of methods for identifying and measuring learning, and the more sophisticated methods for both have had little, if any, use in health technology assessment.

Single operator case series

The systematic reviews identified a hierarchy of techniques that could be applied to simple and complex data. Case series design was used in 95% of the studies identified in the review of the health technology assessment literature and, hence, it is important to understand how to identify and measure changes in performance in case series designs. Some of the statistical methods identified in the reviews were illustrated and explored by applying them to a case series of 190 consecutive laparoscopic fundoplication procedures performed by a single surgeon, from which the following conclusions were drawn.

10. Splitting the series into thirds or quarters can be misleading.
11. The various curves that were identified can be compared statistically using the AIC (although there may well be alternatives).
12. Finding measures that are reliable proxies for learning can be difficult. Measures of patient outcome tend to be dichotomous rare events (such as complications) and therefore relatively intractable to statistical analysis; no method was identified that was entirely suitable for these. Methods are available for analysing continuous process measures, such as time to complete an operation, but the relationship of these variables to learning may be weak.
13. Adjustment for case-mix may be required in analyses of learning. Variation in case-mix may obscure learning effects by introducing background noise into the analysis and case-mix changes over time can confound any learning effect.

Multiple case series designs

Ascertaining the correct learning curve shape of an individual operator is important but does not give an indication of how that operator differs from another operator. To begin to compare operators it is necessary to use data from a number of operators. Some of the techniques identified in the reviews were illustrated and explored using multiple case series of laparoscopic cholecystectomy procedures performed by ten surgeons.

It is apparent that the structure of multiple case series data in the simplest scenario is hierarchical in nature. That is, the operators are a higher level of the hierarchy and the many procedures are at a lower level. The institution adds an extra level to the hierarchy in multiple institution studies. It was decided to illustrate only one of the complex methods (multilevel models), because it appeared to fulfil the necessary characteristics of modelling the hierarchy and enabling flexibility in the type of shape of learning. The multilevel models statistical package MLWin[93] was used for the model fitting.

The following conclusion was drawn from the analysis of the cholecystectomy data.

14. Assessing learning in multiple operators is complex and requires a multilevel approach. These derive measures of individual operator performance and of how each operator compares with other operators. Of the possible methods identified for doing this, only multilevel modelling was tested within this project.

Incomplete data from an RCT

Investigators sometimes believe that their randomised trial may be compromised or biased by learning effects in one arm of the trial. To begin to address this issue, investigators need to identify and measure any learning effect in that arm of the trial. Such assessment is commonly complicated, however, by incomplete case series on the operators in the trial (not every case is necessarily included in a randomised trial). The analysis of complex structured data was extended to randomised trial data derived from the laparoscopic procedure arm of a multicentre randomised trial in groin hernia repair. The randomised trial data were supplemented by information on non-randomised operations performed over the period of the trial.

The data were analysed using multilevel models and the following conclusions were drawn.

15. The coding of the experience variable plays an important role in ascertaining the rate of learning.
16. The addition of the non-randomised data changed the learning curve parameters, suggesting that coding experience using only randomised trial data may spuriously overestimate the rate of learning.
17. While complete data on non-randomised cases is optimal, collection of such data may be too resource intensive to be feasible in a pragmatic randomised trial. A compromise could be to limit collection of non-randomised data to determine the position of the randomised cases in the operator's full series.

Implications for health technology assessment

1. Reliable assessment of learning effects is most likely to come from prospectively collected data on multiple operators or institutions.
2. Reports of studies of learning should, as a minimum, describe the number and experience of the operators, the data source, the proportion of procedures performed by individual operators (to ensure that no one operator dominates) and the level of care.
3. Any study of a methodological topic related to health technology assessment should consider other fields.
4. Investigators of a methodological topic should consider contacting experts in other fields.
5. Researchers should give due attention to the intensive resource implications of undertaking systematic reviews of methodological issues inside and outside health technology assessment.
6. As a principle, the simplest methods within the hierarchies of statistical methods described in this report should be used in a parsimonious way.
7. The split group method should only be used in the context of a linear trend component, interpreted as statistical rather than graphical evidence of performance change.
8. Given that the power curve has been used extensively across many fields as the shape of learning, any comparison of different curve shapes should be made against the power curve.
9. Autocorrelation may be present in learning curve data and should be investigated using an appropriate statistical method.
10. Investigators should consider, and adjust for, any confounding factors.
11. When there are multiple operators in a study, a method should be used which takes into account the hierarchical nature of the data.
12. Collection of non-randomised data alongside an RCT may, despite possible limitations, aid the assessment of learning curve effects.
13. To analyse trends in operator performance within RCTs requires the positions of the operators on their learning curves to be known.

Recommendations for further research

The implications of this work for future research are explored below, and some of the issues surrounding learning curve effects that require to be addressed are described.

1. Further empirical testing of the techniques identified is required. In particular, the generalisability of the various shapes and methods that were identified needs to be assessed for a wider variety of health technologies.

The shape of learning curves associated with three laparoscopic technologies has been explored empirically in this report. However, minimal access techniques are only one type of technology that exhibits learning effects. The testing of the various shapes should be extended to other health technologies that require skill and learning. These include other surgical techniques, diagnostic methods and where dosage decision-making is necessary.

There are several ways to take the statistical analysis of the shapes of learning forward.

- ***Shifting the axis of the curves*** In the analyses described here it was assumed that the experience axis started at zero. Many of the curves can be transformed by shifting the starting point of experience by a constant, c. For example, the power law denoted by $y = aX^{-b}$ (see chapter 4) can be transformed to $y = a(X + c)$. The possible advantage of this is to get a better fit to the initial points on the curve.
- ***Adding covariates to the curve shapes*** Many results of procedures are confounded by patient characteristics such as age, sex and weight. The interaction of these characteristics with the various curve shapes should be investigated.
- ***Curve shapes within a technology*** As well as comparing the shapes of learning between technologies, a comparison of the shape of individual operators' or institutions' learning of a technology should be investigated.
- ***Prior experience of operators*** Questionnaires could be sent to operators in a particular study asking for prior experience/training in the technology under investigation. The results of the questionnaire could then be used to investigate if prior experience/training influences the learning curve.
- ***Asymptote of learning*** Methods should be explored for estimation of the asymptote for each of the curve shapes (see recommendations on asymptotes below).

A number of methods for complex structured data were identified and require to be tested on data sets with different numbers of operators and procedures.

- ***Applicability of data sets*** Some of the methods may be more appropriate for data that have many operators but relatively few observations per operator (for example, generalised estimating equations). The current data sets could be used in simulation studies to generate data with certain characteristics that could then be used to test the limitations of the various techniques. Preferably more data sets should be obtained.
- ***Dichotomous outcomes*** Many of the complex methods identified cannot be used on dichotomous data. For example, stochastic parameter and latent curve models require the data to be normally distributed. However, multilevel models can be applied to dichotomous data and should be explored further using a number of different data sets.
- ***Bayesian techniques*** Bayesian hierarchical models are commonly used for multilevel modelling and their usefulness for modelling learning curves should be explored.

2. Estimation of the time taken to reach an asymptote should be explored further.

The asymptote of learning is the final level of performance obtained by an operator. An estimate of the asymptote is important for two reasons. First, the number of procedures required to reach the asymptote can be calculated and this gives an indication of the potential costs of learning the new procedure. Second, knowledge of the asymptote enables investigators to discard the cases in the learning phase and, hence, to evaluate the optimal results for a procedure.

There are a number of difficulties with estimating the asymptote of performance, and further research is required on the following issues.

- Large variability in the process and outcome variables can mean that estimates for the asymptote can vary considerably.
- Often a series of datapoints is too short to make a reliable estimate of the asymptote. For example, an RCT may only have several observations per operator.
- The full series of cases for an operator is often not collected; it can be difficult, therefore, to judge if the asymptote has been reached.
- There are no guidelines for what is a clinically important distance from the asymptote, although this is likely to vary from technology to technology.

3. Further research is required on variables that are good proxies for learning.

The variables used most commonly to assess learning are the operation time and the number of complications. Operation time is certainly a dimension of learning but relates to the process of the technique and is not a measure of competence. In contrast, complications are a measure of outcome of a procedure but, often, the complications are too rare for statistical analysis. The complications (and operation times) arise through a variety of factors, only some of which are under the control of the operator. These factors include the severity of disease, patient compliance, financial and institutional constraints, limitations of the technology and personal characteristics of the operator and patient.[122]

Alternative proxies for learning should be investigated. These possibly include surgical near-misses,[67] length of stay, blood use, economy of motion,[66] and disease-specific quality-of-life measures. The statistical techniques identified in chapters 2–6 could be used to explore learning with all of these measures.

4. **Relatively rare, dichotomous outcomes are often the most important measures of performance but are currently the least tractable to analysis. Further methodological research is needed to address this issue.**

Measuring the learning curve statistically using rare events is difficult. The systematic searches described in this report did not identify any promising statistical techniques. As a first step, better reporting of complications is needed, ideally through a central data collection system. This would maximise the number of events.[123] Even if this was set up, there would still be difficulties in determining **individual** learning curves if there are few events per individual.

5. **Further empirical work is required to identify the optimal method for assessing learning curves within RCTs. In principle, a randomised trial design should protect against case-mix drift over time.**

The RCT is rightly recognised as the gold standard design for evaluating the effectiveness and cost-effectiveness of new technologies. Learning effects in the new technology can complicate the evaluation. The analysis of the laparoscopic hernia repair trial (chapter 6) illustrated that the learning curve in the laparoscopic arm of the randomised trial was distorted if no consideration was given to the operations performed between randomised cases. The analysis did not consider the impact of learning curve effects upon the size of differential effects between trial groups. In principle, changes in case-mix will be equalised in the treatment and control arms of the randomised trial, but it is still conceivable that such changes will influence the new treatment arm only (or more so than the standard approach). For example, selection of less obese patients might aid laparoscopic surgery but make relatively little difference to the performance of the standard technique. Therefore, further analyses required to investigate these influences include:

- more extensive testing in other technologies of the influence of observational data in the analysis of learning effects in one arm of a randomised trial
- mathematical simulation techniques employed to investigate the impact of 'missing' information on the estimates of learning effects
- analysis of trial data using multilevel modelling that adjusts for learning effects
- investigation of case-mix influences in both arms of a trial using the techniques identified in this report
- assessment of the impact of differing skill levels on the effect sizes using simulation techniques
- assessment of the effects of varying the number of procedures performed by individual operators.

6. **The impact of learning curve factors on economic evaluation should be explored.**

Increasingly, pragmatic trials have an economic evaluation relating costs to effectiveness. The learning curve will impact not only on the effectiveness but also on costs.[37,124] For example, in laparoscopic surgery the staff costs will be affected by changes in operation time (calculated as (staff costs + overheads) × operation time) and the seniority of the surgeons. However, the net impact of the learning curve on cost is uncertain: more junior staff are less experienced and less costly than senior staff; hence, the impact on cost may not be significant. Hospitalisation costs may also vary with experience if the learning curve impacts upon the length of stay of a patient.

There are certainly at least two complementary analyses that could be performed in this context. The first would be an exploratory sensitivity analysis that measured the average cost of a procedure per hospital as the operation time is varied. The second analysis would use the costs per patient data to look for factors that contributed to the costs, and so the experience variable would be used as a predictor in this analysis. Learning could, therefore, have an important impact on cost-effectiveness and, hence, further work on the influence of learning and skill on the benefits and costs of introducing new technologies is required.

7. **Data sets for addressing these issues are most likely to come from prospective data collection; this should be built into the design of future evaluations of technologies likely to show a learning effect.**

While learning curve effects can potentially confound rigorous evaluations of health technologies, such as in the context of RCTs,

learning curve effects can be explored using data from less rigorous observational studies. It is a challenge to healthcare assessors to come up with scientifically adequate, yet economically viable, research designs. Lilford and colleagues[125] have recently proposed a design, called a **tracker study**, that may provide a solution.

A tracker study consists of a number of randomised comparisons of a new type of technology to standard treatment. These technologies will not necessarily be stabilised but, rather, the tracker study would be able to monitor progress of the technology from specific to general use. In these studies, each clinician would be allowed to randomise between trial arms that they considered were reasonable alternatives. In addition, data would also be collected on an observational basis, because many operators or institutions may use only one of the treatment options. A tracker study would be flexible to changes in procedure and would include all operators and institutions, irrespective of skill or experience. In this way, the equipoise[126] of each individual can be utilised to a maximum, since the time lag between an operator wanting to undergo a rigorous evaluation and the setting up of a randomised trial would be minimised. Also, the data collection would be maximised and, therefore, could aid the development of a technology. Lilford and colleagues[125] expected that the tracker study would be used as an early warning system for treatments that perform poorly and could be used to reject unpromising new treatments.

Of course, the method of analysis of these studies would be complex in relation to the learning curve. In theory, the methods proposed in this report would at least provide a sufficient starting place. The funding arrangements required for a tracker trial would need to be more sophisticated than for a standard evaluation. The commissioning body would require flexible budgets to handle the unknown duration of the study and the research group would have to be flexible to respond to changes in the technology. A compromise, in terms of the learning curve issue, would be for funding bodies to commission observational data collection alongside a clinical trial.

8. A theory-based approach (instead of statistical) should be investigated.

This report has focussed on methods for analysing learning curves statistically; ways to develop these further are suggested above. There is also, however, a large amount of literature on the theoretical aspects of learning, such as skill acquisition.[127] The research agenda could be broadened to consider theoretical aspects of learning in the context of health technology assessment.

9. Parallels and possible overlap between the identification and description of learning and the statistical aspects of quality assurance (identifying a system going out of control) should be explored further.

Quality control is a major component of healthcare, and statistical methods for safety monitoring that take into account operator effects and identify those operators who are performing below standard are required. This was brought to public attention by the deaths of babies undergoing congenital heart surgery in Bristol.[128] Indeed, the surgeon at the centre of the inquiry attributed part of the poor performance to the "learning curve".[129] There is extensive literature on statistical quality control.[130] The main aim of quality control is to identify when a system goes out of control. In contrast, the learning curve effect typically involves the beginning of the series and is concerned with a system coming under control. Nevertheless, the similarity should be investigated further.

Acknowledgements

This project was commissioned by the NHS R&D Health Technology Assessment Programme.

