The Reliability of Statistical Investigations into Surgical Audit Data

John Pollock

Ph.D – The University of Edinburgh – 2002

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my gratitude to Professor Gordon Murray for sharing with me his extensive experience of matters concerning the clinical and statistical aspects of surgical audit and for his considerable patience. I would also like to thank those who made the data available for analysis; The Glasgow Royal Infirmary, The Clinical Resource and Audit Group and the Bristol Royal Infirmary Inquiry.

I owe a particular debt to my wife Leslie for never failing to encourage me to continue with my research during what were some difficult times for us both.

Declaration

Aspects of the work undertaken in Chapter 5 have been published in the Lancet (Aylin *et al* 2001) and by the Bristol Royal Infirmary Inquiry (Murray *et al* 2001(a)).

I declare that this thesis has been entirely composed by myself. I collaborated with other statisticians involved with the Bristol Royal Infirmary Inquiry on aspects of the work involved in Chapter 5.

None of the material contained in this thesis has been submitted for any other degree or professional qualification.

Signed

John Pollock

Abstract

This thesis reviews the established methodology used in the examination of outcomes following surgery of various types. In particular we report on the statistical techniques used in audit studies to assess the relative performance of surgeons and the institutions within which they work.

An analysis is made of a long term sequence of outcomes for patients undergoing surgery for colorectal cancer in a single hospital and we explore the reliability of the inferences on relative performance which can be made. The conclusions drawn from this study are extended by an analysis of outcomes for patients undergoing similar surgery but in several hospitals. We compare traditional statistical approaches with newer computationally intensive techniques which more accurately model the variation between surgeons and institutions. These empirical studies are then used as the basis for an extensive simulation exercise which explores the adequacy of existing statistical methods to detect differences in surgical performance when we know *a priori* that they do or do not exist.

As part of the Bristol Inquiry into paediatric cardiac surgery at that institution we reviewed qualitative aspects of the data which were available for analysis and specifically analysed one set of data on outcomes based on the UK Cardiac Surgical Register.

We conclude with a comment on the realistic aims of statistical analysis of surgical audit data and discuss the implications this has for data analysts, clinicians and policymakers.

Chapter 1 – Review of Literature and Policy Background

1.1 Introduction

There has always been interest within the medical profession on the analysis of data on clinical outcomes. A commonly cited early reference on this issue are the practices used by Florence Nightingale in the Crimean War. She undertook a comparative study between active Army hospitals and UK statistics. She advocated recording outcomes with a view to improving overall performance and efficiency. Given what then happened in clinical practice in the following hundred years, she was clearly ahead of her time. Spiegelhalter (1999) reviews Nightingale's contribution together with the differing contribution of the nineteenth century US surgeon Ernest Codman.

It is now widely held that such analyses can assist in the development of good practice, the identification of sub-standard performance by institutions or individuals and in the efficient allocation of resources within hospitals and regional health authorities. This investigation of past and ongoing performance is often referred to as being part of the audit process. More than this, however, there is a much clearer public demand for information on institutional performance (not just in the medical field). There has been a response from Government intent on fulfilling these demands. This thesis is mainly concerned with the statistical issues involved in the audit process but the analysis of outcomes can only be one feature of what has been described as the 'audit loop'. Where inferences from the analysis of outcome data are uncertain we have a greater requirement to consider 'process' as comprehensively as we do 'outcome' (Crombie, 1993).

As discussed, the process of audit is however not simply about the analysis of outcomes. The standard textbook discussions centre around three elements. Structure, Process and Outcome. Crombie *et al* (1993), Delamothe (1994) and Smith (1992) give a ranging discussion of the topic and Pollock (1993) focuses on the issue from a purely surgical perspective. Structure is the basic framework of resources available to the institution of interest. We will discuss structure very little in this thesis but only comment that the results of audit studies on process and outcome can have a significant lagged effect on structure itself. An example of this would be the way in which audit, in its many forms, has highlighted the need for specialised surgery to be carried out in a smaller number of 'centres of excellence' (Clasby *et al*, 1997). Again considering process and the implications for structure, audit studies have also focused attention on training issues for junior surgeons, as discussed by Aitken *et al* (1999) and McCarthy & Byrne (1997).

Process is a more intangible element of the audit exercise. In some ways it is easy to observe being an examination of the way patients are treated in a comparative fashion but it does have inevitable subjective elements in the analysis.

The final element of audit is the analysis of outcome data. This might include a specific study of mortality or complication rates with a clearly defined definition of negative outcome such as death within 30 days or inpatient mortality. It might however be a less clearly defined outcome such as patient satisfaction which could be assigned a numerical scale (and then be analysed using statistical methods) but is more subjective. It will be seen in this thesis that analysis of surgical outcomes alone can help identify extremes of performance but in general we need to focus on both process and outcome.

In summary, audit is important for many reasons. It provides a framework for monitoring performance and feeds back information on needs and resources to funding authorities. Nixon (1992) and Mannion & Goddard (2001) discuss the impact of audit on practice and resource allocation.

Audit also enables the public to assess the relative performance and efficiency of different institutions. Regrettably this final point is not as straightforward as it may seem and the later chapters in this thesis highlight the potent dangers of taking a narrow focus on outcome alone.

1.2 The scope of this review of literature

This initial chapter reviews the context within which surgical audit work takes place in the National Health Service in the United Kingdom. It is comprised of two main sections. Firstly, we will briefly review both the background to this issue and the current policy of Government and the NHS in England and Scotland towards a range of audit, reporting and performance measurement issues. We will discuss the current high profile of public interest in surgical outcomes and health related performance statistics in general. Secondly, we will review the statistical approaches which have historically been taken towards the solution of problems concerning the determination of relative performance within and between institutions and report on how developments in this area have progressed rapidly in recent years as new techniques based on computationally intensive procedures have become more widely developed and applied. Later chapters in this thesis will be concerned with the specific reliability or otherwise of surgical audits based on the examination of outcomes alone. We will report on the analysis of three separate sets of actual data.

1. A long term series of data on outcomes following colorectal cancer surgery in a single institution.
2. A short term series of data on outcomes following colorectal cancer surgery in a number of institutions..
3. A long term series of data on paediatric cardiac surgery reviewed as part of the Bristol Royal Infirmary Inquiry.

Based on the various issues raised in these particular data analytic exercises we also draw more general conclusions regarding the realistic aims of surgical audit outcome analysis exercises through a comprehensive simulation exercise exploring sample size, effect size and modelling considerations.

1.3 The process of literature review

Our review of literature and policy background covers the main sources of published material which relate to surgical audit issues and health related performance measurement in more general terms. It is however not exhaustive since we make relatively brief comments on practical developments and published material outside the United Kingdom.

This was a conscious decision since the policy background within the United Kingdom differs substantially from overseas countries, and in particular from the United States of America. Where overseas publications are referred to they are more generally papers or texts which consider statistical and modelling issues. Our review of literature has also been prepared with the broader objectives of our research in mind, as discussed in the previous section.

The main methods of obtaining references were as follows :-

1. Initially we considered general historical reference works on all aspects of clinical and surgical audit. These not only assisted in the provision of background information and knowledge but were also a useful source of other referenced material. Some published journal papers also provide appropriate historical references. A previous thesis in this area also assisted in this regard.

2. The statistical references were also taken up at this preliminary stage in our research. Several standard texts were available which document the main basic methods used to compare performance and allow for the variables which influence performance. More recently published texts on hierarchical statistical models also provided substantial amounts of reference material on more complex statistical techniques. Explicit referencing in this thesis of all technical material reviewed has not been made since it would not add specific value in relation to the narrower issues and problems which are considered herein but this preliminary accumulation of technical background was valuable.

3. The above steps enabled us to be relatively confident that when our research began we had read sufficiently widely that we were fully informed of the main historical and statistical issues of consequence. From then as we progressed with our research we maintained this level of currency of knowledge by reviewing papers in the more important journals as they emerged. These include the British Medical Journal, The British Journal of Surgery, the Lancet and Statistics in Medicine. These main journal references then provided occasional direction to less prominent journals for additional material of relevance to our research.

4. On line references were also taken up on a regular basis and indeed much of the material which forms the basis of our discussion below on current Government policy and published NHS performance indicators for both England & Wales and Scotland are available from the internet. The vast resource of information and background from the Bristol Royal Infirmary is also fully published on the World-Wide-Web. Various on line statistical resources were also utilized. The Mailbase listings service offers a topical overview of current modelling problems and links to University maintained journal pre-print services. The listings 'BUGS' and 'Multilevel' were of particular help in accumulating ongoing knowledge and reference material.

1.4 Early practical references for surgical audit

Audit data might be collected prospectively, with the ultimate statistical analysis in mind, or the data might be collected retrospectively from records kept for some other purpose. Historically the integrity of data has been a major concern and in many cases coping with unrecorded or inconsistent items of information is a major barrier to the statistician attempting to make sound inferences from their analytical efforts. The increased attention being paid within the medical profession to audit related exercises has however led to the formation of specific audit groups at both local and regional levels and there can be little doubt that the quality of data available to investigators is steadily improving.

The recent high profile within the wider media of specific cases of surgical under-performance seems guaranteed to keep audit related issues at the front of many clinician's and administrator's minds over the coming years.

As discussed earlier the earliest references on the issue tend to be the work and thoughts of Nightingale and Codman. An earlier more contemporary reference was the plea for the introduction of structured surgical audit from Dudley (1974). A reference on institutional differences in death rates outside the United Kingdom is Moses & Mosteller (1968). Another early reference on audit was prepared by Gough *et al* (1980). They surveyed the work of a particular surgical 'firm' (a unit of practitioners working together consistently within a single hospital) and reported a full discussion of their conclusions as they related to outcome measures. The authors were admirably frank about specific episodes where they felt performance (in terms of either mortality or morbidity) was sub standard. They advocated prospective audit and highlighted the demands that audit, if performed properly, places on NHS resources. Gilmore *et al* (1980) were concerned more with case mix and volumes but statistics were used to compare two institutions.

Dunn & Fowler (1992) discussed the Confidential Comparative Audit Service of the Royal College of Surgeons of England. Emberton *et al* (1991) also describes the background to the Comparative Audit. The Dunn & Fowler paper reviewed the analysis of confidential returns submitted by general surgeons. A combined audit involving computing resources across several regions was also discussed by Dunn (1992). A total number of 147,882 admissions were considered. The authors strongly advocated the practice of comparative audit and the contributors were enthusiastic about continuing to contribute to the ongoing audit.

Dunn (1986 and 1988) had previously reviewed audit in various contexts exploring issues as varied as computerisation and construction of forms. The advantage of this audit was the scale of the numbers involved. The disadvantages were the voluntary nature of submission and the difficulties of allowing for complex mixes of patients. Importantly data recording issues were a matter of significant concern. Some basic comparative mortality analyses were performed and surgeons could identify from the output how they were performing and could also examine their mix of patients as an informal way of adjusting for obvious case mix effects. There were statistics for average performance but no explicit calculation of measures of variability.

Matthews *et al* (1986) reported on a study of outcomes following surgery for oesophagael cancer and concluded that such specialised surgery should be concentrated in areas where there is significant experience and activity. The practical and computing considerations involved are discussed by Ellis (1987 and 1989). Ruckley (1984) discussed the problems raised by the long delays that can take place before data is analysed and reported to practitioners involved in audit exercises. The recent advance in uses of cumulative sum techniques, see below, are an attempt to address this issue. Deans *et al* (1987) made an early comment on the need to allow for case mix in making institutional comparisons.

Murray *et al* (1995) and Hayes (1995) extended the work of the Confidential Comparative Audit Service of the Royal College of Surgeons of England. Explicit statistical modelling of the case mix was attempted since, despite the large numbers of patients involved, there was a clear prior reservation about making inferences based on crude rates alone. A substantial change in relative performance occurred when case mix was allowed for relative to crude unadjusted comparisons. The data were presented in a more informative way using confidence intervals for measures of relative mortality and using triangle diagrams to graphically illustrate case mix variability.

One problem identified by the authors was the fact that the analyses were performed on 'aggregate level' data as opposed to being performed on 'patient level' data. The latter is of course to be preferred (since it enables more efficient allowance for case mix and more accurate quantification of statistical variability) but is not always available. The recent work undertaken by statistical experts for the Bristol Royal Infirmary Inquiry also had to address such issues. (The Bristol Royal Infirmary 2001 (a), (b), Murray *et al, 1999* and Spiegelhalter,1999).

The more general problems which arise in analysing aggregate data are also well researched in a more technical sense by statisticians involved in hierarchical modelling. One can contrast inferences made on aggregate data for higher levels in a multilevel model with the (sometimes quite different) inferences that can be made where full information is available at the lowest level of the hierarchical model (Goldstein, 1995).

Porter *et al* (1998) looked specifically at surgeon-related factors in the context of patients with colorectal cancer. Their conclusions based on a large study were that outcome (measured both in terms of recurrence and death) is improved if surgery is concentrated in institutions where clinicians have explicit training and experience with colorectal cancer. The clinical and pathological factors influencing survival for patients with colorectal cancer are covered in Deans *et al* (1994) and Spence (1994).

Pollock (1993) considered the value of audit data in appraising types of surgery and compared the problem to that faced by those working in clinical trials. The ethical problems are arguably more prominent in the field of surgery as opposed to, for example, drug testing. Neugebauer *et al* (1991) also considered the problems of randomised controlled trials in a surgical context.

In the 1980's the first National Confidential Enquiry into Perioperative Deaths involved an individual level analysis where deaths were classified into 'avoidable' or 'unavoidable' (Buck *et al,* 1987) and there have been regular developments in this area most recently The Report of the National Confidential Enquiry into Perioperative Deaths (1999). One of the past recommendations of this series of studies was for a reduction in night time emergency surgery and the involvement of senior staff where possible. There are however conflicting opinions on aspects of the recommendations (Cook *et al*, 1997).

There has been ongoing interest in the factors influencing survival following surgery. Houghton (1994) reviewed the relationship between volume of surgery and outcome at the hospital and surgeon level and highlighted clinical concerns about whether the relationships which are observed, specifically that higher volumes tend to be associated with better outcomes, are in fact causal or related to other factors (including referral practices and population profiles).

In this thesis several chapters are concerned with audit problems where the underlying population are patients undergoing surgery for colorectal cancer. As a result we note some more specific references below. In many respects audit has always had a high profile in this particular surgical discipline. Early studies focused on the clinical and pathological variables that influenced survival (Dukes, 1932 &1940 and Gill, 1978). Others noted the concern over the high percentage of patients presenting with advanced disease (Clarke *et al*, 1980). Importantly, from an audit perspective, Fielding *et al* (1980) used the results of a study of complications to highlight surgeon related variability in outcomes (expressed in terms of both morbidity and mortality).