The authors wish to thank Jane Andrew for her help with the data abstraction and Jonathan Cook for his help with the data analyses in chapter 4. We also wish to thank Zygmunt Krukowski for supplying us with the data sets in chapters 4 and 5 and for his comments on specific analyses.

Our thanks are also due to the experts in the various fields for taking time to reply to our questionnaire.

Finally, the authors are indebted to the referees for their perseverance in reading the report and the quality of their comments.

The Health Services Research Unit is funded by the Chief Scientist Office of the Scottish Executive Health Department. The views expressed are those of the authors.

References

1. Banta D, Luce PA. Health care technology and its assessment: an international perspective. Oxford: Oxford University Press; 1993.
2. Spodick DH. The surgical mystique and the double standard. Controlled trials of medical and surgical therapy for cardiac disease: analysis, hypothesis, proposal. *Am Heart J* 1973;**85**:579–83.
3. Spodick DH. Numerators without denominators. There is no FDA for the surgeon. *JAMA* 1975;**232**:355–6.
4. van der Linden W. Pitfalls in randomized surgical trials. *Surgery* 1980;**87**:258–62.
5. Love JW. Drugs and operations: some important differences. *JAMA* 1975;**232**:37–8.
6. Stirrat GM, Farrow SC, Farndon J, Dwyer N. The challenge of evaluating surgical procedures. *Ann R Coll Surg Engl* 1992;**74**:80–4.
7. Bouchard S, Barkun AN, Barkun JS, Joseph L. Technology assessment in laparoscopic general surgery and gastrointestinal endoscopy: science or convenience? *Gastroenterology* 1996;**110**:915–25.
8. Mowatt G, Bower DJ, Brebner JA, Cairns JA, Grant AM, McKee L. When is the 'right' time to initiate an assessment of a health technology? *Int J Technol Assess Health Care* 1998;**14**:372–86.
9. Neugebauer E, Troidl H, Spangenberger W, Dietrich A, Lefering R. Conventional versus laparoscopic cholecystectomy and the randomized controlled trial. *Br J Surg* 1991;**78**:150–4.
10. Buxton MJ. Problems in the economic appraisal of new health technology: the evaluation of heart transplants in the UK. In Drummond MF, editor. Economic appraisal of health technology in the European Community. Oxford: Oxford Medical Publications; 1987. p.103–18.
11. Health technology assessment in surgery: the role of the randomised controlled trial. London: Medical Research Council; 1994.
12. Russell I. Evaluating new surgical procedures. *BMJ* 1995;**311**:1243–4.
13. Border P. Minimal access ('keyhole') surgery and its implications. London: Parliamentary Office of Science and Technology, 1995.
14. Chalmers TC. Randomization of the first patient. *Med Clin North Am* 1975;**59**:1035–8.
15. Bull C, Yates R, Sarkar D, de Leval MR. Scientific, ethical, and logistical considerations in introducing a new operation: a retrospective cohort study from paediatric cardiac surgery. *BMJ* 2000;**320**:1168–73.
16. Gross M. Innovations in surgery. A proposal for phased clinical trials. *J Bone Joint Surg Br* 1993;**75**:351–4.
17. Buchwald H. Surgical procedures and devices should be evaluated in the same way as medical therapy. *Control Clin Trials* 1997;**18**:478–87.
18. Cuschieri A. Whither minimal access surgery: tribulations and expectations? *Am J Surg* 1995;**169**:9–19.
19. Altman DG, Royston JP. The hidden effect of time. *Stat Med* 1988;**7**:629–37.
20. Tai CT, Chen SA, Chiang CE, Chiou CW, Kuo BI, Wu TJ, *et al.* The effects of accumulated experience on radiofrequency ablation of accessory pathways. *Jpn Heart J* 1995;**36**:729–39.
21. Harkki-Siren P, Sjoberg J. Evaluation and the learning curve of the first one hundred laparoscopic hysterectomies. *Acta Obstet Gynecol Scand* 1995;**74**:638–41.
22. Fisher KS, Matteson KM, Hammer MD. Laparoscopic cholecystectomy: the Springfield experience. *Surg Laparosc Endosc* 1993;**3**:199–203.
23. Peters JH, Ellison EC, Innes JT, Liss JL, Nichols KE, Lomano JM, *et al.* Safety and efficacy of laparoscopic cholecystectomy. A prospective analysis of 100 initial patients. *Ann Surg* 1991;**213**:3–12.
24. Hughes GB. The learning curve in stapes surgery. *Laryngoscope* 1991;**101**:1280–4.
25. Archie JP Jr. Learning curve for carotid endarterectomy. *South Med J* 1988;**81**:707–10.
26. Kelsey SF, Mullin SM, Detre KM, Mitchell H, Cowley MJ, Gruentzig AR, *et al.* Effect of investigator experience on percutaneous transluminal coronary angioplasty. *Am J Cardiol* 1984;**53**:56–64C.
27. Higashihara E, Baba S, Nakagawa K, Murai M, Go H, Takeda M, *et al.* Learning curve and conversion to open surgery in cases of laparoscopic adrenalectomy and nephrectomy. *J Urol* 1998;**159**:650–3.
28. Kockerling F, Schneider C, Reymond MA, Scheidbach H, Konradt J, Barlehner E, *et al.* Early results of a prospective multicenter study on 500 consecutive cases of laparoscopic colorectal surgery. *Surg Endosc* 1998;**12**:37–41.

29. Witt PD, Wahlen JC, Marsh JL, Grames LM, Pilgram TK. The effect of surgeon experience on velopharyngeal functional outcome following palatoplasty. Is there a learning curve? *Plast Reconstr Surg* 1998;**102**:1375–84.

30. Starkes JL, Payk I, Hodges NJ. Developing a standardized test for the assessment of suturing skill in novice microsurgeons. *Microsurgery* 1998;**18**:19–22.

31. Horvath KD, Gray D, Benton L, Hill J, Swanstrom LL. Operative outcomes of minimally invasive saphenous vein harvest. *Am J Surg* 1998;**175**:391–5.

32. Meehan JJ, Georgeson KE. The learning curve associated with laparoscopic antireflux surgery in infants and children. *J Pediatr Surg* 1997;**32**:426–9.

33. Smith JE, Jackson AP, Hurdley J, Clifton PJ. Learning curves for fibreoptic nasotracheal intubation when using the endoscopic video camera. *Anaesthesia* 1997;**52**:101–6.

34. See WA, Cooper CS, Fisher RJ. Predictors of laparoscopic complications after formal training in laparoscopic surgery. *JAMA* 1993;**270**:2689–92.

35. Rege RV, Joehl RJ. A learning curve for laparoscopic splenectomy at an academic institution. *J Surg Res* 1999;**81**:27–32.

36. Danford DA, Kugler JD, Deal B, Case C, Friedman RA, Saul JP, *et al.* The learning curve for radiofrequency ablation of tachyarrhythmias in pediatric patients. Participating members of the Pediatric Electrophysiology Society. *Am J Cardiol* 1995;**75**:587–90.

37. Smith DB, Larsson JL. The impact of learning on cost: the case of heart transplantation. *Hosp Health Serv Admin* 1989;**34**:85–97.

38. Vossen C, Van Ballaer P, Shaw RW, Koninckx PR. Effect of training on endoscopic intracorporeal knot tying. *Hum Reprod* 1997;**12**:2658–63.

39. Davis Z, Jacobs HK, Zhang M, Thomas C, Castellanos Y. Endoscopic vein harvest for coronary artery bypass grafting: technique and outcomes. *J Thorac Cardiovasc Surg* 1998;**116**:228–35.

40. Atherton DP, O'Sullivan E, Lowe D, Charters P. A ventilation-exchange bougie for fibre optic intubations with the laryngeal mask airway. *Anaesthesia* 1996;**51**:1123–6.

41. Gilchrist BF, Vlessis AA, Kay GA, Swartz K, Dennis D. Open versus laparoscopic cholecystectomy: an initial analysis. *J Laparoendosc Surg* 1991;**1**:193–6.

42. Johnson C, Roberts JT. Clinical competence in the performance of fiber optic laryngoscopy and endotracheal intubation: a study of resident instruction. *J Clin Anesthesia* 1989;**1**:344–9.

43. Yuen PM, Rogers MS. Laparoscopic management of ovarian masses: the initial experience and learning curve. *Aust N Z J Obstet Gynaecol* 1994;**34**:191–4.

44. Konrad C, Schupfer G, Wietlisbach M, Gerber H. Learning manual skills in anesthesiology: is there a recommended number of cases for anesthetic procedures? *Anesth Analg* 1998;**86**:635–9.

45. Ou CS, Beadle E, Presthus J, Smith M. A multicenter review of 839 laparoscopic-assisted vaginal hysterectomies. *J Am Assoc Gynecol Laparosc* 1994;**1**:417–22.

46. Laffel GL, Barnett AI, Finkelstein S, Kaye MP. The relation between experience and outcome in heart transplantation. *N Engl J Med* 1992;**327**:1220–5.

47. Ng DT, Rowe NA, Francis IC, Kappagoda MB, Haylen MJ, Schumacher RS, *et al.* Intraoperative complications of 1000 phacoemulsification procedures: a prospective study. *J Cataract Refract Surg* 1998;**24**:1390–5.

48. Lawrence K, McWhinnie D, Goodwin A, Doll H, Gordon A, Gray A, *et al.* Randomised controlled trial of laparoscopic versus open repair of inguinal hernia: early results. *BMJ* 1995;**311**:981–5.

49. Behrens E, Schramm J, Zentner J, Konig R. Surgical and neurological complications in a series of 708 epilepsy surgery procedures. *Neurosurgery* 1997;**41**:1–9.

50. Blumenthal PD, Gaffikin L, Affandi B, Bongiovanni A, McGrath J, Glew G. Training for Norplant implant removal: assessment of learning curves and competency. *Obstetr Gynecol* 1997;**89**:174–8.

51. Buchman CA, Chen DA, Flannagan P, Wilberger JE, Maroon JC. The learning curve for acoustic tumor surgery. *Laryngoscope* 1996;**106**:1406–11.

52. Heinz G, Kratochwill C, Schmid S, Kreiner G, Siostrzonek P, Pacher R, *et al.* Sinus node dysfunction after orthotopic heart transplantation: the Vienna experience 1987–1993. *Pacing Clin Electrophysiol* 1994;**17**:2057–63.

53. Bennett CL, Stryker SJ, Ferreira MR, Adams J, Beart RW Jr. The learning curve for laparoscopic colorectal surgery. Preliminary results from a prospective analysis of 1194 laparoscopic-assisted colectomies. *Arch Surg* 1997;**132**:41–4.

54. Ghosh PK, Choudhary A, Agarwal SK, Husain T. Role of an operative score in mitral reconstruction in dominantly stenotic lesions. *Eur J Cardiothorac Surg* 1997;**11**:274–9.

55. Agachan F, Joo JS, Weiss EG, Wexner SD. Intraoperative laparoscopic complications. Are we getting better? *Dis Colon Rectum* 1996;**39**:S14–19.

56. Dunphy BC, Shepherd S, Cooke ID. Impact of the learning curve on term delivery rates following laparoscopic salpingostomy for infertility associated with distal tubal occlusive disease. *Hum Reprod* 1997;**12**:1181–3.

57. Bubolz B, Case CL, McKay CA, O'Connor BK, Knick BJ, Gillette PC. Learning curve for radio-frequency catheter ablation in pediatrics at a single institution. *Am Heart J* 1996;**131**:956–60.

58. Turjman F, Massoud TF, Sayre J, Vinuela F. Predictors of aneurysmal occlusion in the period immediately after endovascular treatment with detachable coils: a multivariate analysis. *Am J Neuroradiol* 1998;**19**:1645–51.

59. Wijnberger, van der SC. Learning in medicine: chorionic villus sampling. *Fetal Diagn Ther* 1998;**13**:83.

60. Woods JR, Saywell RMJ, Nyhuis AW, Jay SJ , Lohrman RG, Halbrook HG. The learning curve and the cost of heart transplantation. *Health Serv Res* 1992;**27**:219–38.

61. Jowell PS, Baillie J, Branch MS, Affronti J, Browning CL, Bute BP. Quantitative assessment of procedural competence. A prospective study of training in endoscopic retrograde cholangio-pancreatography. *Ann Intern Med* 1996;**125**:983–9.

62. Kopacz DJ, Neal JM, Pollock JE. The regional anesthesia "learning curve". What is the minimum number of epidural and spinal blocks to reach consistency? *Reg Anesth* 1996;**21**:182–90.

63. Parry BR, Williams SM. Competency and the colonoscopist: a learning curve. *Aust N Z J Surg* 1991;**61**:419–22.

64. Schlup MT, Williams SM, Barbezat GO. ERCP: a review of technical competency and workload in a small unit. *Gastrointest Endosc* 1997;**46**:48–52.

65. Molloy M, Archer SB, Hasselgren PO, Dalton BJ, Bower RH. Cholangiography during laparoscopic cholecystectomy: cumulative sum analysis of an institutional learning curve. *Gastroenterology* 1998;**114**:S0163.

66. Darzi A, Smith S, Taffinder N. Assessing operative skill. *BMJ* 1999;**318**:887–8.

67. de Leval MR, Francois K, Bull C, Brawn W, Spiegelhalter D. Analysis of a cluster of surgical failures: application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg* 1994;**107**:914–24.

68. Steiner SH, Cook RJ, Farewell VT. Monitoring paired binary surgical outcomes using cumulative sum charts. *Stat Med* 1999;**18**:69–86.

69. Goddard CC, Gilbert RJ, Needham G, Deans HE. Routine receiver operating characteristic analysis in mammography as a measure of radiologists' performance. *Br J Radiol* 1998;**71**:1012–17.

70. Gutzwiller F, Chrzanowski R, Paccaud F. Data bases for the assessment of medical technologies: examples from Europe. *Int J Technol Assess Health Care* 1988;**4**:65–73.

71. Wright TP. Factors affecting the cost of airplanes. *J Aeronaut Sci* 1936;**3**:122–36.

72. Fitts PM, Posner MI. Human performance. Belmont (CA): Brooks Cole; 1969.

73. Hammond NV. Principles from the psychology of skill acquisition. In Gardiner MM, Christie B, editors. Applying cognitive psychology to user-interface design. Chichester: Wiley; 1987. p.163–87.

74. Newell A, Rosenbloom P. Mechanisms of skill acquisition and the law of practice. In Anderson JR, editor. Cognitive skills and their acquisition. Hillsdale, NY: Lawrence Erlbaum Associates; 1981.

75. De Jong JR. The effect of increasing skill on cycle time and its consequences for time standards. *Ergonomics* 1957;**1**:51–60.

76. Eyring JD, Johnson DS, Francis DJ. A cross-level units-of-analysis approach to individual differences in skill acquisition. *J Appl Psychol* 1993;**78**:805–14.

77. Logan GD. Shapes of reaction-time distributions and shapes of learning curves: a test of the instance theory of automaticity. *J Exp Psychol Learn Mem Cogn* 1992;**18**:883–914.

78. Buck J, Cheng S. Instructions and feedback effects on speed and accuracy with different learning-curve models. *IIE Trans* 1993;**25**:34–47.

79. Bailey C, Mcintyre E. Some evidence on the nature of relearning curves. *Account Rev* 1992;**67**:368–78.

80. Spears WD. Measurement of learning and transfer through curve fitting. *Hum Factors* 1985;**27**:251–66.

81. Zaraiskaya YI, Aleksandrova EA, Lukashev AO, Shvyrkova NA. Features of active avoidance learning in rats with streptozotocin diabetes. *Neurosci Behav Physiol* 1994;**24**:167–9.

82. Ziegler A, Kastner C, Blettner M. The generalised estimating equations: an annotated bibliography. *Biometric J* 1998;**40**:115–39.

83. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 1998;**17**:1261–91.

84. Goldstein H, Healy MJR, Rasbash J. Multilevel time series models with applications to repeated measures data. *Stat Med* 1994;**13**:1643–55.

85. Rice N, Leyland A. Multilevel models: applications to health data. *J Health Serv Res Policy* 1996;**1**:154–64.

86. Ullman JB. Structural equation modelling. In: Tabachnick BG, Fidell LS, editors. Using multivariate statistics. New York: HarperCollins College Publishers; 1996. p. 709–811.

87. Browne MW, Du-Toit S-HC. Models for learning data. In: Collins LM, Horn JL, editors. Best methods for the analysis of change: recent advances, unanswered questions, future directions. Washington DC: American Psychological Association; 1991. p. 47–68.

88. Baloff N, Becker SW. On the futility of aggregating individual learning curves. *Psychol Rep* 1967; **20**:183–91.

89. Hayes KJ. The backward curve: a method for the study of learning. *Psychol Rev* 1953;**60**:269–75.

90. Sidman M. A note on functional relations obtained from group data. *Psychol Bull* 1952;**49**:263–9.

91. Estes WK. The problem of inference from curves based on group data. *Psychol Bull* 1956;**53**:134–40.

92. Delaney PF, Reder LM, Staszewski JJ, Ritter FE. The strategy-specific nature of improvement: the power law applies by strategy within task. *Psychol Sci* 1998;**9**:1–7.

93. Goldstein H, Rasbash J, Plewis I, Draper D, Browne W, Yang M, *et al.* A user's guide to MLwiN. London: University of London; 1998.

94. Nissen R. Gastropexy and fundoplication in surgical treatment of hiatus hernia. *Am J Dig Dis* 1961;**6**:954–61.

95. Dallemagne B, Weerts JM, Jehaes C, Markiewicz S, Lombard R. Laparoscopic Nissen fundoplication: preliminary report. *Surg Laparosc Endosc* 1991;**1**:138–43.

96. Watson DI, Baigrie RJ, Jamieson GG. A learning curve for laparoscopic fundoplication. Definable, avoidable, or a waste of time? *Ann Surg* 1996;**224**:198–203.

97. Munro W, Brancatisano R, Adams IP, Falk GL. Complications of laparoscopic fundoplication: the first 100 patients. *Surg Laparosc Endosc* 1996;**6**:421–3.

98. Deschamps C, Allen MS, Trastek VF, Johnson JO, Pairolero PC. Early experience and learning curve associated with laparoscopic nissen fundoplication. *J Thorac Cardiovasc Surg* 1998;**115**:281–4.

99. Trastek VF, Deschamps C, Allen MS, Miller DL, Pairolero PC, Thompson AM. Uncut Collis–Nissen fundoplication: learning curve and long-term results. *Ann Thorac Surg* 1998;**66**:1739–43.

100. Diggle PJ. Time series: a biostatistical introduction. Oxford, UK: Oxford University Press; 1990.

101. Scheffe H. The analysis of variance. New York: Wiley, 1959.

102. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second international symposium on inference theory. Budapest: Akademiai Kiado; 1973. p. 267–81.

103. Lindsey J, Jones B. Generalized linear models applied to medical data. *Stat Med* 1998;**17**:59–68.

104. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978;**6**:461–4.

105. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Chicago: Rand McNally College Publishing; 1979.

106. Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* 1997;**350**:1128–30.

107. Sarli L, Pietra N, Sansebastiano G, Cattaneo G, Costi R, Grattarola M, *et al.* Reduced postoperative morbidity after elective laparoscopic cholecystectomy: stratified matched case–control study. *World J Surg* 1997;**21**:872–9.

108. Vanek VW, Rhodes R, Dallis DJ. Results of laparoscopic versus open cholecystectomy in a community hospital. *South Med J* 1995;**88**:555–66.

109. Kane RL, Lurie N, Borbas C, Morris N, Flood S, McLaughlin B, *et al.* The outcomes of elective laparoscopic and open cholecystectomies. *J Am Coll Surg* 1995;**180**:136–45.

110. Cagir B, Rangraj M, Maffuci L, Herz BL. The learning curve for laparoscopic cholecystectomy. *J Laparoendosc Surg* 1994;**4**:419–27.

111. Orlando R, Russell JC, Lynch J, Mattie A. Laparoscopic cholecystectomy: a statewide experience. The Connecticut Laparoscopic Cholecystectomy Registry. *Arch Surg* 1993;**128**:494–8.

112. Sariego J, Spitzer L, Matsumoto T. The "learning curve" in the performance of laparoscopic cholecystectomy. *Int Surg* 1993;**78**:1–3.

113. Moore MJ, Bennett CL. The learning curve for laparoscopic cholecystectomy. The Southern Surgeons Club. *Am J Surg* 1995;**170**:55–9.

114. Goldstein H. Multilevel statistical models. New York: Wiley; 1995.

115. The MRC Laparoscopic Groin Hernia Trial Group. Laparoscopic versus open repair of groin hernia: a randomised comparison. *Lancet* 1999;**354**:185–90.

116. Liem MS, van Steensel CJ, Boelhouwer RU, Weidema WF, Clevers GJ, Meijer WS, *et al.* The learning curve for totally extraperitoneal laparoscopic inguinal hernia repair. *Am J Surg* 1996;**171**:281–5.

117. Quilici PJ, Greaney EM, Quilici J, Anderson S. Transabdominal preperitoneal laparoscopic inguinal herniorrhaphy: results of 509 repairs. *Am Surg* 1996;**62**:849–52.

118. Champault G, Rizk N, Catheline JM, Riskalla H, Boutelier P. [Inguinal hernia. Pre-peritoneal laparoscopic surgery vs. the Stoppa procedure. A prospective randomized trial: 100 cases] [in French]. *J Chir (Paris)* 1996;**133**:274–80.

119. Tummala M, Rao BV. Study of lms and ser algorithms for fir filtering. *Comput Elect Eng* 1987;**13**:169–75.

120. Kald A, Anderberg B, Smedh K, Karlsson M. Transperitoneal or totally extraperitoneal approach in laparoscopic hernia repair: results of 491 consecutive herniorrhaphies. *Surg Laparosc Endosc* 1997;**7**:86–9.

121. Campbell M, Grimshaw J, Steen N. Sample size calculations for cluster randomised trials. *J Health Serv Res Policy* 2000;**5**:12–16.

122. Deusinger SS. Analyzing errors in practice: a vehicle for assessing and enhancing the quality of care. *Int J Technol Assess Health Care* 1992;**8**:62–75.

123. Bates DW, Gawande AA. Error in medicine: what have we learned? *Ann Intern Med* 2000;**132**:763–7.

124. Baltussen R, Ament A, Leidl R. Making cost assessments based on RCTs more useful to decision-makers. *Health Policy* 1996;**37**:163–83.

125. Lilford RJ, Braunholtz DA, Greenhalgh R, Edwards SJL. Trials and fast changing technologies: the case for tracker studies. *BMJ* 2000;**320**:43–6.

126. Lilford RJ, Jackson J. Equipoise and the ethics of randomization. *J R Soc Med* 1995;**88**:552–9.

127. Anderson JR. Acquisition of cognitive skill. *Psychol Rev* 1982;**89**:369–406.

128. Treasure T. Lessons from the Bristol case. *BMJ* 1998;**316**:1685–6.

129. Ramsay S. Bristol surgeon attributes poor performance to "learning curve". *Lancet* 1999;**354**:1980.

130. Montgomery DC. Introduction to statistical quality control. New York: Wiley; 1991.

Included papers identified in health technology assessment literature: phase 1 of project

Agachan F, Joo JS, Weiss EG, Wexner SD. Intraoperative laparoscopic complications. Are we getting better? *Dis Colon Rectum* 1996;**39**(10 suppl):S14–19.

Agachan F, Joo JS, Sher M, Weiss EG, Nogueras JJ, Wexner SD. Laparoscopic colorectal surgery. Do we get faster? *Surg Endosc* 1997;**11**:331–5.

Akinboboye O, Sumner J, Gopal A, King D, Shen Z, Bardfeld P, *et al.* Visual estimation of ejection fraction by two-dimensional echocardiography: the learning curve. *Clin Cardiol* 1995;**18**:726–9.

Alimi Y, Orsoni P, Hartung O, Berdah S, Lonjon T, Cador L, *et al.* Laparoscopic replacement of the abdominal aorta. Experimental study in the pig. *J Mal Vasc* 1998;**23**:191–4.

Archie JP Jr. Learning curve for carotid endarterectomy. *South Med J* 1988;**81**:707–10.

Asplund OA, Davies DM. Vertical scar breast reduction with medial flap or glandular transposition of the nipple-areola. *Br J Plast Surg* 1996;**49**:507–14.

Atherton DP, O'Sullivan E, Lowe D, Charters P. A ventilation-exchange bougie for fibre-optic intubations with the laryngeal mask airway. *Anaesthesia* 1996;**51**:1123–6.

Baldwin BJ, Schusterman MA, Miller MJ, Kroll SS, Wang BG. Bilateral breast reconstruction: conventional versus free TRAM. *Plast Reconstr Surg* 1994;**93**:1410–16.

Baumbach A, Bittl JA, Fleck E, Geschwind HJ, Sanborn TA, Tcheng JE, *et al.* Acute complications of excimer laser coronary angioplasty: a detailed analysis of multicenter results. *J Am Coll Cardiol* 1994;**23**:1305–13.

Beek FJ, Kaatee R, Beutler JJ, van der Ven PJ, Mali WP. Complications during renal artery stent placement for atherosclerotic ostial stenosis. *Cardiovasc Intervent Radiol* 1997;**20**:184–90.

Begos DG, Arsenault J, Ballantyne GH. Laparoscopic colon and rectal surgery at a VA hospital. Analysis of the first 50 cases. *Surg Endosc* 1996;**10**:1050–6.

Behrens E, Schramm J, Zentner J, Konig R. Surgical and neurological complications in a series of 708 epilepsy surgery procedures. *Neurosurgery* 1997;**41**:1–9.

Bennett CL, Stryker SJ, Ferreira MR, Adams J, Beart RW Jr. The learning curve for laparoscopic colorectal surgery. Preliminary results from a prospective analysis of 1194 laparoscopic-assisted colectomies. *Arch Surg* 1997;**132**:41–4.

Bickel A, Rappaport A, Hazani E, Eitan A. Laparoscopic cholecystectomy for acute cholecystitis performed by residents in surgery: a risk factor for conversion to open laparotomy? *J Laparoendosc Adv Surg Tech A* 1998;**8**:137–41.

Biro Z. Complications during the learning curve of phacoemulsification. *Ann Ophthalmol Glaucoma* 1998;**30**:370–4.

Bishoff, Jay T, Lloyd E, Robert A, Kenneth D, Louis R. Laparoscopic live donor nephrectomy: evaluation of the learning curve for a new operation. *J Endourol* 1998;**12**(suppl 1):S103.

Bittner R, Leibl B, Kraft K, Daubler P, Schwarz J. Laparoscopic hernioplasty (TAPP) – complications and recurrences in 900 operations. *Zentralbl Chir* 1996;**121**:313–19.

Bittner R, Leibl B, Kraft K, Butters M, Nick G, Ulrich M. Laparoscopic cholecystectomy in therapy of acute cholecystitis: immediate versus interval operation. *Chirurg* 1997;**68**:237–43.

Black JH, Hickey SA, Wormald PJ. An analysis of the results of myringoplasty in children. *Int J Pediatr Otorhinolaryngol* 1995;**31**:95–100.

Blanckaert J, Sallet G. Lasik learning curve: clinical study of 300 myopic eyes. *Bull Soc Belge d'Ophtalmol* 1998;**268**:7–12.

Blanton CL, Schallhorn SC, Tidwell JL. Radial keratotomy learning curve using the American technique. *J Cataract Refract Surg* 1998;**24**:471–6.

Block MS, Kent JN. Long-term follow-up on hydroxylapatite-coated cylindrical dental implants: a comparison between developmental and recent periods. *J Oral Maxillofac Surg* 1994;**52**:937–43.

Blumenthal PD, Gaffikin L, Affandi B, Bongiovanni A, McGrath J, Glew G. Training for Norplant implant removal: assessment of learning curves and competency. *Obstetr Gynecol* 1997;**89**:174–8.

Bogers AJJ, Frohn-Mulder IME, Witsenburg M, de Jong PL, de Joosten KFM, *et al.* Initial experience with the Norwood procedure for aortic atresia in the hypoplastic left heart. *Cardiovasc Eng* 1998;**3**(2):67–70.

Bongiorni MG, Soldati E, Arena G, Mariani M. Removal of chronic pacing and ICD leads: importance of the learning curve using transvenous techniques. *Eur Heart J* 1998;**19**(abstract suppl.):327.