The Large Bowel Cancer Project (Phillips *et al,* 1984), in a prospective study, reviewed a large population of patients undergoing 'curative' surgery for colorectal cancer. The factors influencing survival were identified and a study was made of factors influencing recurrence of disease. Considerable variation in recurrence was noted between surgeons. The Lothian and Borders large bowel cancer project (The Consultant Surgeons and Pathologists of the Lothian and Borders Health Boards, 1995) noted the large variation in mortality between surgeons and the improved morbidity of patients treated by more experienced and active surgeons.

Chapuis *et al* (1985) used more complex statistical methods (the Cox proportional hazards model) to model the effects of variables influencing survival for patients with colorectal cancer. The expected strong influences of pathological variables were identified.

McArdle & Hole *et al* (1990) identified the advanced stage of disease at presentation in colorectal cancer and the same authors investigated surgeon related variability in outcomes using a methodology and data which we will review and extend in this thesis (McArdle & Hole, 1991). In later related studies, using a larger combined set of data, it was noted that despite improvements in surgical technique and perioperative care outcomes remained poor (McArdle *et al,* 1996)

Of specific more recent interest from the surgical audit perspective was the review by Holm *et al* (1997) of surgeon related variability in outcomes following treatment for rectal surgery. The extent of disease present in patients at diagnosis and surgery remains advanced as evidenced by population based audits in several regions but there are some more encouraging trends in various aspects of treatment (Mella *et al,* 1997). A study by Singh *et al* (1997) identified no differences between surgeons in a small audit involving just 267 patients and 4 surgeons.

An interesting recent high profile comment on surgical performance was made by Professor Michael Baum, a breast cancer specialist. Professor Baum published in the Lancet what in previous decades would be seen as a quite extraordinary exercise of self criticism listing where he felt he had erred seriously in the past. This was, naturally, given more general press coverage as well (Peterkin, 2001).

This raises a further implication of the audit process, litigation. Attempts are being made to place claimants and defendants in a less adversarial situation (The Bristol Royal Infirmary Inquiry, 2001 (b)) but it is doubtful whether this will happen given the seemingly endless upwards spiral in the numbers and value of claims in all aspects of modern life. Developments are clearly led in this context by the US.

The NHS Executive estimate that several billion pounds of claims are in progress. Statistical evidence is likely to be increasingly called upon in cases of substantial public interest or financial value. It is of importance that the uncertainty of estimates of performance (and in particular the uncertainty attaching to rankings in 'league tables') can be communicated effectively to the Courts and that these are taken into fair account by the Judiciary.

1.5 Policy background – The Bristol Royal Infirmary Inquiry

Two things in particular have led to this explosion of interest in comparative statistics. They are inter-related. Firstly, as part of government policy on openness, such information has been made more freely available. This has also happened at a time when several major events have increased public focus on the quality of care. The Inquiry into the quality of paediatric surgery at Bristol Royal Infirmary (The Bristol Royal Infirmary Inquiry, 2001 (a)) is the most notable but by no means the only public event that has led to public interest in comparative medical statistics.

The background to the Inquiry and the analyses undertaken (partly reported in this thesis) are to be found on the Inquiry Web Site. When the final report was published in July this year the whole scale of the project became known more clearly than before.

The Guardian (19 July 2001) published a table highlighting some useful statistics which although rather trite bear repeating.

- The Inquiry cost £14 million pounds
- There were 96 days of oral evidence heard
- There were 577 written statements
- There were 673,963 pages of medical records examined.
- There were 42,071 documents submitted
- The final report was 12,000 pages long

The implications of the Bristol Inquiry will start to be felt in the coming years. The Government took the opportunity to announce on the same day the Inquiry Report was published that new policy initiatives were to be effected.

The Report itself recommends the establishment of the Council for the Quality of Healthcare and the Council for the Regulation of Health Care Professionals. These should establish an integrated method for the establishment of standards (structure and process) and for the monitoring of performance (outcomes). It recommends clearer publication of results from surgical units both within and between institutions.

It suggests that there should be a further restriction in the number of centres where cardiac surgery on the very young should be carried out. It suggests minimum numbers of operations per surgeon per annum to establish an 'experience' qualification and advises that at least two such surgeons be available in any one institution. This will formalise the trends towards specialised centres of surgery which have been gradually developing in any event.

An important part of the work done for the Inquiry was purely concerned with data quality issues and these are more fully reported in Lawrence & Murray (2000). Poor data will lead to not only inaccurate point estimates of relative performance but, almost as importantly, inaccurate and biased estimates of statistical variability. As will be discussed the new NHS Data Quality Indicators are an attempt to introduce some stability in coding procedures.

We will periodically return to discussion of data quality and coding issues throughout this thesis as we analyse actual sets of surgical audit data (including data from the Bristol Inquiry). There are particular problems with retrospective audits where the data may have been recorded for a quite different purpose. With prospectively organised audits becoming the standard method of data acquisition data quality will inevitably improve and resulting concerns about bias in results lessened.

The costs of increased Government initiated audit and performance measurement based activity have yet to be fully researched, but will inevitably be substantial.

1.6 Policy background – Scotland

In Scotland, in many respects, the collection of data on process and outcomes and the publication of comparative statistics has often led other parts of the UK. The Lothian Surgical Audit, started in 1979 collected data manually at first and then as computerisation became available widened coverage to include a larger number of outcome measures including mortality and complications. It has highlighted issues on data collection and data quality and has strongly emphasised the communication aspects of statistical work on comparative audit problems (Gruer, 1986). The NHS in Scotland now has a very extensive commitment to both the publication of statistics and to the audit process in general (Carter *et al* , 1995)

The Information and Statistics Division for the NHS in Scotland facilitates many audit exercises. In 1996 a report on commissioning Cancer Services in Scotland suggested that there was a clear requirement for the establishment of a monitoring system for cancer services and that (prospective) audit was a key feature of their proposals. Local data should they said be readily comparable with national statistics, where available (Clinical Resource and Audit Group, 2000)

The Scottish Cancer Therapy Network deals, amongst other matters, with audit of four major cancers (lung, breast, colorectal and ovarian cancer). One key point is that a central agency provides guidance in data definitions, coding procedures and collection mechanisms.

The central agency supports contributing Trusts with provision of computer software (commonly set up within Microsoft Access) together with assistance on data interpretation and definitions. It also directly assists some trusts with management services. The ISD works closely with the Clinical Standards Board for Scotland and the Clinical Resource and Audit Group (Clinical Standards Board for Scotland, 2001).

Scotland has a very comprehensive source of information on clinical data that continues to be actively developed. The ISD manages a number of clinical databases the most important being national systems based around hospital discharge data (the Scottish Morbidity Records). They work to establish linked data resources. The Scottish Morbidity Records (SMR, not to be confused with the Standardised Mortality Ratio) have been recorded since 1961 and are one of the oldest complete national sources of clinical data in the world. Whilst originally developed for largely administrative reasons the SMR are now used more widely. Of particular interest is their use in the development of performance indicators. Data are collected on standardised forms and input onto the Patient Administration System before being submitted to the ISD (Clinical Indicators Support Team, 2001).

The key benefit of the data linkage facility is that individual patients can be traced through multiple hospital episodes. This facilitates audits of, for example, recurrence and readmission. Crucially linkage is extended to registers of deaths and cancer registries. All these facilities both extend the scope of potential statistical analyses and improve the quality of the data which are being used.

The Clinical Resource and Audit Group (CRAG) has provided data for analysis in this thesis and we are grateful to them for giving us this opportunity. The objective of CRAG is to promote clinical effectiveness through a variety of audits of process and outcome. It funds the very recently established Clinical Indicators Support Team at the ISD. The 'flagship' publication of the CRAG is the extensive report on Clinical Outcome Indicators most recently published in December 2000 (Clinical Resource and Audit Group, 2000)

This eighth document in an annual series enables wide public access to a range of data on clinical outcomes and makes use of the extensive sources of data available in Scotland. The 2000 edition does importantly qualify the performance information with a comment that case mix is not properly allowed for or described in the report.

Quite properly CRAG do not allow the reader to infer conclusions from the report alone but do acknowledge that the more striking observations might merit further study. The 2000 report covers a comprehensive range of indicators including those concerning cancer services and emergency admissions.

An impressive feature of the data collection aspect of the work undertaken by CRAG is the sample checking of the SMR records. This enables the researcher (and reader) to be confident about the accuracy of the recorded information and, as importantly, on the consistency of recording across institutions. It appears that the information is also very up to date with SMR data generally submitted to the ISD within three months.

As an example, and in particular since this particular very important clinical issue is considered in detail later in this thesis, we will briefly outline the range of the discussion of and analysis of statistics on colorectal cancer in the latest edition of Clinical Outcome Indicators. It covers the following areas.

- General statistics on incidence and mortality rates.
- International comparisons of incidence and mortality rates
- Trends in mortality rates over time
- Variations in mortality rates between health boards (with confidence intervals)
- Comment on therapy and treatment.
- Comment on explanatory variables including social depravation.

Clearly very comprehensive information is therefore now being made publicly available. The more recent formation of the Clinical Standards Board for Scotland (in April 1999) is a further move towards openness in the NHS (Clinical Standards Board for Scotland, 2001).

Attempts are being made to determine standards for processes in many areas and the membership of the Board is more widely sourced than might have occurred in previous decades.

The most recent development in Scotland at Government level has been the formation of the Clinical Indicators Support Team (2000). This group monitors and maintains the extensive list of Outcome Indicators and is working on independent validation of the indicators used through linkage to the other sets of data. They also look at exactly how Boards and Trusts are using the indicators that are published, an important issue not always at the front of peoples' minds.

1.7 Policy Background - England

The National Institute for Clinical Excellence (incorporating the National Centre for Clinical Audit) now leads and supports many initiatives in the field of clinical audit (National Institute of Clinical Excellence, 2001). Their definition of audit bears repeating. It is 'the monitoring of interventions or care received by patients against agreed standards'. They provide guidance on what best practice might be defined as being and advise on basic methods which enable health care workers to monitor their own performance and adjust their practice as required.

In their document *'A First Class Service' : Quality in the New NHS* ' the Government stressed the importance of monitoring process and assessing performance (Department of Health, 1998)

In England for the last two years we have seen publication of the NHS Performance Indicators (Department of Health, 2001 (b))

Like their Scottish counterparts, the Clinical Outcome Indicators, these are a vast publicly available resource of data on comparative performance. They are being updated to reflect the prevailing list of Government priorities (as detailed in the White Paper, *Saving Lives*). It is proposed clear national standards are established across the disciplines in the NHS through the work of the National Institute for Clinical Excellence. Standards are to be reviewed in three ways ; by the Commission for Health Improvement, through the NHS Performance Assessment Framework and through patient surveys. The NHS Performance Framework (NHS Executive, 1999) introduced a series of what were called High Level Performance Indicators and a series of Clinical and Data Quality Indicators.

As in Scotland there is emphasis on the fact that the indicators only draw attention to areas of concern or interest. They are not a definitive guide to relative performance. The work of this thesis emphasises the common sense of this statement.

The NHS Executive envisage a circular series of audit exercises, really an extension of the traditional audit "loop" as discussed by various authors (Dudley 1974, Crombie and Davies, 1993). The sequence followed is:

1. *Health Improvement Initiatives ->*
2. *Fair access to care ->*
3. *Effective Delivery of Care ->*
4. *Efficiency of Resources ->*
5. *User Experience assessment ->*
6. *Explanation of Outcomes ->*
7. *Health Improvement Initiatives (again)*

The latest edition of indicators (*Quality and Performance in the NHS : NHS performance indicators July 2000*) is a substantial document. The stated aim is to identify extremes of performance and to investigate these with a view to modifying future results.

The Department of Health web site lists the large number of Clinical Indicators which are monitored. These include information on waiting times, length of hospital stay, cancelled operations, rates of surgery, infant mortality, and survival rates from a number of cancers. Graphs are used to illustrate conclusions. A level of confidence attaching to the performance estimate is shown and a ranking by mean values is prepared. There are of course negative comments that can be made regarding the rank order issue. Even graphs with confidence intervals for the performance measure fail to identify the true uncertainty attaching to the rank order. An attempt is made to standardise data with respect to age and sex. No attempt is made to adjust for socio-economic factors.

The indicators are mainly based on the hospital episode statistics (HES) supplied to the Department of Health (Department of Health, 2001 (a)). The Bristol Royal Infirmary Inquiry has highlighted some concerns over data coverage and completeness with HES data (Murray *et al,* 1999 and Lawrence & Murray, 2000). A study of data quality led to the exclusion of certain institutions from certain aspects of the study. As discussed at the Inquiry clinicians in general have little confidence in the accuracy of HES data despite this being collected at the individual as opposed to the aggregate level. A list of Data Quality Indicators is produced in an attempt to place a quantitative measure on this partially qualitative component of the audit exercise.

The NHS Performance Indicators discuss data considerations and this is an area of considerable concern for the audit specialist. The NHS Centre of Coding and Classification is attempting to standardise practice. A substantial part of the statistical work undertaken for the Bristol Inquiry was purely related to data issues (which we reviewed in Chapter 5). It is envisaged that closer communication between clinicians and the coding staff will be required to ensure quality and consistency in the recording of audit data.

Complete data on hospital in-patient episodes based on HES are now available to the public through the Department of Health Web Site. Over 12 Million records with over 50 fields of data are recorded annually. It is administered by the Statistics Division in the Department of Health. Surgical data are coded by OPCS 4 codes and ICD10 codes.

In summary, as in Scotland, the pace at which performance statistics have been prepared and made available for public consumption has quickened significantly in recent years. Data quality concerns remain and there are certain reservations about the presentational aspects of some national reports but there seems no doubt that further resources will be devoted to this area in future.

1.8 The statistical background to audit problems – types of outcome

In audit exercises we are generally dealing with observational studies, as opposed to randomised clinical trials for use in assessing the effect of a particular treatment or intervention. This makes the analysis of data more troublesome since it becomes more difficult to interpret conclusions definitively in the presence of often confounding explanatory variables. As we will discuss below the particular hierarchical nature of many types of audit data further complicates matters.

Outcome data can take many forms and numerous examples can be drawn from branches of the health care system. The vast number of currently monitored outcomes was reported earlier when we discussed the NHS Performance Statistics. As an illustration however we highlight several examples below.