Bracke FA, Meijer A, Van Gelder B. Learning curve characteristics of pacing lead extraction with a laser sheath. *Pacing Clin Electrophysiol* 1998;**21**:2309–13.

Brimacombe J. Analysis of 1500 laryngeal mask uses by one anaesthetist in adults undergoing routine anaesthesia. *Anaesthesia* 1996;**51**:76–80.

Browning DJ, Cobo LM. Early experience in extracapsular cataract surgery by residents. *Ophthalmology* 1985;**92**:1647–53.

Bruch H-P, Schiedeck THK, Schwandner O. Laparoscopic colorectal surgery: a five-year experience. *Dig Surg* 1999;**16**:45–54.

Bubolz B, Case CL, McKay CA, O'Connor BK, Knick BJ, Gillette PC. Learning curve for radiofrequency catheter ablation in pediatrics at a single institution. *Am Heart J* 1996;**131**:956–60.

Buchman CA, Chen DA, Flannagan P, Wilberger JE, Maroon JC. The learning curve for acoustic tumor surgery. *Laryngoscope* 1996;**106**:1406–11.

Burstein FD, Cohen SR, Huang MH, Sims CA. Applications of endoscopic surgery in pediatric patients. *Plast Reconstr Surg* 1998;**102**:1446–51.

Cagir B, Rangraj M, Maffuci L, Herz BL. The learning curve for laparoscopic cholecystectomy. *J Laparoendosc Surg* 1994;**4**:419–27.

Calafiore AM, Di Giammarco G, Teodori G, Gallina S, Maddestra N, Paloscia L, *et al.* Midterm results after minimally invasive coronary surgery (LAST operation). *J Thorac Cardiovasc Surg* 1998;**115**:763–71.

Callaghan JJ, Heekin RD, Savory CG, Dysart SH, Hopkinson WJ. Evaluation of the learning curve associated with uncemented primary porous-coated anatomic total hip arthroplasty. *Clin Orthop* 1992;**(282)**:132–44.

Cartwright PC, Snow BW, Mansfield JC, Hamilton BD. Percutaneous endoscopic trigonoplasty: a minimally invasive approach to correct vesicoureteral reflux. *J Urol* 1996;**156**:661–4.

Champault G, Rizk N, Catheline JM, Riskalla H, Boutelier P. Inguinal hernia. Pre-peritoneal laparoscopic surgery vs. the Stoppa procedure. A prospective randomized trial: 100 cases. *J Chir* 1996;**133**:274–80.

Chaney MA, Nikolov MP, Tuchek M, Bakhos M. An institution's initial experience with Port-Access minimally invasive cardiac surgery. *J Cardiothorac Vasc Anesth* 1998;**12**:617–19.

Chen MH, Murphy EA, Levison J, Cohen JR. Laparoscopic aortic replacement in the porcine model: a feasibility study in preparation for laparoscopically assisted abdominal aortic aneurysm repair in humans. *J Am Coll Surg* 1996;**183**:126–32.

Chung JY, Sackier JM. A method of objectively evaluating improvements in laparoscopic skills. *Surg Endosc Ultrasound Intervent Tech* 1998;**12**:1111–16.

Ciric I, Ragin A, Baumgartner C, Pierce D. Complications of transsphenoidal surgery: results of a national survey, review of the literature, and personal experience. *Neurosurgery* 1997;**40**:225–36.

Cobb TK, Knudson GA, Cooney WP. The use of topographical landmarks to improve the outcome of Agee endoscopic carpal tunnel release. *Arthroscopy* 1995;**11**:165–72.

Cohen TJ, Reid PR, Mower MM, Mirowski M, Aarons D, Juanteguy J, *et al.* The automatic implantable cardioverter-defibrillator. Long-term clinical experience and outcome at a hospital without an open-heart surgery program. *Arch Intern Med* 1992;**152**:65–9.

Colombo A, Maiello L, Almagor Y, Thomas J, Zerboni S, Di Summa M, *et al.* Coronary stenting: single institution experience with the initial 100 cases using the Palmaz-Schatz stent. *Cathet Cardiovasc Diagn* 1992;**26**:171–6.

Contini S, Franze A, DallValle R, Bertele A, Zinicola R. Laparoscopic fundoplication (LF): results in the first 31 patients and influence of the learning curve. *Br J Surg* 1998;**85**:114.

Copeland KL, Carpenter RJJ, Fenolio KR, Ledbetter DH. Integration of the transabdominal technique into an ongoing chorionic villus sampling program. *Am J Obstetr Gynecol* 1989;**161**:1289–94.

Dahl MT, Gulli B, Berg T. Complications of limb lengthening. A learning curve. *Clin Orthop* 1994; **(301)**:10–18.

Danchin N, Daclin V, Juilliere Y, Dibon O, Bischoff N, Pinelli G, *et al.* Changes in patient treatment after abrupt closure complicating percutaneous transluminal coronary angioplasty: a historic perspective. *Am Heart J* 1995;**130**:1158–63.

Danford DA, Kugler JD, Deal B, Case C, Friedman RA, Saul JP, *et al.* The learning curve for radiofrequency ablation of tachyarrhythmias in pediatric patients. *Am J Cardiol* 1995;**75**:587–90.

Daniels GFJ, Garnett JE, Carter MF. Ureteroscopic results and complications: experience with 130 cases. *J Urol* 1988;**139**:710–13.

Dashow L, Friedman I, Kempner R, Rudick J, McSherry C. Initial experience with laparoscopic cholecystectomy at the Beth Israel Medical Center. *Surg Gynecol Obstetr* 1992;**175**:25–30.

Davis Z, Jacobs HK, Zhang M, Thomas C, Castellanos Y. Endoscopic vein harvest for coronary artery bypass grafting: technique and outcomes. *J Thorac Cardiovasc Surg* 1998;**116**:228–35.

Delaney KA, Hessler R. Emergency flexible fiberoptic nasotracheal intubation: a report of 60 cases. *Ann Emerg Med* 1988;**17**:919–26.

Demmy TL, Curtis JJ, Boley TM, Walls JT, Nawarawong W, Schmaltz RA. Diagnostic and therapeutic thoracoscopy: lessons from the learning curve. *Am J Surg* 1993;**166**:696–700.

Deschamps C, Allen MS, Trastek VF, Johnson JO, Pairolero PC. Early experience and learning curve associated with laparoscopic nissen fundoplication. *J Thorac Cardiovasc Surg* 1998;**115**:281–4.

Dunphy BC, Shepherd S, Cooke ID. Impact of the learning curve on term delivery rates following laparoscopic salpingostomy for infertility associated with distal tubal occlusive disease. *Hum Reprod* 1997;**12**:1181–3.

Effert P, Boeckmann W, Wolff J, Jakse G. Laparoscopic lymphadenectomy in prostate carcinoma. Experiences with 120 patients. *Urologe A* 1996;**35**:413–17.

Fahlenkamp D, Loening SA, Turk J, Lindeke A, Muller W, Deger S. Complications of laparoscopic interventions in urology. *Urologe A* 1996;**35**:238–45.

Fahy GJ, Kleman JM, Wilkoff BL, Morant VA, Pinski SL. Low incidence of lead related complications associated with nonthoracotomy implantable cardioverter defibrillator systems. *Pacing Clin Electrophysiol* 1995;**18**:172–8.

Feaster SJ. After the learning curve: improved catheter performance in the NICU with a new technology and technique. *Neonat Intensive Care* 1991;**4**(3):12–15.

Felix E, Scott S, Crafton B, Geis P, Duncan T, Sewell R, *et al.* Causes of recurrence after laparoscopic hernioplasty. A multicenter study. *Surg Endosc* 1998;**12**:226–31.

Ferzli G, Kiel T. Evolving techniques in endoscopic extraperitoneal herniorrhaphy. *Surg Endosc* 1995;**9**:928–30.

Fisher KS, Matteson KM, Hammer MD. Laparoscopic cholecystectomy: the Springfield experience. *Surg Laparosc Endosc* 1993;**3**:199–203.

Fockens P, van den Brande JH, van Dullemen HM, van Lanschot JJ, Tytgat GN. Endosonographic T-staging of esophageal carcinoma: a learning curve. *Gastrointest Endosc* 1996;**44**:58–62.

Foucher G, Allieu Y, Buch N. Endoscopic treatment of carpal tunnel syndrome with Agee's technic. A report of 280 cases. *Rhumatologie* 1995;**47**(2):47–51.

Fowler DL, White SA, Anderson CA. Laparoscopic colon resection: 60 cases. *Surg Laparosc Endosc* 1995;**5**:468–71.

Friedl W, Colombo-Benkmann M, Dockter S, Machens HG, Mieck U. Gamma nail osteosynthesis of per- and subtrochanteric femoral fractures. 4 years experiences and their consequences for further implant development. *Chirurg* 1994;**65**:953–63.

Friedman RL, Fallas MJ, Carroll BJ, Hiatt JR, Phillips EH. Laparoscopic splenectomy for ITP. The gold standard. *Surg Endosc* 1996;**10**:991–5.

Gertzbein SD, Robbins SE. Accuracy of pedicular screw placement *in vivo*. *Spine* 1990;**15**:11–14.

Ghosh PK, Choudhary A, Agarwal SK, Husain T. Role of an operative score in mitral reconstruction in dominantly stenotic lesions. *Eur J Cardiothorac Surg* 1997;**11**:274–9.

Gilchrist BF, Vlessis AA, Kay GA, Swartz K, Dennis D. Open versus laparoscopic cholecystectomy: an initial analysis. *J Laparoendosc Surg* 1991;**1**:193–6.

Gimbel HV, Penno EE, van Westenbrugge JA, Ferensowicz M, Furlong MT. Incidence and management of intraoperative and early postoperative complications in 1000 consecutive laser *in situ* keratomileusis cases. *Ophthalmology* 1998;**105**:1839–47.

Giuliano AE, Kirgan DM, Guenther JM, Morton DL. Lymphatic mapping and sentinel lymphadenectomy for breast cancer. *Ann Surg* 1994;**220**:391–8.

Goldberg JD, Porter AE, Golbus MS. Current assessment of fetal losses as a direct consequence of chorionic villus sampling. *Am J Med Genet* 1990;**35**:174–7.

Goldberg SL, Renslo R, Sinow R, French WJ. Learning curve in the use of the radial artery as vascular access in the performance of percutaneous transluminal coronary angioplasty. *Cathet Cardiovasc Diagn* 1998;**44**:147–52.

Gordon MK, Lawrence-Brown MM, Hartley D, Sieunarine K, Holden A, MacSweeney ST, *et al.* A self-expanding endoluminal graft for treatment of aneurysms: results through the development phase. *Aust N Z J Surg* 1996;**66**:621–5.

Gormally SM, Clarke TA, Krishnan A, Ali M, Matthews TG. Surfactant therapy in respiratory distress syndrome: the effect of a learning curve in improving outcome. *Ir J Med Sci* 1993;**162**:458–61.

Grandjean JG, Boonstra PW, den Heyer P, Ebels T. Arterial revascularization with the right gastroepiploic artery and internal mammary arteries in 300 patients. *J Thorac Cardiovasc Surg* 1994;**107**:1309–15.

Guazzoni G, Montorsi F, Bergamaschi F, Bellinzoni P, Centemero A, Consonni P, *et al.* Open surgical revision of laparoscopic pelvic lymphadenectomy for staging of prostate cancer: the impact of laparoscopic learning curve. *J Urol* 1994;**151**:930–3.

Guillard N, Lefevre T, Spaulding C, Funck F, Py A, Chalet Y, *et al.* Coronary angiography by the left radial approach: a bicentric prospective pilot study. *Arch Mal Coeur Vaiss* 1997;**90**:1349–55.

Harkki-Siren P, Sjoberg J. Evaluation and the learning curve of the first one hundred laparoscopic hysterectomies. *Acta Obstetr Gynecol Scand* 1995;**74**:638–41.

Harris RD, Schned AR, Heaney JA. Staging of prostate cancer with endorectal MR imaging: lessons from a learning curve. *Radiographics* 1995;**15**:813–29.

Harston WE, Tilley S, Rodeheffer R. Safety and success of the beginning percutaneous transluminal coronary angioplasty program using the steerable guidewire system. *Am J Cardiol* 1986;**57**:717–20.

Hatzinikolaou H, Smeets JM, Rodriguez L-M, Vrouchos G, Grecas G, Lonkandwala Y, *et al.* Results of radiofrequency catheter ablation of accessory pathways in 301 consecutive symptomatic patients. Success rate per year. *Hellenic J Cardiol* 1995;**36**:152–8.

Hawasli A, Lloyd LR. Laparoscopic cholecystectomy. The learning curve: report of 50 patients. *Am Surg* 1991;**57**:542–4.

Hawatmeh AS, Walters FP, Wooster CA. Early clinical experience with percutaneous cryosurgical ablation of the prostate. *Missouri Med* 1995;**92**:705–9.

Heintz A, Junginger T. Learning curve after 50 retroperitoneoscopic adrenalectomies. *Minim Invasive Ther Allied Technol* 1998;**7**:273–4.

Heinz G, Kratochwill C, Schmid S, Kreiner G, Siostrzonek P, Pacher R, *et al.* Sinus node dysfunction after orthotopic heart transplantation: the Vienna experience 1987–1993. *Pac Clin Electrophysiol* 1994;**17**:2057–63.

Heller JG, Carlson GD, Abitbol JJ, Garfin SR. Anatomic comparison of the Roy–Camille and Magerl techniques for screw placement in the lower cervical spine. *Spine* 1991;**16**(10 suppl):S552–7.

Hershman MJ, Rosin RD. Laparoscopic laser cholecystectomy: our first 200 patients. *Ann R Coll Surg Engl* 1992;**74**:242–7.

Higashihara E, Baba S, Nakagawa K, Murai M, Go H, Takeda M, *et al.* Learning curve and conversion to open surgery in cases of laparoscopic adrenalectomy and nephrectomy. *J Urol* 1998;**159**:650–3.

Hindricks G. Incidence of complete atrioventricular block following attempted radiofrequency catheter modification of the atrioventricular node in 880 patients. Results of the multicenter European radiofrequency survey (MERFS). *Eur Heart J* 1996;**17**:82–8.

Horvath KD, Gray D, Benton L, Hill J, Swanstrom LL. Operative outcomes of minimally invasive saphenous vein harvest. *Am J Surg* 1998;**175**:391–5.

Hughes GB. The learning curve in stapes surgery. *Laryngoscope* 1991;**101**:1280–4.

Hunter JG, Sackier JM, Berci G. Training in laparoscopic cholecystectomy. Quantifying the learning curve. *Surg Endosc* 1994;**8**:28–31.

Hvass U, Chatel D, Assayag P, Ouroudji M, Pansard Y, Lenormand C, *et al.* The O'Brien–Angell stentless porcine valve: early results with 150 implants. *Ann Thorac Surg* 1995;**60**(2 suppl):S414–17.

Irazuzta J, Zhang J, Pandit S. Decrease in costs for management of lower airway disease in the pediatric intensive care unit. *South Med J* 1998;**91**:655–9.

Ivancev K, Malina M, Lindblad B, Chuter TA, Brunkwall J, Lindh M, *et al.* Abdominal aortic aneurysms: experience with the Ivancev–Malmo endovascular system for aortomonoiliac stent-grafts. *J Endovasc Surg* 1997;**4**:242–51.

Jacobs LK, Shayani V, Sackier JM. Determination of the learning curve of the AESOP robot. *Surg Endosc* 1997;**11**:54–5.

Janetschek G, Hobisch A, Holtl L, Bartsch G. Retroperitoneal lymphadenectomy for clinical stage I nonseminomatous testicular tumor: laparoscopy versus open surgery and impact of learning curve. *J Urol* 1996;**156**:89–93.

Jaroudi KA, Arora M, Hamilton CM, Sieck UV, Willemsen W, Sheth KV, *et al.* Results of 2426 *in vitro* fertilization cycles from King Faisal Specialist Hospital and Research Centre, 1986–1992. *Ann Saudi Med* 1995;**15**:36–42.

Jawad AJ, Kurban K, Ei-Bakry A, Al-Rabeeah A, Seraj M, Ammar A, *et al.* Laparoscopic cholecystectomy for cholelithiasis during infancy and childhood: cost analysis and review of current indications. *World J Surg* 1998;**22**:69–74.

John TG, Banting SW, Pye S, Paterson-Brown S, Garden OJ. Preliminary experience with intracorporeal laparoscopic ultrasonography using a sector scanning probe. A prospective comparison with intraoperative cholangiography in the detection of choledocholithiasis. *Surg Endosc* 1994;**8**:1176–80.

Johnson C, Roberts JT. Clinical competence in the performance of fiberoptic laryngoscopy and endotracheal intubation: a study of resident instruction. *J Clin Anesth* 1989;**1**:344–9.

Jones RA. Complications of laparoscopic hysterectomy: 250 cases. *Gynaecol Endosc* 1995;**4**:95–9.

Jones RM, Fletcher DR, MacLellan DG, Lowe AW, Hardy KJ. Laparoscopic cholecystectomy: initial experience. *Aust N Z J Surg* 1991;**61**:261–6.

Jowell PS, Baillie J, Branch MS, Affronti J, Browning CL, Bute BP. Quantitative assessment of procedural competence. A prospective study of training in endoscopic retrograde cholangiopancreatography. *Ann Intern Med* 1996;**125**:983–9.

Kald A, Anderberg B, Smedh K, Karlsson M. Transperitoneal or totally extraperitoneal approach in laparoscopic hernia repair: results of 491 consecutive herniorrhaphies. *Surg Laparosc Endosc* 1997;**7**:86–9.

Kane RL, Lurie N, Borbas C, Morris N, Flood S, McLaughlin B, *et al.* The outcomes of elective laparoscopic and open cholecystectomies. *J Am Coll Surg* 1995;**180**:136–45.

Kaplan EL, Reid HF, Johnson DR, Kunde CA. Rapid antigen detection in the diagnosis of group A streptococcal pyoderma: influence of a "learning curve effect" on sensitivity and specificity. *Pediatr Infect Dis J* 1989;**8**:591–3.

Katritsis D, Webb-Peploe MM. Cannulation of the coronary sinus via the femoral vein – a new technique. *Clin Cardiol* 1997;**20**:446–8.

Keeling NJ, Menzies D, Motson RW. Laparoscopic exploration of the common bile duct – beyond the learning curve. *Surg Endosc* 1999;**13**:109–12.

Keerl R, Weber R, Drees G, Draf W. Individual learning curves with reference to endonasal micro-endoscopic pan-sinus operation. *Laryngorhinootologie* 1996;**75**:338–43.

Keerl R, Schauss F, Weber R. Role of the multimedia technology in the evolution of learning of endonasal surgery of the sinuses. *Rev Laryngol Otol Rhinol (Bord)* 1997;**118**:129–32.

Kelsey SF, Mullin SM, Detre KM, Mitchell H, Cowley MJ, Gruentzig AR, *et al.* Effect of investigator experience on percutaneous transluminal coronary angioplasty. *Am J Cardiol* 1984;**53**:56–64C.

Kerbl K, Clayman RV, Petros JA, Chandhoke PS, Gill IS. Staging pelvic lymphadenectomy for prostate cancer: a comparison of laparoscopic and open techniques. *J Urol* 1993;**150**:396–8.

Kim PC, Wesson D, Superina R, Filler R. Laparoscopic cholecystectomy versus open cholecystectomy in children: which is better? *J Pediatr Surg* 1995;**30**:971–3.

Kimball BP, Bui S, Cohen EA, Carere RG, Adelman AG. Early experience with directional coronary atherectomy: documentation of the learning curve. *Can J Cardiol* 1993;**9**:177–85.

Klan R, Dieckmann KP, Meier T, Handke A. Laparoscopic vs. open surgical lymphadenectomy in prostate cancer. Methodological comparison. *Urologe A* 1994;**33**:128–32.

Klee LWG. Laparoscopic spermatic vein ligation in dogs. *J Endourol* 1991;**5**:341–4.

Kleman JM, Castle LW, Kidwell GA, Maloney JD, Morant VA, Trohman RG, *et al.* Nonthoracotomy- versus thoracotomy-implantable defibrillators. Intention-to-treat comparison of clinical outcomes. *Circulation* 1994;**90**:2833–42.

Kockerling F, Schneider C, Reymond MA, Scheidbach H, Konradt J, Barlehner E, *et al.* Early results of a prospective multicenter study on 500 consecutive cases of laparoscopic colorectal surgery. *Surg Endosc* 1998;**12**:37–41.

Konrad C, Schupfer G, Wietlisbach M, Gerber H. Learning manual skills in anesthesiology: is there a recommended number of cases for anesthetic procedures? *Anesth Analg* 1998;**86**:635–9.

Kopacz DJ, Neal JM, Pollock JE. The regional anesthesia "learning curve". What is the minimum number of epidural and spinal blocks to reach consistency? *Reg Anesth* 1996;**21**:182–90.

Kopelman D, Hashmonai M, Ehrenreich M, Assalia A. Thoracoscopic sympathectomy for hyperhidrosis: is there a learning curve? *Surg Laparosc Endosc* 1998;**8**:370–5.

Krahenbuhl L, Frei E. Laparoscopic inguinal hernia repair: an individualized approach? *Dig Surg* 1997; **14**:82–7.

Kullman E, Borch K, Svanvik J, Anderberg B. Differences in outcome of acute and elective laparoscopic cholecystectomy. *Dig Surg* 1997;**14**:398–403.

Laffel GL, Barnett AI, Finkelstein S, Kaye MP. The relation between experience and outcome in heart transplantation. *N Engl J Med* 1992;**327**:1220–5.

Lang GS, Ruckle HC, Hadley HR, Lui PD, Stewart SC. One hundred consecutive laparoscopic pelvic lymph node dissections: comparing complications of the first 50 cases to the second 50 cases. *Urology* 1994;**44**:221–5.

Langeron O, Lenfant F, Aubrun F, Riou B, Coriat P. Assessment of the learning curve of a new light wand device (Trachlight™) for tracheal intubation. *Ann Fr Anesth Reanim* 1997;**16**:229–33.

Lawrence K, McWhinnie D, Goodwin A, Doll H, Gordon A, Gray A, *et al.* Randomised controlled trial of laparoscopic versus open repair of inguinal hernia: early results. *BMJ* 1995;**311**:981–5.

Lay L, Zamboni WA, Texter JH, Zook EG. Analysis of hypospadias and fistula repair. *Am Surg* 1995;**61**:537–8.

Lefevre T, Louvard Y, Morice MC, Dumas P, Loubeyre C, Krol M, *et al.* Stenting of bifurcation lesions: influence of the learning curve on the seven-month outcome [abstract]. *Eur Heart J* 1998;**19**(suppl):626.

Leibl B, Schwarz J, Daubler P, Kraft K, Bittner R. Endoscopic hernia surgery (TAPP) – gold standard in management of recurrent hernias? *Chirurg* 1996;**67**:1226–30.

Leibl BJ, Schmedt CG, Schwarz J, Daubler P, Kraft K, Schlossnickel B, *et al.* A single institution's experience with transperitoneal laparoscopic hernia repair. *Am J Surg* 1998;**175**:446–51.

Levin DC, Ganz P, Friedman P, Abben R, Garnic JD, Boxt LM. Percutaneous transluminal coronary angioplasty with an over-the-wire system. *Radiology* 1985;**155**:323–6.

Liem MS, van Steensel CJ, Boelhouwer RU, Weidema WF, Clevers GJ, Meijer WS, *et al.* The learning curve for totally extraperitoneal laparoscopic inguinal hernia repair. *Am J Surg* 1996;**171**:281–5.

Liem MS, van der GY, van Steensel CJ, Boelhouwer RU, Clevers GJ, Meijer WS, *et al.* Comparison of conventional anterior surgery and laparoscopic surgery for inguinal-hernia repair. *N Engl J Med* 1997;**336**:1541–7.

Lin RT, Maloney RK. Flap complications associated with lamellar refractive surgery. *Am J Ophthalmol* 1999; **127**:129–36.

Long JP, Fallick ML, LaRock DR, Rand W. Preliminary outcomes following cryosurgical ablation of the prostate in patients with clinically localized prostate carcinoma. *J Urol* 1998;**159**:477–84.

Louvard Y, Pezzano M, Scheers L, Koukoui F, Marien C, Benaim R, *et al.* Coronary angiography by a radial artery approach: feasibility, learning curve. One operator's experience. *Arch Mal Coeur Vaiss* 1998;**91**:209–15.

Madhavan KK, Forsythe JLR, Garden OJ. Starting a new clinical service – is a learning curve a necessity? *Br J Surg* 1998;**85**:5.

Manga P, Friedman B, Singh S, Pocock WA, Barlow JB. Percutaneous balloon mitral valvuloplasty using trefoil or bifoil balloon catheter: immediate results and emphasis on the learning curve. *J Cardiovasc Technol* 1992;**10**:241–8.

Marks SC. Learning curve in endoscopic sinus surgery. *Otolaryngol Head Neck Surg* 1999;**120**:215–18.

Marshall JB. Technical proficiency of trainees performing colonoscopy: a learning curve. *Gastrointest Endosc* 1995;**42**:287–91.

Masket S, Gokmen F. Efficacy and safety of intracameral lidocaine as a supplement to topical anesthesia. *J Cataract Refract Surg* 1998;**24**:956–60.

May J, White GH, Waugh R, Stephen MS, Chaufour X, Yu W, *et al.* Adverse events after endoluminal repair of abdominal aortic aneurysms: a comparison during two successive periods of time. *J Vasc Surg* 1999;**29**:32–7.

McIvor NP, Freeman JL, Salem S, Elden L, Noyek AM, Bedard YC. Ultrasonography and ultrasound-guided fine-needle aspiration biopsy of head and neck lesions: a surgical perspective. *Laryngoscope* 1994;**104**:669–74.

McMullen K, Hicks TC, Ray JE, Gathright JB, Timmcke AE. Complications associated with ileal pouch-anal anastomosis. *World J Surg* 1991;**15**:763–6.

Meehan JJ, Georgeson KE. The learning curve associated with laparoscopic antireflux surgery in infants and children. *J Pediatr Surg* 1997;**32**:426–9.

Meinke AK, Kossuth T. What is the learning curve for laparoscopic appendectomy? *Surg Endosc* 1994;**8**:371–5.

Minowada S, Higashihara E, Kameyama S, Oshi M, Homma Y, Aso Y. Advantage of a smaller caliber fiberscope and learning curve on transurethral lithotripsy. *J Urol* 1992;**147**:1243–4.

Moffat DA, Hardy DG, Grey PL, Baguley DM. The operative learning curve and its effect on facial nerve outcome in vestibular schwannoma surgery. *Am J Otol* 1996;**17**:643–7.

Molloy M, Archer SB, Hasselgren PO, Dalton BJ, Bower RH. Cholangiography during laparoscopic cholecystectomy: cumulative sum analysis of an institutional learning curve. *Gastroenterology* 1998;**114**:S0163.

Moore MJ, Bennett CL. The learning curve for laparoscopic cholecystectomy. *Am J Surg* 1995;**170**:55–9.

Moore RG, Averch TD, Schulam PG, Adams JB, Chen RN, Kavoussi LR. Laparoscopic pyeloplasty: experience with the initial 30 cases. *J Urol* 1997;**157**:459–62.

Morgenstern L, McGrath MF, Carroll BJ, Paz-Partlow M, Berci G. Continuing hazards of the learning curve in laparoscopic cholecystectomy. *Am Surg* 1995;**61**:914–18.

Morris N, Borbas C, Flood S, Kane R, McLaughlin B, Nemanich G, *et al.* The laparoscopic learning curve: experience with elective cholecystectomy [abstract]. Annual Meeting of International Society of Technology Assessment in Health Care 1994; Baltimore, MD: 148.

Mounce RE, Nakamuta H, Lovejoy C. Canal master instrumentation: an *in vitro* study of separation frequency. *J Endodont* 1993;**19**:1–3.

Munro W, Brancatisano R, Adams IP, Falk GL. Complications of laparoscopic fundoplication: the first 100 patients. *Surg Laparosc Endosc* 1996;**6**:421–3.

Mussurakis S, Buckley DL, Coady AM, Turnbull LW, Horsman A. Observer variability in the interpretation of contrast enhanced MRI of the breast. *Br J Radiol* 1996;**69**:1009–16.

Nakagawa K, Murai M, Deguchi N, Baba S, Tachibana M, Nakamura K, *et al.* Laparoscopic adrenalectomy: clinical results in 25 patients. *J Endourol* 1995;**9**:265–7.