1. Length of stay in hospital for patients of various types.
2. Proportion of patients experiencing recurrence of symptoms after a particular drug treatment.
3. Numbers of deaths following cardiac surgery
4. Survival times following surgery for colorectal cancer.
5. Recurrence of symptoms or tumours following surgery of a particular type
6. Presence or absence of a wound infection following abdominal surgery.
7. Number of units of blood used in general surgery

Examples 1, 4 and 7 deal with a continuous outcome measure. Examples 2, 3, 5, and 6 deal with what are effectively binary outcomes. Most analyses of audit data fall into these two main categories. A key feature of some types of outcome data is that the variable being measured, say, survival time, may be censored.

For example for some reason we may not know the actual time of death but only that it was beyond a certain point in time. Equally we may simply have lost the individual to follow up. The mechanism of censoring might be random or it might be a feature of the study itself.

The type of outcome measure determines the types of statistical methods that can be used to prepare measures of absolute and relative performance and to place measures of uncertainty on these point estimates.

1.9 Factors influencing survival and problems to be addressed

For the statistician to make valid inferences about individual surgeons or the population of surgeons an attempt must be made to model the effect of variables that influence the outcome under consideration. It can be clearly demonstrated that certain factors predispose individuals to good or bad outcomes. These fall into three broad categories (at least when considering the surgical audit example):

1. Variables that are inherent qualities of the patient involved. These would include age, sex, social class, and location of residence.

2. Variables which reflect the nature of the disease under consideration. Taking colorectal cancer as an example this might be the Dukes' Stage (Dukes 1932, 1940) or the level of differentiation of the tumour which is exhibited. Broadly speaking, these covariates would reflect either the extent or aggressiveness of the disease.

3. Variables reflecting the treatment of the individual under consideration. Continuing the colorectal example this might include the surgeon and hospital involved, the seniority of the surgeon (and possibly any assistant), the procedure undertaken (e.g. a curative or palliative resection) and the admission category (e.g. emergency or elective surgery).

The questions that statisticians may be addressing are many and varied :-

- They will be interested in quantifying the performance of the surgeons (or hospitals) under consideration having properly adjusted for the different risk characteristics of the patients involved (or having stratified them into homogeneous groups with similar characteristics). Attempts may also be made to illustrate the confidence attaching to any point estimate of performance, such as a risk or hazard ratio.

- They may be interested in detecting whether individuals (or hospitals) are significantly different from their peer group or certain individuals within the group. This involves testing a particular hypothesis that a particular surgeon differed from the others involved in the study.

- They will wish to consider the rank order of the practitioners (or possible other functions of the primary outcome measure). Of particular interest will be the confidence attaching to rank orders.

- Also of interest is the power of the statistical method being used to detect differences in performance where they do exist and the determination of the probability of incorrect inferences being made about surgeons, institutions or the population as a whole. What is the chance (given a particular data set up and statistical method) that we are detecting a significant difference in surgical performance when in fact none exists? Given that we know a difference exists what are our chances of detecting this difference ?

- An examination may also be made of the sensitivity of performance measures to the statistical method being used and the set of explanatory variables being used to adjust for case mix.

- A further examination of the residuals after the model has been fitted will illustrate the adequacy of the fitted model, the influence of unusual, 'outlying', observations and may highlight anomalous data entries caused by input or recording errors.

- The fitted statistical model may be used for prognostic purposes and be used to assess the expected impact of the principal explanatory variables of interest.

- They may be interest in assessing whether the data recorded are accurate and that they have been gathered consistently across the institutions of interest. They will wish to explore whether data considerations alone might invalidate the statistical inferences being made about relative performance.

1.10 Statistical background

Statisticians must be aware that they are modelling an inherently variable and complex process. Any statistical model is just that, an attempt to model the features of some underlying process. More complex models may explain more of the underlying variation but might do so in a way that makes interpretation of the parameters of the fitted model difficult. The objective should be to have a relatively simple model where the underlying effects of various variables are measured with as much accuracy and as little bias as possible. More than this one would expect some variability between surgeons. When comparing one surgeon with 'the rest' we may be unwise to treat the peer group as being a homogeneous group.

The basic statistical techniques used to answer some of the above questions are well established and there are many standard texts on medical statistics, including Altman (1991) and Breslow & Day (1980). We comment on these methods further below but before proceeding to this a final, and most important, feature of the data should be mentioned. This is the fact that surgical audit data are often hierarchical in nature. The individual cases are not necessarily independent of each other. It would seem reasonable to assume from prior considerations that a team of surgeons operating within a particular hospital might share some characteristics in common. Equally patient characteristics may differ from another hospital in another part of the country. The individual patients can be seen as the first level in a hierarchical structure, the second level being surgeons and the third level being hospitals. The different levels in the structure are nested within each other.

New techniques of analysis are complex and have indeed only become feasible with the rapid increase in available computing power we have seen recently. This comment would particularly apply to more complex techniques such as Bayesian hierarchical modelling, see below.

It is, however, to be remembered that the interpretation and communication of the results of a study must be in a format that clinicians, and increasingly the general public, can understand. Recent media attention on league tables of death rates with insufficient emphasis on the uncertainty of such estimates should serve as a warning to researchers in this field. This would be particularly the case for studies publishing rank orders with no comment on the level of confidence attaching to the ranks. Even displays of performance measures with an associated confidence interval ordered sequentially (as used in the NHS performance indicators) can be misleading since these fail to notify the reader to the true uncertainty attaching to the rank order. Some functions of the outcome measures such as ranks are notoriously variable and in many cases it is very difficult to say with confidence that a particular surgeon or hospital is even within the bottom quartile let alone make a judgement that the surgeon or hospital is 'bottom of the league'.

Advances in computing power have led to a development in the application of Bayesian approaches to analysing complex sets of data. The WinBUGS software which can be used to analyse problems from a Bayesian perspective is discussed below and lends itself naturally, through the particular simulation based features of the estimation procedure, to inferences on rank orders. It is expected that this new methodology will become more widely used and reported in future than it is at present, particularly given the fact that the results are so easily interpreted by the general public (Marshall & Spiegelhalter, 1998 and Parry *et al,* 1998). Not only did these papers demonstrate the theoretical uncertainty of ranks but also demonstrated the observed lack of stability in rankings in successive years. This is an area we consider ourselves in Chapter 2 of this thesis.

It would be false to imply that the particular choice of model used is in a sense uniquely correct. Of course, some models are better than others, and in particular models accounting for the hierarchies in the data have less bias and better estimates of error than single level analyses trying to solve the same problems, but different statistical approaches are still possible even within the same model structure. We make no attempt to contrast the Bayesian approach to hierarchical modelling (as used in practice in this thesis) with the classical or frequentist approach used in other software packages, such as MLWin.

Basic regression techniques for the selection of optimal subsets of explanatory variables are well documented and essentially a statistical test is used to assess the significance of the alteration in fit when variables are successively added or removed. These various techniques do not necessarily give the same set of selected variables at the end of the exercise; some judgement is still required. Many studies have literally dozens of items of information, many of which can be regarded as measures of the severity or aggressiveness of the disease under consideration.

Chapter 2 of this thesis considers some of these issues in a basic individual level analysis of data on outcomes following surgery for colorectal cancer from Glasgow Royal Infirmary between 1974 and 1984. Many adequate but slightly different models can be developed which capture the major case mix effects and explain (via case mix adjustments) a similar amount of the total variation in outcomes (survival times in this example). Different models may give rise to different performance measures.

We are interested in the reliability of inferences made from audit data and would hope that the measures of performance that we use are reproducible in different circumstances. A formal definition of reproducibility of outcomes based on a different data or modelling setup is difficult to structure. We would certainly wish to see that under different circumstances that measures of performance did not differ to an extent that alterations in levels of statistical significance arose.

It would raise concerns if a surgeon was identified as being significantly different from his or her peers using one set of variables to account for case mix variation when no such divergent performance was identified using a different set of covariates chosen to achieve the same purpose. A partially qualitative judgement depending on the nature and size of the study could be made about an acceptable level of divergence in performance measures based on differences in risk measures.

Studies where the outcome measure is a censored survival time raise particular concerns as the researcher often has to make a choice as to when to stop following up patients. The longer patients are followed up the more information is obtained and the lower the percentage of cases censored by other mechanisms, both features increasing the power of the analysis.

However, this naturally causes a delay in the reporting of results. Investigations have been made (reported in Chapter 3) where the sensitivity of performance measures to the timing of the final censoring event is examined. In some cases, see Chapter 3, the decision to censor all cases at a fixed follow up duration is taken for other reasons.

Outcomes thought of as coming from a binomial distribution and related modelling techniques are well documented in texts on Logistic Regression (Hosmer & Lemeshow,1989). In the same way that one models the log odds ratio as a linear combination of explanatory variables to explain as much of the case mix variability as possible logistic regression models extend in a straightforward way to surgical audit problems.

In analysing survival time data the underlying process is modelled using methods similar to those outlined above for generalised linear models adapted for the censored features of the response data. There are two principal methods of analysis used, namely, parametric and semi-parametric methods. The semi-parametric Cox proportional hazards model is most commonly used in medical studies (Cox, 1972 and Cox & Oakes, 1984)

No matter what the form the outcome data take, analyses have elements in common. An attempt is made to determine the relationship between a response and a set of explanatory variables by modelling data collected on individuals. Any group effects (say a surgeon or hospital effect) are allowed for by the inclusion in the fitting process of numerous dummy binary variables for the surgeons of interest (e.g. 1 if a case is operated on by surgeon X and 0 otherwise).

The resulting parameter estimates give the estimated effects of treatment by the surgeon in question (relative to his peers or to a chosen baseline hospital depending on the method used). This approach can however be criticised since the method does not allow for explicit modelling of the dependence between certain groups or clusters of patients.

Where such dependence exists the residual terms are no longer independent of each other violating the underlying assumptions of the model fitting process. In particular the standard errors of estimates will be incorrect and the tendency will be for there to be an excessive amount of higher level units who are judged to be 'significantly different from average'. The standard errors of estimates are understated and the probability of Type 1 errors in hypothesis tests increases. The greater the degree of correlation within hierarchical data the worse will be the problem (Goldstein, 1995)

This is a crucial point when one considers the area of performance measures and league tables for hospitals or surgeons. Goldstein & Speigelhalter (1996) review issues related to performance measures and these ideas have been developed further by other researchers (Deely & Smith, 1998). In the surgical audit data analysis that follows in this thesis the rank orders of surgeons are of particular interest. The stability of such outcome indicators is reviewed and confidence intervals for ranks are developed. These techniques collectively will illustrate the unreliability of some presentations of data and results and the tendency to identify fewer under or over performers compared with basic techniques assuming independence of all patients.

Multilevel techniques can also give the researcher insight into the relationships between the lower and higher levels in the study and in some cases can help partition the variation between the hierarchies. A discussion of multilevel or random effects models introduces the concept of shrinkage. In a model involving hierarchies estimates for some individual surgeons may be unreliable purely because of low numbers of patients involved. By pooling data across surgeons we 'borrow strength' from the population as a whole. The shrinkage we see in a random effects model of relative risk estimates for surgeons or hospitals is an important feature and effectively the parameter estimates obtained in a random effects model are a compromise between those obtained from fixed effect regressions for individual and aggregate data. The performance measure for a surgeon with few cases whose crude estimate of performance is 'extreme' is shrunken considerably towards the population mean performance but a surgeon with large numbers of cases with the same degree of divergent performance will shrink to a lesser extent. This does raise concerns when shrunken estimates are used to construct rank orders of performance measures. These will inevitably be difficult to interpret and communicate to the public (Kreft & De Leeuw, 1998).

The statistical techniques which have emerged in recent years for analysing sets of hierarchical data have been formulated largely in the field of social science and in particular by educational researchers. The standard texts in the field by Goldstein (1995), Longford (1993), Bryk & Raudenbush (1992) and Kreft & De Leeuw (1998) all illustrate the theoretical development of these models with educational examples of pupils nested within classes which are nested within schools (and regions).
The parallels of these nested structures within the health care areas are obvious and recently much more interest has been shown in such models in the medical literature. Leyland (1995) produced one of the first significant applications of new techniques in multilevel or random effects modelling to medical applications when he examined the relationship between length of stay and readmission rates for certain diagnoses in Scotland. He standardized data with respect to certain explanatory variables and modeled the hierarchical structure of the data. A useful review article on heirarchical models in medicine is Rice & Leyland (1996).

The determination of power calculations is of particular interest. As always greater power is obtained if we have large numbers of observations, but consideration also needs to be given to the number of higher level units and the numbers of lower level units within these higher level units. Such problems are almost always approached using simulation methods. Results will also depend on the strength of the effects involved in the study (i.e. very powerful effects underlying the data generating process will mean less data is required to achieve a given level of confidence).

1.11 Classical or Bayesian Methods of Inference

Many statisticians have long held a preference, in many cases on purely philosophical grounds, to the Bayesian approach to statistics. Historically whilst the two contrasting approaches to inference have both had their enthusiastic proponents it has been the classical statisticians who have led the way in the analysis of actual complex sets of data in a variety of disciplines, including medicine. In particular the development of the theory for the well known class of generalized linear models and proportional hazards models dominated much of the methodological literature over the 1970's and 1980's. The availability and steady improvement of statistical computer software (e.g. GLIM) meant that the theoretical developments in classical inference could be implemented in practical data analysis context.

The main reason that Bayesian statisticians were less able to address complex multi-dimensional problems was the very significant computational difficulties associated with their approach. In particular Bayes' theorem relates a posterior distribution to the product of a prior distribution and the likelihood of the observed data where the scaling factor involves an integral expression. In addition many of the features of the posterior distribution which are of interest can be explored only after the evaluation of other complex integral expressions. The integration problem is therefore central to Bayesian inference. Standard references on Bayesian methods are Carlin & Louis (1996), Gelman *et al,*(1995) and O'Hagan (1994).

Bayesian theory taught to undergraduates pursuing degrees with a meaningful statistical content focuses almost exclusively on problems where analytic solutions to the integration problems are available, most commonly using conjugate prior distributions for the parameters of interest. In realistic problems such attractive closed form solutions are, regrettably, very rare indeed. In the examples considered in this thesis we are particularly interested in the marginal posterior distributions of certain parameters and especially in summary measures of parameters, particularly expectations, measures of dispersion and rank order statistics.

Clearly when addressing such an important and sensitive topic as the comparison of surgical performance it is crucial to reassure those examining or interpreting the results of any analysis that no bias has been introduced at a personal level by the statistician involved. This is raised by classical statisticians as a concern when Bayesian methods are used since they require the incorporation of a prior distribution for the various parameters in the model. In the majority of the examples that are considered in this thesis an uninformative prior has been used to alleviate any such concerns

Even when an uninformative prior is chosen we will see that the use of Bayesian Methods in analyzing hierarchical data can alter the inferences drawn quite substantially. In a hierarchical model where there is no longer independence of patients within surgeons we see shrinkage of the estimates for, say, hazards ratios, towards the overall mean effect.