Nash A, Burrell CJ, Ring NJ, Marshall AJ. Evaluation of an ultrasonically guided venepuncture technique for the placement of permanent pacing electrodes. *Pacing Clin Electrophysiol* 1998;**21**:452–5.

Nataf P, Kirsch W, Hill AC, Anton T, Zhu YH, Ramadan R, *et al.* Nonpenetrating clips for coronary anastomosis. *Ann Thorac Surg* 1997;**63**(6 suppl):S135–7.

Nathanson MH, Gajraj NM, Newson CD. Tracheal intubation in a manikin: comparison of supine and left lateral positions. *Br J Anaesth* 1994;**73**:690–1.

Ng DT, Rowe NA, Francis IC, Kappagoda MB, Haylen MJ, Schumacher RS, *et al.* Intraoperative complications of 1000 phacoemulsification procedures: a prospective study. *J Cataract Refract Surg* 1998;**24**:1390–5.

Nobuyoshi M. Percutaneous transluminal coronary angioplasty: technical and anatomical considerations. *J Cardiogr Suppl* 1986;**10**:55–61.

Nottle PD, Wale RJ, Johnson WR. Percutaneous laparoscopic cholecystectomy: the first fifty. *Aust N Z J Surg* 1991;**61**:254–60.

O'Donohoe MK, Sultan S, Colgan MP, Moore DJ, Shanik GD. Outcome of the first 100 femoropopliteal angioplasties performed in the operating theatre. *Eur J Vasc Endovasc Surg* 1999;**17**:66–71.

Orlando R, Russell JC, Lynch J, Mattie A. Laparoscopic cholecystectomy. A statewide experience. *Arch Surg* 1993;**128**:494–8.

O'Sullivan DC, Averch TD, Cadeddu JA, Moore RG, Beser N, Breitenbach C, *et al.* Teleradiology in urology: comparison of digital image quality with original radiographic films to detect urinary calculi. *J Urol* 1997; **158**:2216–20.

Ou CS, Beadle E, Presthus J, Smith M. A multicenter review of 839 laparoscopic-assisted vaginal hysterectomies. *J Am Assoc Gynecol Laparosc* 1994;**1**:417–22.

Panetta TF, Hunt JP, Buechter KJ, Pottmeyer A, Batti JS. Duplex ultrasonography versus arteriography in the diagnosis of arterial injury: an experimental study. *J Trauma* 1992;**33**:627–35.

Parikh D, Johnson M, Chagla L, Lowe D, McCulloch P. D2 gastrectomy: lessons from a prospective audit of the learning curve. *Br J Surg* 1996;**83**:1595–9.

Parry BR, Williams SM. Competency and the colonoscopist: a learning curve. *Aust N Z J Surg* 1991;**61**:419–22.

Perry KGJ, Hess LW, Roberts WE, Allbert JR, Floyd RC, McCaul JF, *et al.* Cordocentesis (funipuncture) by maternal–fetal fellows: the learning curve. *Fetal Diagn Ther* 1991;**6**:87–92.

Peters JH, Ellison EC, Innes JT, Liss JL, Nichols KE, Lomano JM, *et al.* Safety and efficacy of laparoscopic cholecystectomy. A prospective analysis of 100 initial patients. *Ann Surg* 1991;**213**:3–12.

Phipps JH, John M, Lewis BV. Laparoscopic treatment of tubal ectopic pregnancy – a series of 62 cases. *Gynaecol Endosc* 1992;**1**:191–4.

Picano E, Lattanzi F, Orlandini A, Marini C, l'Abbate A. Stress echocardiography and the human factor: the importance of being expert. *J Am Coll Cardiol* 1991;**17**:666–9.

Poulin EC, Mamazza J. Laparoscopic splenectomy: lessons from the learning curve. *Can J Surg* 1998;**41**:28–36.

Prasad S. Phacoemulsification learning curve: experience of two junior trainee ophthalmologists. *J Cataract Refract Surg* 1998;**24**:73–7.

Prince RB, Tax RL, Miller DH. Conversion to small-incision phacoemulsification: experience with the first 50 eyes. *J Cataract Refract Surg* 1993;**19**:246–50.

Querleu D, Lanvin D, Elhage A, Henry-Buisson B, Leblanc E. An objective experimental assessment of the learning curve for laparoscopic surgery: the example of pelvic and para-aortic lymph node dissection. *Eur J Obstet Gynecol Reprod Biol* 1998;**81**:55–8.

Quilici PJ, Greaney EM, Quilici J, Anderson S. Transabdominal preperitoneal laparoscopic inguinal herniorrhaphy: results of 509 repairs. *Am Surg* 1996;**62**:849–52.

Radomski SB, Herschorn S. Laparoscopic Burch bladder neck suspension: early results. *J Urol* 1996;**155**:515–18.

Rae AJ, Belzberg A, Cleator IM, Caglar M. Use of the ^{14}C breath test in the treatment of *Helicobacter pylori*. *Can J Gastroenterol* 1995;**9**:191–4.

Rassweiler J, Fornara P, Weber M, Janetschek G, Fahlenkamp D, Henkel T, *et al.* Laparoscopic nephrectomy: the experience of the laparoscopy working group of the German Urologic Association. *J Urol* 1998;**160**:18–21.

Rassweiler JJ, Seemann O, Frede T, Henkel TO, Alken P. Retroperitoneoscopy: experience with 200 cases. *J Urol* 1998;**160**:1265–9.

Rassweiler JJ, Seemann O, Henkel T, Tschada R, Potempa D, Alken P. Retroperitoneoscopy. Technique and experiences with the first 100 patients. *Urologe A* 1996;**35**:185–95.

Rau HG, Buttler E, Meyer G, Schardey HM, Schildberg FW. Laparoscopic liver resection compared with conventional partial hepatectomy – a prospective analysis. *Hepatogastroenterology* 1998;**45**:2333–8.

Ravintharan T, Lim PH, Chng HC. Ureterorenoscopy: factors influencing success. *Singapore Med J* 1991;**32**:151–3.

Regan JJ, Yuan HS, McAfee PC. Laparoscopic fusion of the lumbar spine: minimally invasive spine surgery – a prospective multicenter study evaluating open and laparoscopic lumbar fusion. *Spine* 1999;**24**:402–11.

Rege RV, Joehl RJ. A learning curve for laparoscopic splenectomy at an academic institution. *J Surg Res* 1999;**81**:27–32.

Reissman P, Cohen S, Weiss EG, Wexner SD. Laparoscopic colorectal surgery: ascending the learning curve. *World J Surg* 1996;**20**:277–81.

Resch H, Povacz P, Wambacher M, Sperner G, Golser K. Arthroscopic extra-articular Bankart repair for the treatment of recurrent anterior shoulder dislocation. *Arthroscopy* 1997;**13**:188–200.

Richardson MC, Bell G, Fullarton GM. Incidence and nature of bile duct injuries following laparoscopic cholecystectomy: an audit of 5913 cases. *Br J Surg* 1996;**83**:1356–60.

Rihal CS, Nishimura RA, Holmes DR, Jr. Percutaneous balloon mitral valvuloplasty: the learning curve. *Am Heart J* 1991;**122**:1750–6.

Roon AJ, Hoogerwerf D. Intraoperative arteriography and carotid surgery. *J Vasc Surg* 1992;**16**:239–43.

Rosen CA. Complications of phonosurgery: results of a national survey. *Laryngoscope* 1998;**108**:1697–703.

Rosen DMB, Cario GM, Carlton MA, Lam AM, Chapman M. An assessment of the learning curve for laparoscopic and total laparoscopic hysterectomy. *Gynaecol Endosc* 1998;**7**:289–93.

Rothenberg SS. Experience with 220 consecutive laparoscopic Nissen fundoplications in infants and children. *J Pediatr Surg* 1998;**33**:274–7.

Rutter MJ, Furneaux CE, Morton RP. Craniofacial resection of anterior skull base tumours: factors contributing to success. *Aust N Z J Surg* 1998;**68**:350–3.

Salai M, Mintz Y, Giveon U, Chechik A, Horoszowski H. The "learning curve" of total hip arthroplasty. *Arch Orthopaed Trauma Surg* 1997;**116**:420–2.

Sanders R, Fortin P, DiPasquale T, Walling A. Operative treatment in 120 displaced intraarticular calcaneal fractures. Results using a prognostic computed tomography scan classification. *Clin Orthop* 1993;(290):87–95.

Sariego J, Spitzer L, Matsumoto T. The "learning curve" in the performance of laparoscopic cholecystectomy. *Int Surg* 1993;**78**:1–3.

Sarli L, Pietra N, Sansebastiano G, Cattaneo G, Costi R, Grattarola M, *et al.* Reduced postoperative morbidity after elective laparoscopic cholecystectomy: stratified matched case–control study. *World J Surg* 1997;**21**:872–9.

Sathe S, Vohra J, Chan W, Wong J, Gerloff J, Riters A, *et al.* Radiofrequency catheter ablation for paroxysmal supraventricular tachycardia: a report of 135 procedures. *Aust N Z J Med* 1993;**23**:317–24.

Savage DD, Garrison RJ, Kannel WB, Anderson SJ, Feinleib M, Castelli WP, *et al.* Considerations in the use of echocardiography in epidemiology. The Framingham Study. *Hypertension* 1987;**9**:II40–4.

Schertz RD, Baskin WN, Frakes JT. Flexible fiberoptic sigmoidoscopy training for primary care physicians: results of a 5-year experience. *Gastrointest Endosc* 1989;**35**:316–20.

Schlup MT, Williams SM, Barbezat GO. ERCP: a review of technical competency and workload in a small unit. *Gastrointest Endosc* 1997;**46**:48–52.

Schneider A, Krause N, Kuhne-Heid R, Kamprath S, Endisch U, Merker A, *et al.* Laparoscopic para-aortic and pelvic lymph node excision – initial experiences and development of a technique. *Zentralbl Gynakol* 1996;**118**:498–504.

Schneider JE, Mann T, Cubeddu MG, Arrowood ME. Transradial coronary stenting: a United States experience. *J Invasive Cardiol* 1997;**9**:569–74.

Scholefield JH, Berry DP, Armitage NM, Wastie ML. Magnetic resonance imaging in the management of fistula *in ano*. *Int J Colorect Dis* 1997;**12**:276–9.

Schuman ES, Standage BA, Ragsdale JW, Gross GF. Reinforced versus nonreinforced polytetrafluoro-ethylene grafts for hemodialysis access. *Am J Surg* 1997;**173**:407–10.

See WA, Cooper CS, Fisher RJ. Predictors of laparoscopic complications after formal training in laparoscopic surgery. *JAMA* 1993;**270**:2689–92.

Senagore AJ, Luchtefeld MA, Mackeigan JM. What is the learning curve for laparoscopic colectomy? *Am Surg* 1995;**61**:681–5.

Serra A, Bonan R, Lefevre T, Cequier A, Petitclerc R, Leclerc Y, *et al.* Percutaneous mitral valvuloplasty. An analysis of the immediate results. *Rev Esp Cardiol* 1991;**44**:174–83.

Seward HC, Dalton R, Davis A. Phacoemulsification during the learning curve: risk/benefit analysis. *Eye* 1993;**7**:164–8.

Sheridan MB, Nicholson DA, Martin DF. Trans-abdominal ultrasonography as the primary investigation in patients with suspected Crohn's disease or recurrence: a prospective study. *Clin Radiol* 1993;**48**:402–4.

Silver RK, MacGregor SN, Sholl JS, Hobart ED, Waldee JK. An evaluation of the chorionic villus sampling learning curve. *Am J Obstet Gynecol* 1990;**163**:917–22.

Simons AJ, Anthone GJ, Ortega AE, Franklin M, Fleshman J, Geis WP, *et al.* Laparoscopic-assisted colectomy learning curve. *Dis Colon Rectum* 1995;**38**:600–3.

Sinclair BG, Sandor GG, Farquharson DF. Effectiveness of primary level antenatal screening for severe congenital heart disease: a population-based assessment. *J Perinatol* 1996;**16**:336–40.

Smith DB, Larsson JL. The impact of learning on cost: the case of heart transplantation. *Hosp Health Serv Admin* 1989;**34**:85–97.

Smith JE, Jackson AP, Hurdley J, Clifton PJ. Learning curves for fibreoptic nasotracheal intubation when using the endoscopic video camera. *Anaesthesia* 1997;**52**:101–6.

Smith RS, Kern SJ, Fry WR, Helmer SD. Institutional learning curve of surgeon-performed trauma ultrasound. *Arch Surg* 1998;**133**:530–5.

Solomon MJ, McLeod RS, Cohen EK, Simons ME, Wilson S. Reliability and validity studies of endoluminal ultrasonography for anorectal disorders. *Dis Colon Rectum* 1994;**37**:546–51.

Spangenberger W, Klein J, Troidl H. Laparoscopic cholecystectomy – initial experiences and results. *Langenbecks Arch Chir Suppl Kongressbd* 1990;1361–8.

Starkes JL, Payk I, Hodges NJ. Developing a standardized test for the assessment of suturing skill in novice microsurgeons. *Microsurgery* 1998;**18**:19–22.

Steingart RM, Wassertheil-Smoller S, Budner N, Tobin J, Wachspress J, Lense L, *et al.* The clinical use of nuclear exercise tests. *Int J Technol Assess Health Care* 1988;**4**:613–22.

Still RM, Walsh DJ. A pilot study to assess the learning curve in transcervical endometrial resection; possible implications for postgraduate accreditation. *Gynaecol Endosc* 1992;**1**:111–13.

Stoddard MF, Hammons RT, Longaker RA. Doppler transesophageal echocardiographic determination of aortic valve area in adults with aortic stenosis. *Am Heart J* 1996;**132**:337–42.

Stolier AJ. Stereotactic breast biopsy: a surgical series. *J Am Coll Surg* 1997;**185**:224–8.

Stone GW, Rutherford BD, McConahay DR, Johnson WLJ, Giorgi LV, Ligon RW, *et al.* Procedural outcome of angioplasty for total coronary artery occlusion: an analysis of 971 lesions in 905 patients. *J Am Coll Cardiol* 1990;**15**:849–56.