The dangers of exclusive use of fixed effect analyses are highlighted. In particular one tends to find more statistically significant differences from average. As an example, if we only calculate hazard rates and confidence intervals based on a full independence assumption then the results for surgeons or hospitals with few cases can be highly unreliable since the expected values and confidence intervals are dependent almost exclusively on the sample for that higher level unit alone. If however we assume that surgeons in general might be expected to have similar outcomes then the means of the performance measures of the surgeons with few cases are shrunk towards the overall mean.

In later chapters we will see that the additional attractions of random effects models do come at the cost of reductions in statistical power for some studies of interest.

The WinBugs program (BUGS stands for Bayesian Inference Using Gibbs Sampling) and associated documentation have been developed by various authors at the MRC Biostatistics Unit in Cambridge and more recently at Imperial College in London (Spiegelhalter *et al* 1995, 1996 (a) & (b) and 1999 (b)).

1.12 A more formal description of Bayesian models

We will now briefly consider the theoretical development of Bayesian models. Consider a study where the model parameters are themselves considered random quantities (and the missing data might be considered here as well) and we observe some data generated by the assumed underlying probability model. We can denote the prior distribution for a set of model parameters as $p(\theta)$ and the likelihood $p(D/\theta)$ and these are combined using Bayes' Theorem to give an expression for the posterior distribution denoted $p(\theta/D)$

$$p(\theta/D) = p(D/\theta)p(\theta) / \int p(D/\theta)\, p(\theta)d\theta$$

This is the conditional (posterior) distribution of θ given our observed data D.

We will often be interested in aspects of this posterior distribution and may be interested in the marginal posterior distributions of parameters of interest (obtained by integrating out the other parameters). In particular we will often wish to consider the posterior expectations of functions since these enable us to address specific questions of inference. It is through these two aspects, the marginal posterior distributions and posterior expectations that we can consider the assessment of institutional performance within both a hierarchical and a Bayesian framework.

The posterior expectation of a function is expressed as follows

$$E[f(\theta)/D] = \int f(\theta)\, p(\theta/D)d\theta$$

We choose functions f(θ) which will give us insight into the features of the underlying process under consideration. Trivially if f(θ)= θ then we have an expression for the posterior expected value of θ.

In the above expressions we have to evaluate complex, possibly multi-dimensional, integrals and this is the area of the problem where what have come to be know as Markov Chain Monte Carlo (MCMC) methods are used. It is to be noted that MCMC methods are not the only solution to the integration problem but they are increasingly seen as the most attractive alternative, albeit a computationally intensive one. Alternative methods include Laplace approximations and some non stochastic numerical integration routines (Gilks *et al 1996*).

In essence when using MCMC techniques the integral is estimated by the mean of a simulated sample of values from $f(\theta)$. Following the example above we have,

$\hat{E}[f(\theta)|D] = \sum f(\theta i)/n$ where (θ_i : i=1to n) are a sample from the posterior distribution of θ.

The difficulty lies in obtaining a representative sample of values from the posterior density $p(\theta/D)$. How MCMC methods assist in this matter is to draw samples from a Markov Chain which crucially has $p(\theta/D)$ as its stationary distribution.

They are not independent samples from the distribution but this does not halt the development of an accurate approximate integral since we are just looking at the sample mean of a series of simulated values.

The literature and applications of these stochastic processes is vast and we will not cover the theory in any depth but the specific ideas behind these applications were first considered by Metropolis *et al (*1953) within the context of a physics application and were developed to suit more obviously statistical applications by Hastings (1970). Thereafter very little happened until the rapid development of computing power in the 1990's enabled the interesting but largely theoretical ideas of Metropolis and Hastings to be applied to meaningful problems.

Gilks *et al* (1993) discuss the Bayesian MCMC solution of specific medical examples and the development of modelling techniques are detailed more fully in Gilks *et al* (1994) and Best *et al* (1999). Comprehensive reviews of current theory and practice in this rapidly developing area can be found in Brooks (1998) and Gamerman (1997).

In some ways MCMC methods can seem unusual since most work with stochastic processes involves the determination of the unknown stationary distribution given some underlying probability model with a prespecified transition distribution (or transition matrix in a discrete example). The application of MCMC techniques requires, by contrast, the determination of the mechanism which will produce the required known stationary distribution. There are a number of different generation processes but the most commonly used in current practice is the Gibbs Sampler. This is discussed below but a final observation before considering technical matters is to appreciate that it will take time for any, even well defined, Markov Chain to approach its stationary distribution. Not all of the sampled values can be used to assess the numerical approximation to the required integral. Since we are only interested in values when the chain has approached stationary the initial values must be discarded. This is known as the burn in time.

If the first m samples are discarded before the average is calculated then we evaluate instead the following statistic.

$\hat{E}[f(\theta)/D]=\sum f(\theta_i)/(n-m)$ for i=m+1 to n

A number of issues are raised with this aspect of the problem, most obviously the choice of mechanism to generate the values, the determination of stationarity and the length of the initial `burn in'. The truncated summation above is sometimes referred to as the ergodic average which converges to the required expected value under the conditions of the ergodic theoem.

To emphasise matters the stationary distribution is known in advance, what is to be determined is a procedure to generate a Markov Chain with the required stationary distribution. As we increase the sampled realisations of the process we generate (in the limit) samples from the stationary distribution of primary interest, that is our posterior distribution.

There are a number of different algorithms which can be used in the solution of the MCMC problem but the most commonly used ones are the Metropolis Hastings Algorithm and a special case of this method, the Gibbs sampler. It is these methods which are commonly incorporated into the currently available software and in particular the WinBUGS software discussed and used in this thesis.

Another useful by product of the Gibbs sampler is the ease at which it enables study of the important marginal densities involved in any particular multi-dimensional problem. It is the study of marginal densities that is particularly important in the analysis of surgical audit data as we wish to assess the impact of individuals or specific characteristics on clinical outcomes. We simulate samples from the distributions of interest and make inferences through use of sample averages and other elements of the appropriate posterior distributions. We are particularly interested in mean values and measures of variability.

Confidence intervals (or more correctly, Bayesian credibility intervals) can also be calculated by identifying which observations are in the appropriate tails of the marginal distribution e.g. for a 90% credible interval for a particular factor we determine the 5th and 95th centiles from the observed marginal data. The available software packages also enable the illustration of the whole shape of the distribution, the 'kernel density'.

One particular problem requires further comment, the suggested length of any burn-in before obtaining our random sample from our stationary (posterior) distribution of interest. There is the associated problem of low long to sample after the decided burn in period. In some cases the sampler can give the impression of convergence whereas it has in fact not done so.

One solution is to make several runs from different starting values which are fairly well dispersed and to examine the results for convergence in the same areas. Some formal tests for convergence are available but in many cases if computing resources are extensive then a large burn-in and sampling period can be tolerated.

1.13 Some specific statistical references

The range of statistical methods used in audit has expanded enormously over the last 30 years. For some time audit was considered a helpful voluntary exercise and one that might highlight interesting features of process and be a useful educational tool but the idea of comparative studies of actual outcomes was an uncomfortable development for some. As discussed earlier in this Chapter, in the 1980's interest gathered pace significantly with a major study, the National Confidential Inquiry into Perioperative Deaths (Buck *et al* 1987) and more widespread discussions in the medical literature.

Efforts to compare mortality rates were rather basic in the formative days of audit. The National Confidential Inquiry into Perioperative Deaths proceeded to examine cases individually and then allocated these to 'avoidable' and 'unavoidable'. The most recent work done by this Inquiry group represent a considerably more structured approach to analysing surgical outcomes.

Pressure then started to develop on audit specialists to deal more comprehensively with the risk characteristics of patients. This was generally achieved using scoring systems (Playforth *et al,* 1990). The POSSUM scoring system (Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (Copeland *et al* 1991) uses a scoring system for features of severity in general surgery and then uses a logistic regression models to allow for the prior risks when modelling mortality or morbidity. The explanatory variables include 12 physiological parameters (age, cardiac history, blood pressure etc) and 6 operative parameters (severity, multiple procedures, blood loss etc).

The system has been used in practical audits including a comparative audit of colorectal resection (Sagar *et al,* 1994), vascular audit (Copeland *et al* 1993) and a prospective analysis of general surgery (Copeland *et al*, 1995). Hartley & Sagar (1994) for completeness and comparison also considered 'gut feeling' as a predictor of postoperative outcome.

More recently in 1998 this system has been extended to the P – POSSUM scoring system (Wijesinghe *et al*, 1998 and Prytherch *et al* 1998). The extensions of the original POSSUM models were to account for the seeming over prediction of death, particularly in low risk patients. Some studies however indicate that even with P-POSSUM the risk of death is still overestimated.

Other scoring systems based on multivariate logistic regression models include the Parsonnet risk stratification system for cardiac surgery (Parsonnet *et al,* 1989). This scoring system stratifies patients into levels of predicted operative mortality. The model includes 14 risk factors and determines the probability of death within 30 days. The authors advocated the use of scoring systems and associated regression models in exercises comparing surgical and institutional performance. Nashef *et al* (1992) further examined the Parsonnet scoring system in an audit of patients undergoing open heart surgery in 1991. They highlighted the relative ease of collection of the data required for the scoring system and concluded that the Parsonnet system did indeed make it possible to allow for case mix in UK based audit studies.

Some reservations exist with surgeons about the optimal list of factors entering the scoring system but this does not invalidate the broad applicability of the method.

The APACHE, APACHE 2 and APACHE 3 scoring systems for intensive care patients are largely developed in similar way (Knaus 1981, 1985 and 1991). A more general comment on the issues arising from the use of scoring and predictive models can be found in Knaus (1988). As before personal and physiological items of data are recorded and a risk of negative outcome (death) is calculated from a logistic regression model (although the first development used subjective weights).

Jones *et al* (1992) compared POSSUM and APACHE 2 in their relative success at predicting outcome from a surgical high dependency unit. They concluded that the former scoring system was to be preferred but this should be seen in the context of the comments regarding under reporting of mortality discussed above. Further discussions on allowing for pre-operative status are to be found in Tremblay (1993) and Schein (1988).

Jones & Cossart (1999) comprehensively review most available models for general surgery and conclude that the POSSUM method (and its derivative) is still the most appropriate for use in practice. Successful models have also been developed to predict outcome following severe head injury (Murray, 1986). In this case the models actually predict outcome very well, a consequence of the high mortality rate and clear negative outcomes associated with injuries of increasing severity as measured by the Glasgow Coma Scale.

McArdle and Hole (1991) published a comparative analysis of outcomes after surgical treatment of colorectal cancer within Glasgow Royal Infirmary highlighting the wide variation in outcomes between surgeons within a single hospital. They observed significant differences between surgeons even after adjusting for case mix in a Cox proportional hazards regression procedure.

League tables then came to the fore in 1993. Studies of waiting times caused particular consternation. Whilst such tables have very clear disadvantages these early publications did focus attention on many issues such as resource allocation and regional social effects.

In 1994 a major step forwards was taken with the publication, in Scotland, of tables of mortality rates for the first time. Since then there has been an explosion of interest in comparative statistics of all kinds. There remains considerable clinical reservation over the use of 'league tables' (Sanderson & McKee, 1998) and, as discussed earlier, this is backed up by statistical analyses as reported by Marshall & Speigelhater (1998).

As discussed earlier in this chapter developments in the area of surgical audit in the United States of America have often proceeded at different rates than in the United Kingdom and in the US publication of outcome data started far earlier. Audit results are required for a purpose. Amongst other things, results are there to enable surgical performance to be assessed. In particular they allow for extreme out-performance or under–performance to be identified and acted upon in a timely fashion. This causes the inevitable problem for the data analyst. Ideally we would prefer the study to be able to identify such extremes of performance quickly but due to the variability of surgical outcomes a considerable number of cases are required before the study has enough statistical power to enable conclusive inferences to be drawn.

One recent approach drawing on statistical techniques developed in the area of industrial quality control is the use of cumulative sum techniques where a continuous audit process is envisaged with new cases incrementing a surgeon's case load and statistically significant extremes of performance are looked for on an ongoing basis (as opposed to periodically at the end of a pre-specified period of case acquisition or follow up). Initially formulated by Williams *et al* (1992) the method has been enhanced and extended by various other authors.

Poloniecki *et al* (1998) discuss allowing for case mix in the CUSUM process (using the Parsonett scoring system in the context of cardiac surgery). Specific attention was focused on the number of operations required to detect an excess of mortality, that is a study of the statistical power to detect differences in performance. A conclusion drawn was that league tables of performance were structurally unreliable since large variations in mortality are observed even when the underlying case mix does not vary. Steiner *et al* (2000) extend the CUSUM procedure further making adjustments to take account of the evolving case mix of the surgeon under consideration. Importantly they explore the effects of run length in their analysis. CUSUM methods may perhaps be a rather idealised solution to the audit process given the complex mix of factors involved but they are certainly a useful early warning system which might indicate that more thorough analysis and investigation is required.

Lawrance *et al* (2001) followed the above methodology in a recent observational study where they used cumulative mortality data to assess relative hospital mortality for patients with acute myocardial infarction.

Perhaps as a guide to future trends Jarman *et al* (2001) use publicly available data (the hospital episode statistics) to compare hospital performance and revealed considerable variation in outcome. The authors concluded that the percentage of emergency admissions was a powerful predictor of outcome.

Of specific interest in the context of this thesis is the recent paper by Andersen *et al* (1999) which considered the problem of testing for centre effects in multi-centre comparisons of survival data and highlighted the relative efficiency of random effects models as opposed to fixed effect models. Matsuyama *et al (*1999) performed a similar exercise again exploring the data from a hierarchical perspective.

Important theoretical contributions have been made by statisticians and heath economists working in the United States using techniques from econometric analysis to assess hospital performance and quality. McLellan & Newhouse (1994) discuss the problem of dealing with unobservable characteristics (selection bias) which can influence performance measures in an analysis of data on treatment of acute myocardial infarction. They use instrumental variables based on distance to treatment centres to control for unobserved case mix variation. Gowrisankaran & Town similarly address the problem of selection bias controlling for severity of illness using an instrumental variables approach based on distances from treatment centres. They concluded, after an extensive analysis of outcomes following onset of pneumonia in the elderly, that selection problems remain even after controlling for case mix in traditional models introducing bias into estimates of hospital quality. A working paper also involving the same authors (Geweke *et al*, 2000) continues this theme analysing a similar set of data using models that can incorporate the possibility that the greater the probability of mortality due to unobserved characteristics the more likely is admission to particular centres.

They conclude that patients with greater unobserved severity of illness are admitted to better hospitals. There is a risk there fore that performance measures are biased being poorer for better performing institutions and *vice versa.*

The statistical methods used in audit studies have then advanced as attention has been more keenly focused on institutional performance in general. Before concluding this introduction we state the main statistical methods used in practice with a brief commentary of the effectiveness of the approach .

Presentational and Descriptive Methods :- When communicating results to the public and even amongst experienced practitioners it is important to use effective presentational methods. There is also a requirement to avoid making such presentations misleading. Graphs of comparative mortality rates should be scaled appropriately and importantly any productions of estimates of performance in a graphical way should also include a corresponding measure of uncertainty so that the reader can see the range in which the underlying performance may lie. This is particularly true for the preponderance of league tables now in the public domain. There should be associated commentary on the poor levels of confidence attaching in many cases to rank order estimates.

Fixed Effect Analyses :-Traditionally the measures of relative performance have been the relative risk (or relative hazard for survival times analyses). All of the scoring systems and audits referred to in this chapter, and performed later in this thesis, proceed by adjusting for case mix. This is generally achieved by allowing for the various variables in a regression model. The surgeon or institution specific effects (and confidence intervals) then flow naturally from the structure of the model by fitting a binary variable identifying the surgeon in addition to the case mix variables. These analyses are carried out in the main on widely available software packages. We have used Minitab, S-Plus and SPSS (Minitab Inc, 2000, Mathsoft Inc, 1997 and SPSS Inc, 1998)

Random Effects Analyses :- To fit more realistic models taking into account more complex but intrinsic features of surgical audit data one has to use more recently developed software such as WinBUGS. This software has been used by researchers in the recent past to develop a very important methodology for dealing with rank orders and the confidence attaching to them. (Marshall & Speigelhalter, 1998). It has also been used in the final synthesis of statistical work undertaken for the Bristol Royal Infirmary Inquiry where a cyclical approach was taken by repeatedly excluding an institution from the analysis and then using the remainder to provide a basis for estimating the number of excess deaths for that institution (with an associated confidence interval). These different types of statistical approach will be discussed in the Chapters that follow in this thesis.

This brief review of current practice has shown that that the gradual development in largely voluntary and informal surgical audit up to the early 1990's has been replaced by a massive infrastructure of Government and NHS Initiatives. The establishment of the National Institute for Clinical Excellence and the Clinical Standards Board for Scotland has placed consideration of audit and clinical governance much higher on the agenda for health care professionals. Audit is now a pre-requisite for clinicians of all types. The Bristol Royal Infirmary Inquiry has led to more general public interest and awareness of surgical performance and the recent report by Dame Janet Smith from The Shipman Inquiry (2002) into the criminal activities of Dr Harold Shipman will focus concern on the monitoring and governance of a range of heath practitioners. Audit, and particularly the analysis of surgical outcomes, has come from being almost a side issue in the medical profession to being an integral part of the regulatory environment within which health care professionals work within twenty years. It is hoped that the work undertaken in the production of this thesis will help in some small way to guard against abuse of performance indicators of many kinds now in the public domain.

The submission of this thesis comes at a time when various new policy and statistical developments have emerged. It will be of considerable interest to see how the discussion and monitoring of clinical outcomes evolves in future after such a rapid burst of activity in the last 5 years.

Chapter 2 - A study of the relative performance of surgeons treating patients with colorectal cancer in a single hospital.

2.1 Introduction

In Chapter 1, Section 1.4 we discussed how previous clinical and audit studies highlighted the poor long-term survival experience of patients undergoing surgery for colorectal cancer. Statistical analyses have highlighted the substantial degree of variability in outcome between individual surgeons even after adjustments have been made for the differing case mix handled by the individuals involved (McArdle & Hole, 1991 and Porter *et al*). This chapter presents a case study of the established methodology used in surgical audit investigations examining both the stability of measures of the relative performance of individual surgeons with respect to differing choices of case mix variables and the reliability of assessments made about the performance of any particular individual. Data covering 1128 patients who presented at Glasgow Royal Infirmary with colorectal cancer between 1974 and 1984 were analysed using various statistical methods, principally the Cox proportional hazards model (Cox, 1972). The results of this investigation, which has a very long period of case acquisition and follow up, highlight the fact that conclusions drawn from surgical audit investigations in specialised disciplines such as colorectal cancer surgery can be unreliable. It is notable that this study, comprehensive as it is, is based within one institution. Later chapters in this thesis will address problems concerning the adequacy of surgical audit investigations when we deal with shorter periods but with large numbers of institutions. We conclude that the development of more complex models might increase the reliability of audit studies but at the expense of simplicity, and existing methods for assessing performance based on clinical outcomes are perhaps more suited to areas of surgery where greater numbers of patients are treated by each individual surgeon.

2.2 The factors influencing survival

A number of studies have highlighted the poor survival experience of patients undergoing surgery for colorectal cancer. A significant proportion of patients present as emergencies with advanced disease and longer term outcome remains poor even for those undergoing what was considered by the surgeon to be a curative procedure. Patients undergoing elective procedures exhibit better outcomes but the extent of disease at presentation (and the associated poor outcomes) remains a significant concern. Despite improvements in surgical techniques and peri-operative care, outcome has shown only modest improvement over the last 25 years (Clinical Resource and Audit Group, 2000, McArdle *et al*, 1996).

It is intuitively obvious that characteristics of individual patients strongly influence surgical outcomes and that as a consequence greater insight into the quality of the surgical process is gained if proper adjustments can be made for the mix of cases in any particular study. Statistical techniques and presentational methods can assist in determining those factors of primary importance and in the quantification of the effects of such factors.

When studying patients undergoing surgery for colorectal cancer a number of factors are known to be strong predictors of long term outcome. Patient specific predictors of long term survival include age, sex, the operative procedure involved and admission status (that is elective versus emergency). The disease specific predictors recorded will vary depending on the particular data collected in any individual study and include measures of spread such as the Dukes' Stage determined at pathology , the extent of local invasion of the tumour and the presence of liver metastases. Measures of the aggressiveness of the tumour may also be recorded, an example being the level of differentiation exhibited. (Dukes, 1932 , Deans *et al* 1994 and Spence, 1994),

To a degree when several of these factors are recorded a choice of different covariates can be explored. This is because many recorded variables are in fact related to the underlying severity of the disease in question. They are proxies for each other, many representing the same disease characteristics in different ways. It is therefore of interest to explore whether the choice of variables to be included in the statistical model used to deal with case mix effects influences the case mix adjusted measures of performance.

The surgeon involved can also introduce a degree of variability into long-term outcome and past studies have examined the impact of this important external variable on subsequent survival experience. Previous studies have examined these matters in some detail and highlighted the wide variation in post-operative complications and mortality between different surgeons.

The power of statistical methods to detect differences in performance will depend on many factors: -

- the total case numbers
- the numbers of cases per surgeon and the distribution of these numbers
- the numbers of deaths observed
- the number of cases lost to follow up for various reasons
- the length of the study
- the period of acquisition for patients
- the explanatory variables which have been recorded.

Conclusions about statistical power from other survival studies cannot be easily generalised. In addition in an audit study a balance has to be struck between the need for rapid assessment and feedback on performance and the requirement for an adequate number of cases to be analysed and for an appropriate period of long term follow up.

2.3 Historical Data and Methods

An established methodology has developed for analysing surgical audit data that allows for case mix adjustments. This is primarily based around the computation of hazard ratios using the Cox Proportional Hazards Model (Cox, 1972 and Cox & Oakes, 1984). Considerable residual variability however remains when the actual performance of individual surgeons is examined. This is a consequence of both the inherent variability of surgical outcomes in specialist disciplines and the low number of procedures undertaken by any individual surgeon. The Cox model enables efficient adjustment for case mix to be made and many of the parameter estimates have natural interpretations for the audit specialist. The Cox model is based around survival times as opposed to specific outcomes at a prescribed point in time. It copes with censoring in an efficient manner when a binary logistic model might have to discard useful information relating to such cases.

McArdle and Hole (1991) analysed a data set obtained prospectively from 645 patients with colorectal cancer presenting at Glasgow Royal Infirmary between 1974 and 1979 and reported a wide variation in measures of the relative performance of individual surgeons even after adjustments for case mix had been made.

The measure of relative performance used by McArdle & Hole was the hazard ratio, a measure of the mortality risk for a particular surgeon relative to the other surgeons included in the audit. If a surgeon has a calculated hazard ratio below one this would indicate in the context of this discussion that the survival experience of his patients was better than those of his peer group. The estimated hazard ratios are of course only point estimates of the performance measure. One also has to look at the level of statistical confidence attaching to such an estimate. In later chapters we will see that even this approach has weaknesses and greater insight can be gained by examining the whole distribution of performance measures to examine whether an individual is not just statistically different from average but appears to be an outlier from the distribution of performance measures.

2.4 The objectives of our case study

We have built upon the analysis of McArdle & Hole by, in effect, extending their analysis with the benefit of several years additional data on surgical outcomes. This larger set of data enables us to explore specific issues of interest.

Data on a further 483 patients admitted to the same hospital, the Glasgow Royal Infirmary, between 1980 and 1984 were collected prospectively and the combined data set has been analysed further. In particular our objective has been to examine the adequacy of the prevailing methodology used in these and similar studies with a view to improving the strength of future investigations in related areas. Specifically we consider the following two statistical issues in some detail: -

1. How sensitive are performance measures to the particular choice of factors adopted for case mix adjustment and can we make reliable inferences on performance from the data which have been recorded ?

2. How reproducible are performance measures?

In attempting to answer these questions we have to explore the data in some detail :-

- We must consider the reasonableness of merging the two sets of data and examine the extent of missing data

- The basic results of McArdle & Hole should be able to be reproduced.

- We can then explore the sensitivity of performance measures to choices of covariates used in the Cox regression model.

- The reproducibility of performance measures can be considered with internal comparisons of performance within each surgeon's case load.

2.5 A description of the data and preliminary investigations regarding comparability

The Glasgow Royal Infirmary data set has been described in separate papers and publications and we will not therefore describe it again in detail. In summary, a total of 1128 cases were considered and information for each case included the following variables (McArdle & Hole , 1991 and McArdle *et al*, 1996) :-

Patient specific fields:- sex, age, marital status, occupation, weight, smoking history, date of admission, family history, date of surgery, patient identifier codes, surgeon, GP, presentation (emergency or elective), procedure (e.g. curative resection, palliative resection, laparotomy only etc)

Disease Specific fields:- spread, site of tumour, differentiation and invasion of tumour, Dukes' Stage, evidence of metastasis to liver.

Outcomes :- post-operative complications and cancer recurrence (wound infections, anastomosis leak, haemorrhage etc), discharge dates and information, follow up dates (or evidence of loss to follow up), recurrence information (site and date observed), additional treatment, date of death, cause of death

As discussed the data set under consideration comprised of two parts representing cases from 1974-79 and cases from 1980-84. For the two separate time periods we examined the distributions of the various factor levels for the major explanatory variables of interest and these were shown to be stable. An example of this would be the fact that the proportion of patients being admitted as emergencies or with particular recorded levels of Dukes' Stage were seen to be stable between the two series. We also looked at the stability of the proportions for Leakage, Wound Infection, Type of Procedure, Site of Tumour, Admission Status, Age and Sex. All supported the basic conclusion that there is no obvious concern about analysing a merged set of data. This enabled analyses to proceed using the merged data set without particular concern that the mix of patients was altering markedly over time in such a way as to complicate elements of our statistical investigations.

In fact a global statistical test can be undertaken to assess whether a factor representing the periods 1974-79 or 1980-84 would contribute anything to explaining the residual variability in the model. This did not approach any level of statistical significance and so the conclusions arrived at purely by inspection were seen to be valid. In addition simple statistical tests of the null hypotheses that the proportions in various categories were the same did not reach levels of significance, although this is slightly complicated by the profile of missing data in some of the covariates.

The merged data were then examined in more detail with software packages, primarily SPSS, Splus and Minitab. Additional fields were created to calculate survival times from surgery (or admission for unoperated cases) and to identify particular surgeons (a binary variable coded to highlight the surgeon in question relative to all other surgeons.

2.6 The treatment of missing data

Missing data were of some concern and an inspection confirmed that missing values tended to occur as a result of certain variables not being recorded at the time of the surgical procedure. This was usually because the patient in question had a sufficiently poor prognosis that no operative procedure was in fact undertaken and as a result certain items of information, which would have only been determined after pathology, could not then be recorded.

An analysis of only complete cases discards potentially useful information but replacing missing values with information imputed from the available set of data leads to an underestimate of variability. There is a danger in underestimating variability since it increases the risk of identifying an individual (or institution) as being different from average when in fact no difference exists. Consequently, we imputed values for only the more obvious cases. As an example missing Dukes' Stage values for unoperated cases were recoded as being Stage D before any analysis was performed. This was based on the assumption that the reason no operation was performed was that the cancer was already widely disseminated, which was generally confirmed by the recording of the presence of liver metastases.

Since missing data were largely confined to those patients with more advanced disease, who would have been unlikely to have undergone a tumour resection, any adjustments we made for missing values were in fact of little consequence in the primary analyses, which focused on curative resections.

2.7 The appropriate choice of data and outcome measure

The stated objectives of any given study will influence the choice of an appropriate data set and statistical method. A choice has to be made of both the groups of patients to be included in the study and the outcome measure appropriate to that particular group. Ideally the chosen patients should form a homogeneous group and statistical methods using these data will then offer more reliable indications of discrimination within the pool of surgeons of interest.

In this study we are primarily interested in the extent to which surgeon related variability impacts on long term survival. It might therefore reasonably be argued that the most appropriate subset of the data to analyse would be one which included only cases where the procedure was considered by the surgeon to have been potentially curative with the outcome of interest being death from a cancer related cause. Another valid study might have been an analysis of all emergency admissions where the outcome measure could be death from any cause with a shorter-term outcome measure being death within 30 days.

Whilst making use of homogeneous subgroups from a larger data set using a tailored outcome measure is an attractive proposition it does inevitably involve a reduction in available patient numbers with a consequent decrease in the statistical power of any analysis. In studies of specialised areas of surgical practice low numbers of cases often limit the strength of conclusions drawn and, despite the large data set available from the Glasgow Royal Infirmary, this study is no exception to this general rule.

Whilst larger numbers of patients increase the power of our analysis there is a subsidiary feature of the data which is also of relevance in surgical audit investigations. Apart from wanting to examine a large homogeneous group of patients with an appropriate outcome measure we would also ideally want to be analysing cases which did not include large numbers of patients with outcomes that were likely to be either very good or very poor. When the objective of the audit exercise is to examine the extent to which surgeons can influence long term outcome it is the middle group of patients, those without a very clear prognosis, who offer the greatest scope for statistical methods to differentiate between surgeons.