Stromqvist B, Nilsson LT, Thorngren KG. Femoral neck fracture fixation with hook-pins. 2-year results and learning curve in 626 prospective cases. *Acta Orthopaed Scand* 1992;**63**:282–7.

Surya BV, Provet J, Dalbagni G, Johanson KE, Brown J. Experience with potency preservation during radical prostatectomy. Significance of learning curve. *Urology* 1988;**32**:498–501.

Sutton DN, Wayman J, Griffin SM. Learning curve for oesophageal cancer surgery. *Br J Surg* 1998;**85**:1399–402.

Tavola A, Carones F, Galli L, Fontanella G, Brancato R. The learning curve in myopic photorefractive keratectomy. *J Refract Corneal Surg* 1994;**10**(2 suppl):S188–93.

Taylor TA, Mander J, Manning K, Lawrence-Brown MM. Criteria audit of percutaneous nephrolithotomies. *Aust N Z J Surg* 1990;**60**:849–53.

Teoh TG. Effect of learning curve on the outcome of external cephalic version. *Singapore Med J* 1997;**38**:323–5.

Terachi T, Matsuda T, Terai A, Ogawa O, Kakehi Y, Kawakita M, *et al.* Transperitoneal laparoscopic adrenalectomy: experience in 100 patients. *J Endourol* 1997;**11**:361–5.

Tocchi A, Liotta G, Mazzoni G, Lepre L, Costa G, Maggiolini F, *et al.* Learning curve for "tension-free" reparation of inguinal hernia. *G Chir* 1998;**19**:199–203.

Toogood GJ, Torrie EP, Magee TR, Galland RB. Early experience with stenting for iliac occlusive disease. *Eur J Vasc Endovasc Surg* 1998;**15**:165–8.

Torella F, Biyani CS, Cade D, Powell CS. Laparoscopic nephrectomy for benign disease: the early learning curve. *Minim Invasive Ther* 1997;**6**:73–6.

Toy FK, Moskowitz M, Smoot RTJ, Pleatman M, Bagdasarian A, Polito W, *et al.* Results of a prospective multicenter trial evaluating the ePTFE peritoneal onlay laparoscopic inguinal hernioplasty. *J Laparoendosc Surg* 1996;**6**:375–86.

Trastek VF, Deschamps C, Allen MS, Miller DL, Pairolero PC, Thompson AM. Uncut Collis–Nissen fundoplication: learning curve and long-term results. *Ann Thorac Surg* 1998;**66**:1739–43.

Trias M, Targarona EM, Espert JJ, Balague C. Laparoscopic surgery for splenic disorders – lessons learned from a series of 64 cases. *Surg Endosc* 1998;**12**:66–72.

Tucker PA, Ferguson JJ, Harlan M, Gaos CM, Massumi A. Balloon mitral valvuloplasty: clinical experience at the Texas Heart Institute. *Tex Heart Inst J* 1992;**19**:270–7.

Turjman F, Massoud TF, Sayre J, Vinuela F. Predictors of aneurysmal occlusion in the period immediately after endovascular treatment with detachable coils: a multivariate analysis. *Am J Neuroradiol* 1998;**19**:1645–51.

Unger SW, Glick GL, Landeros M, Cosgrove J, Crooms J, Deziel D, *et al.* Cystic duct leak after laparoscopic cholecystectomy – a multi-institutional study. *Surg Endosc* 1996;**10**:1189–93.

Ungerleider RM, Greeley WJ, Kanter RJ, Kisslo JA. The learning curve for intraoperative echocardiography during congenital heart surgery. *Ann Thorac Surg* 1992;**54**:691–6.

Vanek VW, Rhodes R, Dallis DJ. Results of laparoscopic versus open cholecystectomy in a community hospital. *South Med J* 1995;**88**:555–66.

Venta LA, Salomon CG, Flisak ME, Venta ER, Izquierdo R, Angelats J. Sonographic signs of breast implant rupture. *Am J Roentgenol* 1996;**166**:1413–19.

Vignon P, Rambaud G, Francois B, Cornu E, Gastinne H. Transoesophageal echocardiography for diagnosis of post-traumatic injuries to major intrathoracic vessels in 150 patients: influence of learning curve. *Ann Fr Anesth Reanim* 1998;**17**:1206–16.

Voitk AJ. The learning curve in laparoscopic inguinal hernia repair for the community general surgeon. *Can J Surg* 1998;**41**:446–50.

Vossen C, van Ballaer P, Shaw RW, Koninckx PR. Effect of training on endoscopic intracorporeal knot tying. *Hum Reprod* 1997;**12**:2658–63.

Wass DM, Brown GA, Warren PS, Saville TA. Completed follow-up of 1000 consecutive transcervical chorionic villus samplings performed by a single operator. *Aust N Z J Obstet Gynaecol* 1991;**31**:240–5.

Watkins JL, Etzkorn KP, Wiley TE, DeGuzman L, Harig JM. Assessment of technical competence during ERCP training. *Gastrointest Endosc* 1996;**44**:411–15.

Watson DI, Mathew G, Williams JA. Impact of laparoscopic cholecystectomy in a major teaching hospital: clinical and hospital outcomes. *Med J Aust* 1995;**163**:527–30.

Watson DI, Baigrie RJ, Jamieson GG. A learning curve for laparoscopic fundoplication. Definable, avoidable, or a waste of time? *Ann Surg* 1996;**224**:198–203.

Wijnberger, van der SC. Learning in medicine: chorionic villus sampling. *Fetal Diagn Ther* 1998; **13**(suppl 1):83.

Willems S, Chen X, Hindricks G, Kottkamp H, Rotman B, Haverkamp W, *et al.* Radiofrequency ablation of AV-nodal re-entry tachycardia: experience with selective fast pathway ablation. *Z Kardiol* 1994;**83**:165–72.

Willems TP, van Herwerden LA, Steyerberg EW, Taams MA, Kleyburg VE, Hokken RB, *et al.* Subcoronary implantation or aortic root replacement for human tissue valves: sufficient data to prefer either technique? *Ann Thorac Surg* 1995;**60**(2 suppl):S83–6.

Windsor JA, Pong J. Laparoscopic biliary injury: more than a learning curve problem. *Aust N Z J Surg* 1998;**68**:186–9.

Winfield HN, Hamilton BD, Bravo EL, Novick AC. Laparoscopic adrenalectomy: the preferred choice? A comparison to open adrenalectomy. *J Urol* 1998;**160**:325–9.

Wishner JD, Baker JWJ, Hoffman GC, Hubbard GW, Gould RJ, Wohlgemuth SD, *et al.* Laparoscopic-assisted colectomy. The learning curve. *Surg Endosc* 1995;**9**:1179–83.

Witt PD, Wahlen JC, Marsh JL, Grames LM, Pilgram TK. The effect of surgeon experience on velopharyngeal functional outcome following palatoplasty: is there a learning curve? *Plast Reconstr Surg* 1998;**102**:1375–84.

Wolfe BM, Gardiner BN, Leary BF, Frey CF. Endoscopic cholecystectomy. An analysis of complications. *Arch Surg* 1991;**126**:1192–6.

Woods JR, Saywell RMJ, Nyhuis AW, Jay SJ, Lohrman RG, Halbrook HG. The learning curve and the cost of heart transplantation. *Health Serv Res* 1992;**27**:219–38.

Yang WT, Ahuja A, Tang A, Suen M, King W, Metreweli C. Ultrasonographic demonstration of normal axillary lymph nodes: a learning curve. *J Ultrasound Med* 1995;**14**:823–7.

Yim AP, Liu HP. Complications and failures of video-assisted thoracic surgery: experience from two centers in Asia. *Ann Thorac Surg* 1996;**61**:538–41.

Yu W, Whang I, Averbach A, Chang D, Sugarbaker PH. Morbidity and mortality of early postoperative intra-peritoneal chemotherapy as adjuvant therapy for gastric cancer. *Am Surg* 1998;**64**:1104–8.

Yuen PM, Rogers MS. Laparoscopic management of ovarian masses: the initial experience and learning curve. *Aust N Z J Obstet Gynaecol* 1994;**34**:191–4.

Included papers identified in non-health technology assessment literature: phase 2 of project

Atkinson RC. Mathematical learning theory. In: Hilgard ER, Bower GH, editors. Theories of learning. New York: Appleton Century Croft; 1966.

Bailey CD, McIntyre EV. Some evidence on the nature of relearning curves. *Account Rev* 1992;**67**:368–8.

Baloff N, Becker SW. On the futility of aggregating individual learning curves. *Psychol Rep* 1967;**20**:183–91.

Browne MW, Du-Toit S-HC. Models for learning data. In: Collins LM, Horn JL, editors. Best methods for the analysis of change: recent advances, unanswered questions, future directions. Washington DC: American Psychological Association; 1991. p. 47–68.

Buck JR, Cheng SJ. Instructions and feedback effects on speed and accuracy with different learning-curve models. *IIE Trans* 1993;**25**:34–47.

Cousineau D, Larochelle S. PASTIS: a program for curve and distribution analyses. *Behav Res Methods Instr Comput* 1997;**29**:542–8.

Delaney PF, Reder LM, Staszewski JJ, Ritter FE. The strategy-specific nature of improvement: the power law applies by strategy within task. *Psychol Sci* 1998;**9**:1–7.

Eyring JD, Johnson DS, Francis DJ. A cross-level units-of-analysis approach to individual differences in skill acquisition. *J Appl Psychol* 1993;**78**:805–14.

Fitts PM, Posner MI. Human performance. Belmont, CA: Brooks Cole; 1969.

Goldstein H, Healy MJR, Rasbash J. Multilevel time series models with applications to repeated measures data. *Stat Med* 1994;**13**:1643–55.

Hammond NV. Principles from the psychology of skill acquisition. In: Gardiner MM, Christie B, editors. Applying cognitive psychology to user-interface design. Chichester: Wiley; 1987. p. 163–87.

Logan GD. Shapes of reaction-time distributions and shapes of learning curves: a test of the instance theory of automaticity. *J Exp Psychol Learn Mem Cogn* 1992;**18**:883–914.

Newell A, Rosenbloom P. Mechanisms of skill acquisition and the law of practice. In: Anderson JR, editor. Cognitive skills and their acquisition. Hillsdale, NY: Lawrence Erlbaum Associates; 1981.

Ployhart RE, Hakel MD. The substantive nature of performance variability: predicting interindividual differences in intra-individual performance. *Personnel Psychol* 1998;**51**:859–901.

Spears WD. Measurement of learning and transfer through curve fitting. *Hum Factors* 1985;**27**:251–66.

VanderLeeden R. Multilevel analysis of repeated measures data. *Qual Quantity* 1998;**32**:15–29.

Zaraiskaya YI, Aleksandrova EA, Lukashev AO, Shvyrkova NA. Features of active avoidance learning in rats with streptozotocin diabetes. *Neurosci Behav Physiol* 1994;**24**:167–9.

Ziegler A, Kastner C, Blettner M. The generalised estimating equations: an annotated bibliography. *Biomet J* 1998;**40**:115–39.

Appendix 1

Literature search strategies used in phase 1 of the project†

Search terms used for the identification of studies related to learning curves

MEDLINE, EMBASE, CINAHL, HealthSTAR: Learning adj4 curve$.tw.

MEDLINE (full text): Learning adj4 curve$.tx.

Science Citation Index, Social Science Citation Index, BIOSIS: Learn* and curve* (in title, abstracts and keywords).

Cochrane Library, National Research Register: Learn* and curve* (in all fields).

NHS Economic Evaluation Database: Learning curve$ (all fields).

Current Controlled Trials: Learning (any field, any Register).

Search terms for statistical methods used in assessing the learning curve

Curve analysis; hierarch* model*; multilevel model*; random effect* model*; generali#ed estimat* equation*; latent curve model*.

Other search terms tested but rejected

Skill* and (acquir* or acquisit*); learning rate*; ((operator$ or surgeon$) adj4 experience$).tw.; calibrat* and (skill* or learn*).

† Key: $ = wildcard; adj(n) = adjacent, within n words either side of the other term; tw = textword, searches in title and abstract; tx = full text, * = wildcard; # = substitutes for one character.

Appendix 2

Search terms used in phase 1 and 2 of the project†

Search terms for the identification of articles related to learning

The following search terms were used; searches were from 1989 unless otherwise stated (dates covered by the searches are given in parentheses after the name of each database). Details of the individual databases are given in appendix 3.

MEDLINE (1966–March 1999); EMBASE (1980–February 1999); CINAHL (1982–December 1998); HealthSTAR (1975–November 1998); CAB Abstracts (to September 1998): Learning adj4 curve$.tw.

MEDLINE (full text) (1993–October 1998): Learning adj4 curve$.tx.

Science Citation Index (1981–March 1999); Social Science Citation Index (1981– March 1999); Arts and Humanities Citation Index (to July 1999); BIOSIS (1985–March 1999); Sociofile (to June 1998); PsycLIT (to June 1998): Learn* and curve* (in title, abstracts and keywords).

INSPEC (to September 1995, January–March 1997); ABI/INFORM (to September 1997), Dissertation Abstracts (to March 1998); Index to Theses Great Britain and Ireland (vols. 21–42; to December 1993): Learning curve? (title, abstract).

Ei Compendex Plus/Page One (to March 1999); ISTP (to March 1999); EconLit (to August 1998); IBSS (to June 1999); RSC bibliographic databases (to June 1999); IngentaJournals online (date of search 23 July 1998): learning curve* (title, abstract, keywords).

Search terms for known statistical methods used in assessing the learning curve

Curve analysis; hierarch* model*; multilevel model*; random effect* model*; generali#ed estimat* equation*; latent curve model*.

Curve analysis

MEDLINE (1995–May 1998); EMBASE (January–May 1998); CINAHL (1982–March 1998); HealthSTAR (1996–May 1998); Social Science Citation Index (January–July 1998); Science Citation Index (January–July 1998); BIOSIS (January–July 1998); Ei Compendex Plus/Page One (January–July 1998); IBSS (1995–July 1998); RSC (1995–July 1998); ISI Arts & Humanities Citation Index (January–July 1998); ISTP (January–July 1998): curve analysis.