The choice of an appropriate group of patients and outcome measure also influences the selection of factors to be included in the statistical model designed to allow for the mix of cases under consideration. We have, in the main, followed the explanatory variable selection procedure used by McArdle & Hole, having confirmed the continued applicability of their choice of factors for the merged data set. Stepwise regression procedures for both the Cox model (using survival times) and a logistic model (for 30 day deaths) confirmed the reasonableness of the particular subset of covariates which are used.

In our investigations reported below we have primarily used the set of cases involving curative procedures with the outcome measure being death from a cancer related cause. We have also performed a subsidiary, but more all embracing analysis, involving all cases in the ten-year period where the outcome of interest was death by any cause. An analysis of only emergency cases would be another possible study of interest.

Whilst the basic data are extensively discussed elsewhere (McArdle *et al*, 1996) Appendix 1 lists some summary statistics for further information in addition to the analysis that follows.

2.8 *A basic description of the data*

At this stage it is worth tabulating the numbers of patients, deaths and the associated mortality rates per surgeon :-

Table 2.1

Glasgow Royal Infirmary – Colorectal cancer surgery

Numbers of cases, deaths and mortality rates for 1974-79 and 1980-84

	1974/79			1980/84		
surgeon	cases	deaths	rate	cases	deaths	rate
1	98	16	0.16	73	18	0.25
2	11	2	0.18	100	12	0.12
3	52	14	0.27	34	4	0.12
4	21	3	0.14	62	7	0.11
5	66	8	0.12	12	5	0.42
6	13	2	0.15	62	6	0.10
7	37	11	0.30	24	5	0.21
8	58	9	0.16			
9	52	7	0.13			
10	34	7	0.21	16	2	0.13
11	46	5	0.11			
12	32	4	0.13	10	1	0.10
13				39	5	0.13
14	38	3	0.08			
15	36	5	0.14			
16				23	12	0.52
17	21	2	0.10			
18	14	5	0.36			
19				8	2	0.25
20	5	1	0.20			
21				5	1	0.20
22	2		0.00			
23	1		0.00			
unknown	8		0.00	15	7	0.47
totals	645	104	0.16	483	87	0.18

Several points are clear by inspection even without the benefit of statistical tests of comparison. Firstly, the overall death rates are stable between the periods. Secondly, the variation in crude mortality rates is large between surgeons and within surgeons, between periods. This could of course be explained by different case mix. Thirdly, a number of surgeons have caseloads which are too small to allow the drawing of any firm conclusions about their performance (from an examination of mortality statistics alone).

The table below illustrates the pronounced effect that the severity of disease (as measured by the Dukes's Stage) has on survival times. In general disease staging is used by clinicians to aid them in developing a prognosis and course of treatment for a patient. In this case it clearly splits patients into categories with different outcomes and as such is useful for risk adjustment in a statistical model. It is appropriate to analyse cancer related survival data allowing for the case mix of the surgeons in addition to examining crude survival rates alone.

Table 2.2

The Variation in Survival Percentages by Dukes' Stage

Dukes' Stage	Percentage Alive After 3 Years
A	85%
B	56%
C	38%
D	5%

Similar discrimination between patients is also seen when examining survival rates stratified according to admission status, differentiation and spread. Age is a further variable that explains some differences in outcomes and conveniently can be dealt with as a binary variable denoting age less than or greater than 65.

This cut off point was determined by examining the Kaplan –Meier curves for a variety of patient groups stratified according to bands of ages at admission (or surgery) and performing the associated log rank tests (Altman, 1991). Banding in groups of 5 years of age was generally used in the calculations.

These tests show that survival times are relatively stable across a band of ages up to 65 but there is deterioration in outcomes after that age. Further, within this older second group survival experience is broadly similar with respect to age (adjusting for other variables). In effect there appears to be almost a discrete change in survival experience at age 65 and a binary variable can therefore be used as opposed to modelling age as a continuous variable. In any event the objective was not to improve on the model developed by McArdle & Hole but to examine the reliability of inferences made given that a particular model had been adopted. For completeness however we report in Table 2.3 below the broad profile of survival rates at two years duration. The groupings are extended beyond 5 years at the extremes of age to obtain an adequate sample in each category.

Table 2.3

The variation in mortality rates by age for all cases –2 years after surgery

Age Group	Mortality rate
Less than 50	23%
50-54	27%
55-59	27%
60-64	26%
65-69	31%
70-74	37%
75-79	39%
75-80	39%

We evaluated the consultant hazard ratios fitting a binary variable for the surgeon factor in addition to age (a binary age variable split at age 65), sex, admission category, procedure (where appropriate) and Dukes' Stage. In the interests of using a relatively parsimonious model we did not fit any additional variables for this illustrative exercise. In fact additional available variables explain little of the residual variability (which is substantial).

Due to sample size considerations our analysis was largely confined to the fifteen surgeons who had an aggregate number of patients in excess of 35 when observed over the entire 1974-84 period.

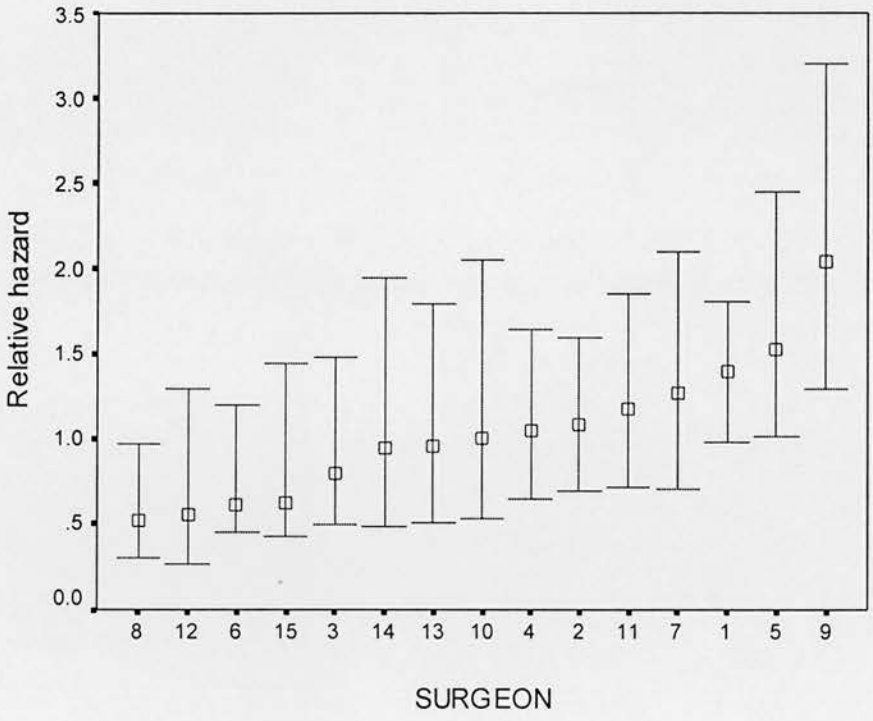
Before examining the results of our case studies we illustrate the general variation in performance measures for the whole set of data. The figure below shows the ordered progression of hazard ratios with associated 95% confidence intervals plotted around the mean hazard per surgeon. The population was those patients undergoing a curative procedure and the outcome measure was death from a cancer related cause. The obvious point of note is that, for the subset of data where we expect the most discrimination between surgeons, the confidence intervals are mostly overlapping. Some surgeons are identified as being significantly different from average in that their confidence intervals exclude unity. In later chapters we will illustrate the effect that overlapping confidence intervals have on the distribution of rank orders when one is examining surgical performance league tables.

Figure 2.1

The relative hazards of surgeons measured over the period 1974 to 1984

(Curative resections)

(Point estimates with individual 95% confidence intervals)



2.9 Results - The Stability of Hazard rates over Time

For each consultant with case numbers over 35 we looked at the stability of the hazard rates when one divides the case load into two distinct parts. For each surgeon we ordered their case load by date of admission and then divided the total series of patients into two halves, denoted the 'early' and 'late' cases (as measured chronologically).

Many combinations of outcome measures and subsets of the data were explored but we report only the main category of interest for an audit exercise, that is the subset of patients undergoing a curative resection. These are the subset of operations where the surgeon believes that the prognosis can be significantly improved by surgery as opposed to a palliative procedure. The outcome measure is death from a cancer-related cause. Case mix is allowed for in the calculations. The table below gives the hazard ratios and rank orders of the surgeons when case mix has been allowed for in the calculations.

Table 2.4

The relative hazards for surgeons in two periods (measured chronologically) (all cases)

Surgeon	Early Hazard	Late Hazard	Early Rank	Late Rank
1	0.99	1.29	6	13
2	0.99	1.05	7	9
3	1.09	0.68	11	3
4	0.99	1.00	8	8
5	1.29	1.09	13	10
6	0.76	1.11	2	12
7	1.49	0.82	15	4
8	0.91	0.63	5	2
9	1.44	1.39	14	14
10	0.61	1.47	1	15
11	1.03	1.10	10	11
12	1.17	0.50	12	1
13	0.99	0.95	9	6
14	0.86	0.96	4	7
15	0.82	0.87	3	5

The hazard ratios for surgeons 7 and 9 were significantly different from unity in period 1 and the hazard ratios for surgeons 1 and 9 were significantly different from unity in period 2. Since there is a possibility that there could be a systematic difference in performance between time periods (say improvement with experience) we decided not to analyse the time dependent data to any further extent preferring a different approach (see below).

What is clear is however that with sample sizes of the order under consideration and for this particular disease type the hazard ratios can diverge quite substantially in different study periods. The rank orders show this feature of the results even more clearly with surgeon 10 falling from top to bottom of a notional 'league table' between the two periods. This is clear practical evidence of the theoretical comments made on league tables in Chapter 1, Section 1.10.

2.10 Results - The Reproducibility of Hazard Rates

We proceeded to determine (for each surgeon) the separate hazard ratios that were obtained by dividing his or her caseload randomly into two groups. The objective of this exercise was to generate an artificial situation where we knew, by construction, that the surgeons' performance did not differ over the two case series. The use of random sampling is of course a key part of many formal statistical tests and underlies the theory behind modern experimental design and, for example, the large and important discipline of clinical trials in their many shapes and forms. In this instance the use of a random allocation of cases can be interpreted as an attempt to remove any possible systematic bias in the data associated with, for example, a general improvement in surgical skill associated with increasing experience. The preliminary conclusions we drew from the analysis of data split into two parts purely on time grounds are open to criticism for this reason.

Calculations were made using all cases with the outcome of interest being death by any cause and for curative procedures only with the outcome of interest being death by a cancer related cause.

For both all cases/all cause mortality and curative resections/cancer related deaths we allocated at random the cases for each surgeon into two approximately equal groups and calculated the individual hazard ratios in each group. This was done separately for each of these two patient and outcome groups to avoid the risk of unequal sample sizes occurring in the more restricted set of data.

The results are conveniently presented in the form of a graph the co-ordinate of any point being the representation of a surgeon's group one and group two hazard ratios. These were computed for the two data sets both allowing for case mix (with the same factors as detailed above) and also using the crude unadjusted data. Only the case mix adjusted results are reported in what follows but the Cox Proportional Hazards Model adjustments, as expected, do indeed redistribute the hazard ratios somewhat and in particular they bring the extreme points back towards the centre of the graph. It is also possible to plot the bounds of the confidence area or region associated with any individual co-ordinate but, for this study the numbers involved and the widths of the intervals make this presentationally rather confusing and so only the point estimate are plotted.

We display the graphs for the two randomised data sets below in Figures 2.2. and 2.3. The data appear to be weakly correlated. The hazard ratios and rank orders of the surgeons are detailed in tables 2.4 and 2.5 and show similar conclusions. Together, the graphs and tables highlight the fact that there are a number of consultants who perform quite differently when their two separate groups are analysed. Looking at the curative resection results the 'extreme' consultants 5, 8 and 9 remain identified as being different from average but consultants 2, 10, 12 and 14 display very different performance in the two randomly selected groups of cases under consideration. To emphasise matters surgeon 2 (who has the second highest number of patients) has a hazard ratio for curative resections of 1.23 in the first group of cases which falls to 0.78 when the second group is analysed.

Table 2.5 - Comparison of hazard ratios in two randomly selected subgroups - curative resections

Surgeon	Group 1 Hazard	Group 2 Hazard	Group 1 Rank	Group 2 Rank	Case Numbers
1	1.02	1.60	8	14	85
2	1.27	0.76	11	4	60
3	0.69	0.89	6	6	36
4	1.21	0.94	10	7	48
5	1.70	1.43	14	12	38
6	0.59	0.76	3	3	41
7	1.38	0.98	12	9	29
8	0.60	0.41	4	1	31
9	2.21	1.76	15	15	34
10	1.47	0.65	13	2	19
11	1.15	1.14	9	10	30
12	0.27	1.16	1	11	21
13	0.87	0.98	7	8	21
14	0.46	1.48	2	13	22
15	0.62	0.77	5	5	19

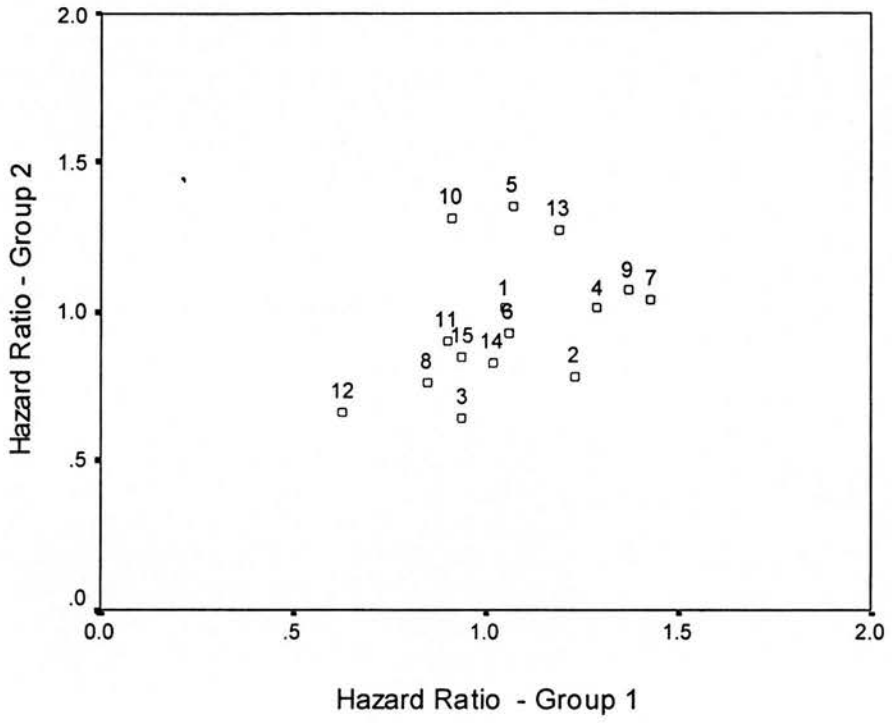
Table 2.6 - Comparison of hazard ratios in two randomly selected subgroups - all cases

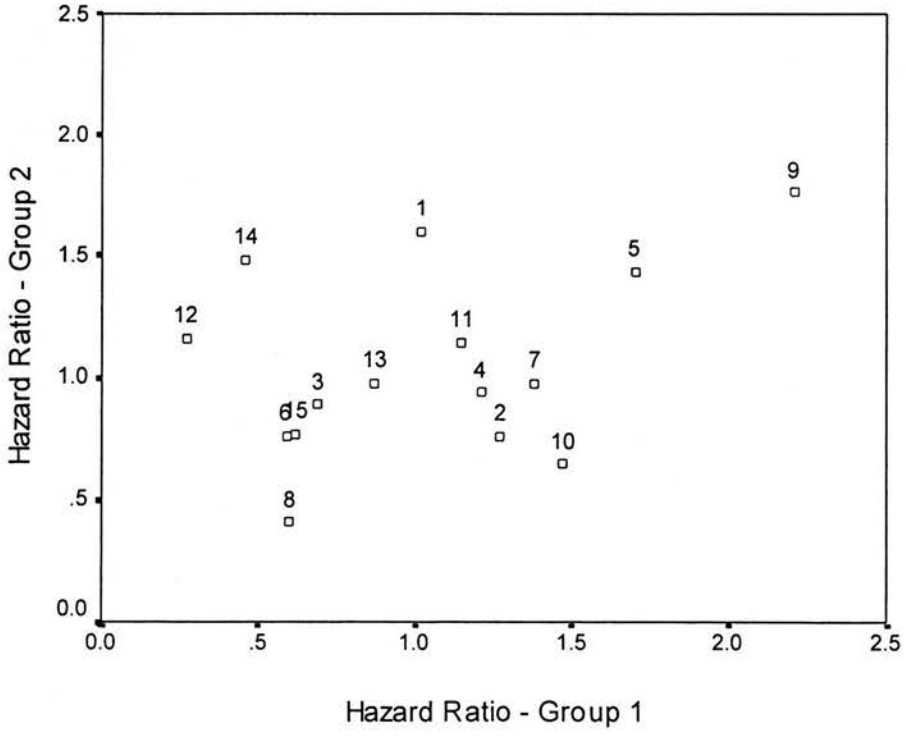
Surgeon	Group 1 Hazard	Group 2 Hazard	Group 1 Rank	Group 2 Rank	Case Numbers
1	1.05	1.01	8	10	171
2	1.23	0.78	12	4	111
3	0.94	0.64	5	1	86
4	1.29	1.01	13	9	83
5	1.07	1.35	10	15	78
6	1.06	0.93	9	8	75
7	1.43	1.04	15	11	61
8	0.85	0.76	2	3	58
9	1.37	1.07	14	12	52
10	0.91	1.31	4	14	50
11	0.90	0.90	3	7	46
12	0.63	0.66	1	2	42
13	1.19	1.27	11	13	39
14	1.02	0.83	7	5	38
15	0.94	0.85	6	6	36

Figures 2.2 and 2.3

The hazard ratios in two randomly selected subgroups

(all cases and curative resections)





Clearly the small numbers of procedures per surgeon has a pronounced effect on the size of the confidence intervals. To explore this in a little more detail we performed the following empirical exercise. Some corresponding theoretical results could also be obtained. We created a series of hypothetical groups of cases by taking random samples of increasing size (from 25 to 550) from the total list of 1128 patients. The total number of patients was fixed at 1128. We then calculated the hazard ratios and associated confidence intervals for these sampled groups as compared with the remaining patients, again using the Cox model making the same allowance for case mix as was made in the previous calculations for individual surgeons. The calculations were made for all cases and for cases involving curative resections alone. The table below (for all cases) demonstrates the effect of increasing sample size on the confidence we might have in any one particular computed hazard ratio.

Table 2.7

The width of confidence intervals as sample size increases

Random Sample of Patients	Width of Confidence Interval
50	0.65
75	0.55
100	0.45
200	0.38

2.11 Results - The sensitivity of hazard ratios to the choice of explanatory variables

We then examined the sensitivity of individual surgeon hazard ratios to different combinations of prognostic factors used in the statistical model designed to make adjustments for case mix. As discussed earlier in the introduction to this Chapter a number of items of recorded information could be used as a proxy for a general measure of 'severity of disease'. In our main investigations we chose to use Dukes' Stage since this classification split the data into four distinct subgroups with clearly different survival experiences. It would however also have been possible to fit the variable for local invasion of tumour or the extent of differentiation of the tumour amongst others. Indeed different studies may record different information and it is therefore of interest to examine how the individual surgeon hazard ratios respond to the precise choice of risk factors available to the researcher.

In Tables 2.7 and 2.8 we detail the individual hazard ratios for the 15 consultants under consideration together with their associated rank orders. This is presented adjusted for two slightly different sets of explanatory variables. It can be seen that there is variation in the rank order depending on the particular choice of risk factors. The results are also illustrated graphically in Fig 2.4. Although those surgeons ranked very highly or poorly remain relatively static in the list, the rank orders of intermediate individuals do tend to be re-distributed. As an example the hazard ratio for surgeon 2 falls from 1.04 to 0.84 and what was an 'average' performance becomes much improved. By contrast, for surgeon 13, what was seen as performance around 10% above average is considered to be one around 10% below average when a slightly different model is fitted.

Table 2.8 - Hazard ratios using different explanatory variables - all cases

Surgeon	Hazard Dukes'	Hazard Local Inv	Dukes Rank	Local Inv Rank	Case Numbers
1	1.11	1.01	12	8	171
2	1.20	1.03	14	11	111
3	1.03	0.95	9	6	86
4	0.91	1.01	6	9	83
5	1.10	1.20	10	15	78
6	0.88	1.00	5	7	75
7	1.30	1.06	15	12	61
8	0.76	0.91	3	5	58
9	1.10	1.18	11	14	52
10	1.18	1.16	13	13	50
11	0.83	0.89	4	4	46
12	0.76	0.67	2	1	42
13	1.03	1.03	8	10	39
14	0.92	0.84	7	2	38
15	0.70	0.85	1	3	36

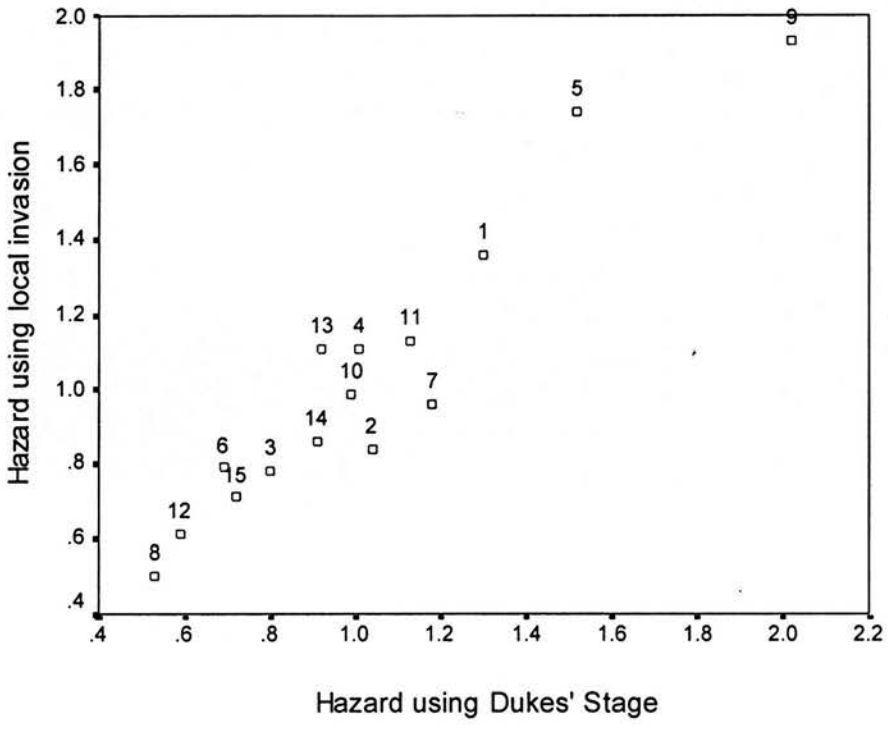
Table 2.9 - Hazard ratios using different explanatory variables - curative resections

Surgeon	Hazard Dukes'	Hazard Local Inv	Dukes Rank	Local Inv Rank	Case Numbers
1	1.30	1.36	13	13	85
2	1.04	0.84	10	6	60
3	0.80	0.78	5	4	36
4	1.01	1.11	9	11	48
5	1.52	1.74	14	14	38
6	0.69	0.79	3	5	41
7	1.18	0.96	12	8	29
8	0.53	0.50	1	1	31
9	2.02	1.93	15	15	34
10	0.99	0.99	8	9	19
11	1.13	1.13	11	12	30
12	0.59	0.61	2	2	21
13	0.92	1.11	7	11	21
14	0.91	0.86	6	7	22
15	0.72	0.71	4	3	19

Figure 2.4

A plot of the hazard ratios derived using different explanatory variables

(curative resections)



2.12 A digression - A brief examination of complications and recurrence data

The data recorded on complications and recurrence are also of interest to the audit specialist. We may be interested in determining what factors contribute to early complications and whether such complications influence long term survival. If complications do influence long term survival then these short-term complications can be used as potential 'surrogate' outcome indicators in an audit study. Given the need for early identification of extreme performance this may well be helpful.

We therefore examined how complication rates varied between surgeons and compared the short and long term rankings of the individuals involved. We also looked at the variation in time to recurrence between surgeons and that of recurrence to death. Finally we considered whether performance measures based on survival data truncated at the two year point differed substantially from the measures using the entire period of follow up.

The main complications identified were wound infection, leakage and 30 day deaths. We examined the complications from various subsets of the data, curative resections, all cases and emergency cases and investigations were then made with and without adjustments for case mix. Investigations were made to determine the explanatory variables of interest and to compare these with those found to be a useful part of the regression model in the earlier study.

An aggressive surgeon might produce more leakage or wound infection but achieve greater tumour clearance enhancing long term survival. It is not clear from prior considerations that the correlation of outcomes, short and long term, will necessarily be positive. Alternatively these complications might be indicative of a generally poor overall surgical performance. We examined the rank orders of the surgeons when one considers their relative performance based on shorter term outcome measures and compared these with the longer term rankings from the other reported study.

We proceeded to fit regression models to see whether the addition of a short term complication variable (e.g. 'wound infection' or 'leak') would contribute significantly to explaining the variation observed in the long term survival experience of patients. They did not add a significant component to the regression model either in addition to a fully fitted model (as one might expect) or when fitted in isolation. This suggested that observing short term complications does not add much to our ability to predict outcomes for the patient either in a positive or negative way. This inference is however qualified by concerns over the much less reliable data quality for recurrence and complication information relative to more traditional measures of performance.

It was noted that the use of case mix adjustments made little difference to the rankings of surgeons when rates of wound infection were analysed relative to the ranks based on unadjusted complication rates. However, when 30 day deaths were analysed the case mix adjusted surgeon rankings, not surprisingly, varied substantially from their unadjusted counterparts. This is evident when the basic summary data are examined e.g. 30 day deaths are 2.5 times more likely for emergency admissions than for elective admissions and similar large effects are seen for other variables.

We examined the relationship between short term surgeon rankings (based on the complications data) and the long term rankings derived earlier. We observed that when the short term measure was wound infection there was no statistically significant correlation between the rankings. The reduction in data numbers and data quality for complications information does however mean that only the most general inferences can be drawn from these tests.

When the times from operation until recurrence were examined the differences between surgeons became, as expected, highly evident and although the data numbers fell (as there were many missing recurrence times) the surgeons previously identified as being at the top and bottom were still identified.

The survival experience for the patient after recurrence displays little variability between surgeons. This is intuitively sensible given the uniformly poor prospects for patients after recurrence has been identified. The rankings based on the analysis of 'survival' to recurrence were significantly correlated with the rankings from the longer term study, a further indication that recurrence could be used as a possible surrogate outcome indicator in audit studies.

Finally, we moved on to an analysis of an additional 'short term' indicator which was constructed from the existing data set. We chose to look at survival up to two years as the indicator of interest but other periods could of course be used. The period has some appeal since many studies do not benefit from the long period of follow up and case acquisition that we see in the Glasgow Royal Infirmary data. In addition for the whole process to be worthwhile the audit results need to be communicated fairly quickly to surgeons.

The survival time from the original data was truncated at 2 years and an appropriate outcome measure variable was constructed. We analysed both all cases/all causes and curative resections/cancer related deaths plotting the appropriate survival curves and calculating log rank statistics and hazard ratios. We confirmed that the selection of covariates used previously also applied to this revised study. The rankings are highly significantly positively correlated ($p=.003$ for the curative resection/cancer related deaths group). A high level of correlation can be expected not just because the first two years experience is part of the entire experience previously studied but because it contains a significant proportion of the actual events of interest i.e. deaths from a cancer related cause.

2.13 Discussion

The ability and experience of the surgeon is a contributory factor when looking at the analysis of long term survival data in a specialised discipline such as treatment of colorectal cancer. It would be preferable if the standard measures of relative surgical performance were robust with respect to modest alterations in the regression model specification (say, because one study might collect different data from another). In addition if the existing analytical methods being used are to give reliable estimates of relative performance then a surgeon's hazard ratio might reasonably be expected to be stable when looking at different groups of patients within his or her overall case load. We have considered both of these questions in this chapter.

The level of insight gained in any audit investigation is arguably greatest when one examines well defined, homogeneous subgroups of patients, although there is an inevitable trade off between the acquisition of a homogeneous group for analysis and the reduced sample size that is the inevitable consequence of any restricted study. The choice of subgroup will of course be influenced by the objectives of the study itself. In this thesis we are interested in examining the power of existing statistical methods to quantify reliably the difference between the performance of individuals involved in surgery, and some natural subgroups can be identified. Initially one might wish to analyse all cases as a starting point to provide a basic frame of reference.