Hierarchical model, multilevel model, random effects model and generalised estimating equations

MEDLINE (1995–May 1998); EMBASE (January–May 1998); CINAHL (1982–March 1998); HealthSTAR (1996–May 1998): (hierarch$ model$ or multilevel model$ or random effect$ model$ or generali#ed estimat$ equation$).tw. Social Science Citation Index (January–July 1998); Science Citation Index (January–July 1998); BIOSIS (January–July 1998); Ei Compendex Plus/Page One (January–July 1998); IBSS (January–July 1998); RSC (1995–July 1998); ISI Arts & Humanities Citation Index (January–July 1998); ISTP (January–July 1998): (hierarch* model*, multilevel model*, random effect* model*, ((generalised, generalized) estimat* equation*).ti,ab,kw.

Latent curve model

MEDLINE (1996–June 1999): latent curve$.tw. Science Citation Index (1981–July 1999): latent curve* (ti,ab,kw).

Search terms for rare events or binary data

These terms were used only in PsycLIT (1996–December 1998); Science Citation Index (1998–January 1999); Ei Compendex (1998–January 1999): (logit or log linear or logistic or probit or dichotomous or binary or categorical or

† Key $ = wildcard; adj(n) = adjacent to or within n words either side of the specified term; tw = textword, searches in title and abstract; tx = full text; * = wildcard; # = substitutes for one character; in DE = in descriptors; ti, ab, kw = searches in title, abstract and keywords.

nominal or ordinal or discrete or rare event* or poisson or error* or adverse event* or count or counts or low frequenc* or adverse outcome* or complication* or defect*) and (learn* or skill* or performanc* or experience*).

Skill acquisition – terms used and dates searched

MEDLINE (January–July 1998); EMBASE (January–May 1998); CINAHL (January–May 1998); HEALTHSTAR (January 1996–July 1998, restricted to non-MEDLINE); CAB Abstracts (January 1973–September 1998): Skill$ adj2 (acquir$ or acquisit$).
ISI Science Citation Index (January–July 1998); ISI Social Science Citation Index (January–July 1998); ISI Arts and Humanities Citation Index (January–July 1998); ISTP (January–July 1998); IBSS (January–July 1998); Ei Compendex Plus (January–July 1998); Ei Page One (January–July 1998); BIOSIS (January–July 1998); RSC (January–July 1998): Skill* + (acquir*, acquisit*).
PsycLIT (January 1993–June 1998): (skill* near2 acquisit*) or (skill* near2 acquir*) or (skill* learning) – all fields including controlled vocabulary but only first 50 assessed out of 851.
Sociofile (1974—June 1998 but only first 20 assessed out of 188); ECONLIT (1969–August 1998): (skill* near2 acquisit*) or (skill* near2 acquir*).
ABI/INFORM (January 1996–September 1997): (skill? Acquir?) or (skill? Acquis?).
INSPEC (January–March 1997): skill* near (acquir* or acquisit*).

Learning effect – search terms used and dates searched

PsycLIT (January 1996–June 1998); Sociofile (1974–June 1998); ECONLIT (1969–August 1998): learn* near 2 effect*.
ABI/INFORM (January 1996–September 1997): learning effect?

Slips and mistakes – search terms used and dates searched

PsycLIT (January 1988–December 1998): (slip* or mistake*) and error*.
INSPEC (January–September 1995): (slip? or mistake?) and error?

Other search terms tested

Learning rate; operator/surgeon experience; calibration and (skill* or learn*); learning methods; maturity model.

Learning rate

MEDLINE (1995–October 1998): learning rate$.tw.
PsycLIT (1996–June1998): learning rate* (all fields).
Sociofile (1974–June 1998); EconLit (1969–August 1998): learn near2 rate*.
ABI/INFORM (1996–September 1997): learning rate?

Operator/surgeon experience

MEDLINE (1993–October 1998): ((operator$ or surgeon$) adj4 experienc$).tw.
INSPEC (January–March 1997): operator* near experience*.

Calibrate and (skill or learn)

MEDLINE (1995–May 1998); CINAHL (1982–March 1998); HealthSTAR (1996–May 1998): calibrat$ and (skill$ or learn$).
ISI Science Citation Index (January–July 1998); ISI Social Science Citation Index (January–July 1998); ISI Arts and Humanities Citation Index (January–July 1998); ISTP (January–July 1998); IBSS (January–July 1998); Ei Compendex Plus/Page One (January–July 1998); BIOSIS (January– July 1998); RSC (January–July 1998): calibrat* + (skill*, learn*).

Learning methods

PyscLIT (1996–June 1998); Sociofile (1974–June 1998); ECONLIT (1969–August 1998): learn* near2 method*
ABI/INFORM (1996–September 1997): learning method?

Maturity model

PyscLIT (1996–June 1998); Sociofile (1974–June 1998); ECONLIT (1969–August 1998): maturity model*.
ABI/INFORM (1996–September 1997): maturity model?

Rejected terms from preliminary work

Learning performance; development of competence; Weibull; ergonomics; power law (of practice); power function of practice; fan effect.

Learning performance

BIOSIS (1994–July 1997); PsycLIT (1991–June 1997); International ERIC (1976–March 1997); ERIC (1982–June 1994); Sociofile (1974–April 1997): learning perform*.

Development of competence

MEDLINE (1993–June 1997); CINAHL (1982–May 1997): develop$ adj2 competen$.
Science Citation Index (1981–July 1997); Social Science Citation Index (1981–July 1997);

Arts & Humanities Citation Index (1981–July 1997); ISTP (1982–July 1997); Ei Compendex Plus (1994–July 1997): develop* competen*.

Weibull
Science Citation Index (1994–July 1997); Social Science Citation Index (1981–July 1997); Arts & Humanities Citation Index (1981–July 1997); ISTP (1982–July 1997); Ei Compendex Plus (1994–July 1997): weibull.

Ergonomics
ASSIA PLUS (Spring 1997); Sociofile (1974–April 1997): ergonom*.
PsycLIT (1991–June 1997): human factors engineering (in DE).

Power law
MEDLINE (1993–June1997); Science Citation Index (1994–July 1997); Social Science Citation Index (1981–July 1997); Arts & Humanities Citation Index (1981–July 1997); ISTP (1982–July 1997); Ei Compendex Plus (1994–July 1997): power law.

Power law of practice
Science Citation Index (1981–July 1997); ISTP (1982–July 1997); Ei Compendex Plus (1994–July 1997): power law practice.

Power function of practice
Science Citation Index (1981–July 1997); Social Science Citation Index (1981–July 1997); Arts & Humanities Citation Index (1981–July 1997); ISTP (1982–July 1997); Ei Compendex Plus (1994–July 1997): power function practice.

Fan effect
Science Citation Index (1981–July 1997); Social Science Citation Index (1981–July 1997); Arts & Humanities Citation Index (1981–July 1997); ISTP (1982–July 1997); Ei Compendex Plus (1994–July 1997): fan effect.

Appendix 3

Details of electronic databases searched in phase 2 of the project

Non-clinical databases

- Social Science Citation Index (Institute for Scientific Information, Philadelphia, PA, USA) via BIDS online.
- Ei Compendex Plus (Computerized Engineering Index, electronic version of *Engineering Index*: Engineering Information Inc. (EI), Hoboken, NJ, USA) on BIDS online.
- Ei Page One (Engineering Information Inc. (EI), Hoboken, NJ, USA) on BIDS online.
- IBSS (International Bibliography of the Social Sciences: British Library of Political and Economic Science, London School of Economics and Political Science, London, UK) on BIDS online.
- RSC bibliographic databases (five of the available databases were searched. Chemical Engineering and Biotechnology Abstracts (CEABA): Royal Society of Chemistry; DECHEMA 1997, Analytical Abstracts (AA), Chemical Business NewsBase, Chemical Safety NewsBase, Mass Spectrometry Bulletin: Royal Society of Chemistry (RSC)) on BIDS online.
- PsycLIT (from PsycINFO database, American Psychological Association, Washington DC, USA) from SilverPlatter Information Ltd on CD–ROM.
- SOCIOFILE (Cambridge Scientific Abstracts, Bethesda, MD, USA, under license from Sociological Abstracts; SOPODA (Social Planning/Policy and Development Abstracts); also includes enhanced dissertation citations from Dissertation Abstracts International) from SilverPlatter Information Ltd on CD–ROM.
- ABI/INFORM (Bell & Howell Information and Learning, Ann Arbor, MI, USA).
- ECONLIT (American Economics Association, Nashville, TN, USA) from SilverPlatter Information Ltd on CD–ROM.
- Dissertation Abstracts (Proquest by UMI, Bell & Howell Information and Learning, Ann Arbor, MI, USA) on CD–ROM.
- Index to Theses, Great Britain and Ireland (Expert Information Ltd, London, UK, on KawareF Retrieval System) on CD–ROM.
- CAB Abstracts (CAB International, Wallingford, UK).
- INSPEC (Institution of Electrical Engineers, London, UK, covering physics, electronics, computing) on CD–ROM).
- NASA Technical Reports Server (http://techreports.larc.nasa.gov/cgi-bin/NTRS/).
- ISI Arts & Humanities Citation Index (Institute for Scientific Information, Philadelphia, PA, USA).
- ISTP (Index to Scientific and Technical Proceedings: Institute for Scientific Information, Philadelphia, PA, USA).

Clinical databases

- MEDLINE (National Library of Medicine, USA; electronic version of *Index Medicus*) from Ovid Technologies Inc. on CD PLUS.
- EMBASE (Elsevier Science Publishers BV, Amsterdam, The Netherlands; electronic version of *Excerpta Medica*) from Ovid Technologies Inc. via BIDS online.
- Science Citation Index (Institute for Scientific Information. Philadelphia, PA, USA) via BIDS online.
- BIOSIS (Biological Abstracts Inc. USA; electronic version of *Biological Abstracts*) on Edina.
- CINAHL (Cumulative Index of Nursing and Allied Health Literature: CINAHL Information Systems, Glendale, CA, USA) from Ovid Technologies Inc., on CD PLUS.
- HealthSTAR (National Library of Medicine, Bethesda, MD, USA, and the American Hospital Association) from Ovid Technologies Inc. on CD PLUS.

Methodology Group

Members

Methodology Programme Director
Professor Richard Lilford
Director of Research and Development
NHS Executive – West Midlands, Birmingham

Chair
Professor Martin Buxton
Director, Health Economics Research Group
Brunel University, Uxbridge

Professor Douglas Altman
Professor of Statistics in Medicine
University of Oxford

Dr David Armstrong
Reader in Sociology as Applied to Medicine
King's College, London

Professor Nicholas Black
Professor of Health Services Research
London School of Hygiene & Tropical Medicine

Professor Ann Bowling
Professor of Health Services Research
University College London Medical School

Professor David Chadwick
Professor of Neurology
The Walton Centre for Neurology & Neurosurgery
Liverpool

Dr Mike Clarke
Associate Director (Research)
UK Cochrane Centre, Oxford

Professor Paul Dieppe
Director, MRC Health Services Research Centre
University of Bristol

Professor Michael Drummond
Director, Centre for Health Economics
University of York

Dr Vikki Entwistle
Senior Research Fellow, Health Services Research Unit
University of Aberdeen

Professor Ewan B Ferlie
Professor of Public Services Management
Imperial College, London

Professor Ray Fitzpatrick
Professor of Public Health & Primary Care
University of Oxford

Dr Naomi Fulop
Deputy Director, Service Delivery & Organisation Programme
London School of Hygiene & Tropical Medicine

Mrs Jenny Griffin
Head, Policy Research Programme
Department of Health
London

Professor Jeremy Grimshaw
Programme Director
Health Services Research Unit
University of Aberdeen

Professor Stephen Harrison
Professor of Social Policy
University of Manchester

Mr John Henderson
Economic Advisor
Department of Health, London

Professor Theresa Marteau
Director, Psychology & Genetics Research Group
Guy's, King's & St Thomas's School of Medicine, London

Dr Henry McQuay
Clinical Reader in Pain Relief
University of Oxford

Dr Nick Payne
Consultant Senior Lecturer in Public Health Medicine
ScHARR
University of Sheffield

Professor Joy Townsend
Director, Centre for Research in Primary & Community Care
University of Hertfordshire

Professor Kent Woods
Director, NHS HTA Programme, & Professor of Therapeutics
University of Leicester

continued

HTA Commissioning Board

Members

Programme Director
Professor Kent Woods
Director, NHS HTA Programme, &
Professor of Therapeutics
University of Leicester

Chair
Professor Shah Ebrahim
Professor of Epidemiology of Ageing
University of Bristol

Deputy Chair
Professor Jon Nicholl
Director, Medical Care Research Unit
University of Sheffield

Professor Douglas Altman
Director, ICRF Medical Statistics Group
University of Oxford

Professor John Bond
Director, Centre for Health Services Research
University of Newcastle-upon-Tyne

Ms Christine Clark
Freelance Medical Writer
Bury, Lancs

Professor Martin Eccles
Professor of Clinical Effectiveness
University of Newcastle-upon-Tyne

Dr Andrew Farmer
General Practitioner &
NHS R&D Clinical Scientist
Institute of Health Sciences
University of Oxford

Professor Adrian Grant
Director, Health Services Research Unit
University of Aberdeen

Dr Alastair Gray
Director, Health Economics Research Centre
Institute of Health Sciences
University of Oxford

Professor Mark Haggard
Director, MRC Institute of Hearing Research
University of Nottingham

Professor Jenny Hewison
Senior Lecturer
School of Psychology
University of Leeds

Professor Alison Kitson
Director, Royal College of Nursing Institute, London

Dr Donna Lamping
Head, Health Services Research Unit
London School of Hygiene & Tropical Medicine

Professor David Neal
Professor of Surgery
University of Newcastle-upon-Tyne

Professor Gillian Parker
Nuffield Professor of Community Care
University of Leicester

Dr Tim Peters
Reader in Medical Statistics
University of Bristol

Professor Martin Severs
Professor in Elderly Health Care
University of Portsmouth

Dr Sarah Stewart-Brown
Director, Health Services Research Unit
University of Oxford

Professor Ala Szczepura
Director, Centre for Health Services Studies
University of Warwick

Dr Gillian Vivian
Consultant in Nuclear Medicine & Radiology
Royal Cornwall Hospitals Trust
Truro

Professor Graham Watt
Department of General Practice
University of Glasgow

Dr Jeremy Wyatt
Senior Fellow
Health Knowledge Management Centre
University College London

Current and past membership details of all HTA 'committees' are available from the HTA website (see inside front cover for details)

Feedback

The HTA programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278