A group of cases which is generally considered to be of interest would be the set of all patients undergoing a curative procedure where the outcome of interest is death from a cancer related cause. This should give an insight into the ability of the surgeon to prolong life when it was assessed at the time of the procedure that this was an achievable objective. A second group of cases which might be useful in a surgical audit investigation would be the set of all emergency admissions with the outcome of interest most likely to be death by any cause. The rationale behind this second choice of subgroup is that patients entering hospital as an emergency would do so without the surgeon having been pre-selected, a possible source of bias in the audit investigation.

We have attempted to show that the precise choice of risk factors does indeed matter in surgical audit investigations. Not all of the movement in hazard ratios and rank orders can be ascribed to small data numbers. The surgeons at the Glasgow Royal Infirmary were coded according to descending numbers of procedures and so whilst it is perhaps not surprising that consultant number 13 moves by 4 places (since he only performed 39 operations of which 21 were deemed to be curative) it is more surprising that consultant number 2 also moved by 4 places (since he performed 111 operations of which 59 were deemed to be curative).

Whilst most colorectal cancer surgeons perform well (and some perform very well) there are practitioners who clearly display disappointing rates of short-term complications and poor long-term outcomes. Of course one has to be careful not to assume that short-term complications are necessarily associated with poor long-term outcomes. It is possible that an adventurous surgeon might create short-term problems for a patient whilst improving his or her long-term outcome as a result of having performed more aggressive clearance of the tumour. The results of past analyses of colorectal cancer data have led researchers to suggest that colorectal cancer surgery should be concentrated within teams in specialist centres where appropriate levels of training and experience are in evidence.

Further, by analysing the cases of surgeons in two parts we have also demonstrated that, even with a carefully selected homogeneous group of cases using an appropriate outcome measure, measured performance of individual surgeons can vary substantially within their own case load. The stability in measured surgical performance which we would have ideally liked to have seen was not evident. Attention must continue to be focused on not just measures of relative risk but on the associated confidence intervals around these point estimates. The confidence intervals for relative performance measures do of course indicate the lack of precision of the estimated relative hazards themselves but the split sample technique illustrates quite graphically exactly what this lack of precision means in practice.

The data on colorectal cancer patients attending Glasgow Royal Infirmary are comprehensive in their coverage of a range of potential explanatory variables and importantly there is a long period of case acquisition and follow up. What this Chapter has attempted to show is that whilst differences in case mix adjusted surgical performance do indeed exist care must continue to be exercised in drawing definite conclusions from any particular study. The hazard ratio measures for individual surgeons vary depending on the particular choice of risk factors and display variability within the caseload even for those individuals carrying out a large number of procedures.

Low case numbers per surgeon will continue to place a constraint on the power of available statistical methods to highlight differences in performance between surgeons in specialised surgical disciplines. This is of particular concern when the analysis is confined to a subset of the data, such as curative procedures or emergency admissions.

We have highlighted several concerns which should then be borne in mind when interpreting the results of surgical audit studies. The established statistical techniques used in audit investigations are powerful in their ability to assess the influence of various prognostic factors influencing surgical outcomes and in the determination of the relative merits of different aspects of surgical practice as a whole. They are also a convenient method for calculating hazard ratio measures for surgeons and efficiently illustrate the divergence which is evident in surgical performance particularly highlighting those practitioners who are exceptionally good or very poor.

However, the lack of stability of calculated hazard ratios within the case load or when using different mixes of explanatory variables (combined with the large confidence intervals inevitably associated with an inherently variable process with small case numbers) do mean that considerable care must be taken in drawing definite conclusions about the relative merits of the majority of specialist surgeons. In particular our case studies highlight the extreme sensitivity of rank orders, particularly for middle ranking surgeons.

The power of available techniques would be seen to be greater when confined to disciplines where a greater volume of cases per surgeon are being handled, for instance in cardiac or general surgery.

Where surgical audit measures are required for specialised surgery departments such as colorectal cancer great care should be taken in the interpretation of studies with small data numbers and increased effort is required in the development of, probably more highly parameterised, models which might make better use of what information is available although at the expense of simplicity.

We concluded that inferences from shorter term periods, say 2 years, bring out broadly similar conclusions to studies with a much longer period of follow up.

Chapter 3

An Audit of Surgical Performance Across Several Hospitals

3.1 Introduction

The previous chapter explored a set of data which enabled one to review the extent of surgeon related variability in outcomes following surgery for colorectal cancer within the context of a single hospital. The data were in the form of admissions to Glasgow Royal Infirmary where the patient had colorectal cancer of varying degrees of severity. The period of case acquisition and follow up was very long, between 1974 and 1984.

In this chapter we will study, once again, the outcomes of patients undergoing surgery for colorectal cancer but within a different setting. The main structural differences are as follows.

- The data are sourced from a number of hospitals.
- The period of case acquisition and follow up is much shorter. Indeed further than this, all data were treated as being censored at the two year point.

We were provided with a file of data derived from a study funded by the Clinical Resource and Audit Group (2000) which covered 2090 cases from 8 hospitals with (47 surgeons) in the West of Scotland. This was the principal source of data used in our analysis. A sample of the form used to collect the data is included as Appendix 1. Our focus in this thesis is on the statistical issues involved in surgical audit but in passing it is worthy of comment that a data acquisition project such as this is a very demanding exercise. It demands significant allocations of both time and resources from practitioners, senior and junior, across a range of disciplines in the National Health Service.

The information collected is substantial and is very similar to that discussed in Chapter 2, Section 2.5. It also contains supplementary information on the site of the tumour, repeated operations and any adjuvant therapy provided to patients. The usual pathology data were recorded. The follow up information is comprehensive and supplementary details are recorded as to recurrence, secondary operative procedures and presence of disease at death.

Another source of information made available to us was the South East of Scotland CRAG Colorectal data for a similar period. There were 1687 cases in the East of Scotland file. In many respects this was a comparable set of data to that provided for the West of Scotland this second source was mainly used to validate some of the assumptions made regarding the effects of the main covariates of interest which we proceed in Chapter 4 to use as a basis for more wide ranging studies of statistical models used to analyse surgical audit problems.

The two sets of data did differ in certain respects. In particular the East of Scotland source contained only operated cases. Disease measures were therefore likely to be more advanced in the West of Scotland including, as the data did, unoperated cases as well as cases where the patient underwent surgery. When validating the West of Scotland data adjustments were made to the files to ensure correspondence between the two sources of information. The East of Scotland data included additional comprehensive information on the seniority of the surgeon and any assisting staff.

Overall however even after making appropriate adjustments to the files we could not be satisfied as to the basic comparability of the two sources. Had this been the case the files could have been merged enabling additional statistical issues to be explored.

3.2 The specific objectives of our case study

Looking forward we approached the data analysis with a number of issues which we wished to consider. These included the following.

- An initial descriptive analysis of case number and case mix variability between hospitals and surgeons.
- An assessment of missing data problems and a determination of the covariates that should be included in models used to describe case mix variability.
- A comparison of inferences made on relative performance using different models, specifically Cox Regression and Logistic Regression.
- An examination of the extent of variation in actual and predicted mortality rates.
- A comparison of inferences made on relative performance using fixed or random effects (multilevel) models.
- A study of the statistical distributions of rank orders for the surgeons in the study.

The underlying objective was to develop the theme addressed in Chapter 2 of assessing the reliability of inferences made on surgical performance. The additional knowledge gained from this further case study was to be used eventually as the basis for a simulation exercise (see Chapter 4).

3.3 Initial Exploratory Analyses

The data, whilst very comprehensive, were not without problems. Text files were imported into the software packages Minitab and SPSS for analysis. The routine importing of the data had altered certain dates into the incorrect century but these were obvious by inspection and adjusted as appropriate.

Additional fields were created for the following variables and identifiers:-

- Fields to identify whether death occurred within a specific time (sometimes there was conflicting information on file from differing sources). To enable a full analysis and comparison of different outcome measures we recorded identifiers for death within 6 months, 1 year or 2 years.
- Fields to identify if a case was lost to follow up
- Binary fields to identify particular surgeons or hospitals.
- Censoring fields to limit survival times to 2 years.

A decision was made to censor all data at 2 years. The basic reasoning behind this decision was that the follow up of patients was not systematic beyond this point. Specifically there were concerns about the reliability of follow up information beyond the two year point as individuals relocated to other geographical areas or changed practitioners. The information on follow up, whilst very comprehensive, clearly becomes less reliable as time passes. An example of the impact of this feature of the data might be a case where an individual is lost to follow up but this fact is not recorded on the audit file. He subsequently dies but this information is lost to the statistician. Mortality rates are underestimated as a consequence.

It is to be recalled however that the analysis of data on colorectal cancer surgery from a single institution which was reported in Chapter 2 showed that audit inferences from data truncated at two years were very similar to those obtained using much longer periods of follow up. Conditional on the fact that the patient does in fact die, the vast majority of times of death are recorded within the two year period. This encouraged us to believe that any inferences we might make on relative performance would not be invalidated by our decision to censor all survival times at 2 years.

The particular set of data also then enabled us to make comparisons of the effectiveness of analyses based on survival times (e.g. Cox regression) with analyses based on binary outcomes (e.g. logistic regression).

3.4 The numbers and types of cases being analysed

As discussed in Chapter 2, Section 2.13 it is of primary importance to analyse audit data taking into account both differences in case mix and case numbers. The confidence attaching to performance measures is highly influenced by the size of the population of patients per surgeon (or hospital) relative to the peer group of interest.

The following table demonstrates the variability of case numbers by hospital and by surgeon. Category 999 is for cases where the surgeon was not recorded and all hospitals have been allocated anonymous codes, A to H.

Table 3.1 : CRAG West of Scotland Data - Case numbers for surgeons in the study

	Hospital							
Surgeon	A	B	C	D	E	F	G	H
1		42						
2		43						
3		9						
4		15						
5		107						
6		27						
7		42						
8		7						
9			15					
10			35					
11			14					
12			9					
13				40				
14				58				
15				43				
16				61				
17				24				
18					97			
19					48			
20					70			
21					107			
22						11		
23						2		
25						22		
26							77	
27							36	
28							35	
29							25	
30							37	1
31								89
32								89
33								45
34								59
35								73
36								51
37								54
38								102
39								36
40								31
41								28
42				26				
43								39
44								32 1
45		41						
46					27			
47		2						
999		19	20	12	15		4	16 44

The most immediately apparent feature of this table is that there is quite pronounced variability in patient numbers between practitioners and institutions (and also between practitioners within institutions). There is also very little cross-classification in the data, that is to say a surgeon who has operated on patients in two or more different hospitals within the period under consideration. This latter point is of importance when one considers the advantages and limitations of using the more advanced random effects/multilevel statistical approaches discussed in Chapter 1, Section 1.10. The low level of cross classification results in the main from the relatively short period of case acquisition in this particular study.

The data on the East of Scotland (not tabulated) were, as discussed earlier, used to validate the choices we made about the variables to use in the statistical models to adjust for case mix and to confirm the essential reasonableness of the proportion of cases which fall into various homogenous groups of patients. In short we wished to be able to infer that the West of Scotland data were typical of a cross section of a population of patients undergoing surgery for colorectal cancer. This is of particular importance when one uses the data as a basis for theoretical statistical work (as will be discussed in Chapter 4). As a point of interest for further research the East of Scotland data exhibited greater cross-classification than did the West of Scotland (where the surgeons were effectively nested within hospitals). It would be of interest to study the impact this has the whole modelling process. In effect cross-classification enables one to have a greater understanding of the components of variation in a hierarchical statistical model. The statistician has the ability to 'explain' more variation as opposed to merely partitioning the variability in a different way between the hierarchies.

Table 3.2 shows the aggregate numbers of cases at the hospital level. Hospital E hospital clearly stands out as being different from the others, both in numbers of cases and the low level of deaths.

Table 3.2

CRAG West of Scotland Data - Cases and Deaths by Hospital

Hospital	Cases	Deaths	Mortality rate
A	354	117	33.1%
B	95	37	38.9%
C	263	101	38.4%
D	363	145	39.9%
E	35	2	5.7%
F	212	85	40.1%
G	442	140	31.7%
H	326	124	38.0%
Overall	2090	751	35.9%

Table 3.3 shows the variation in curative and palliative procedures between hospitals (ranging from 68.9% to 85.5% (curative)). The variation in survival percentages is pronounced.

Table 3.3

Variation in Curative and Palliative Procedures by Hospital

	Curative			Palliative		
	Count	%	2 year survival %	Count	%	2 year survival %
A	201	72	83.6	78	28	59.0
B	71	85.5	77.5	12	14.5	25.0
C	174	79.5	72.5	45	20.5	60.0
D	214	71.6	75.7	84	28.1	47.6
E	27	81.8	92.6	6	18.2	100.0
F	115	68.9	81.7	52	31.1	46.2
G	273	73.8	80.6	97	26.2	59.8
H	191	72.6	78.9	72	27.4	50.0

Table 3.4 shows the variation in the categories of the presentation variable, which was collapsed into a binary variable denoting elective/emergency (previously having had 3 emergency classifications; obstruction, perforation and 'other'). Although not explicitly reported the survival percentages range from 67.1% to 75.1% for elective surgery (excluding hospital E which seemed outlying at 96.2%). The survival percentages range from 41.0% to 60.5% for emergency surgery (excluding hospital E which seemed outlying at 87.5%).

Table 3.4

Variation in Elective and Emergency Cases by Hospital

Hospital	Elective		Emergency (total)	
	Count	%	Count	%
A	233	65.8	121	34.2
B	55	57.9	40	42.1
C	173	66.0	89	34.0
D	234	64.5	129	35.5
E	26	76.5	8	23.5
F	141	65.9	73	34.1
G	280	64.1	157	35.9
H	219	67.2	107	32.8

Table 3.5 below illustrates the variation in resection percentages. Although not explicitly reported the survival percentages range from 0% to 29.2% for surgery where there was no resection (excluding hospital E). The survival percentages range from 67.8% to 76.2% for resections (excluding hospital E).

Table 3.5 – Variation in resection percentages by Hospital

Hospital	No resection		Resection	
	Count	%	Count	%
A	24	7.9	281	92.1
B	9	9.8	83	80.2
C	22	9.2	217	90.8
D	14	4.5	298	95.5
E	2	5.7	33	94.3
F	16	8.5	173	91.5
G	21	5.4	369	94.6
H	18	6.4	265	93.6

Finally Table 3.6 shows the distribution of cases over Dukes' Stage and Table 3.7 shows the corresponding survival percentages which again exhibit considerable variation

Table 3.6

The distribution of cases over Dukes' Stage by Hospital

Hospital	A		B		C		D	
	Count	%	Count	%	Count	%	Count	%
A	14	4.7	138	46	80	27	68	23
B	3	3.9	42	55	18	23	14	18
C	5	2.1	117	50	58	25	53	23
D	7	2.2	153	49	82	26	71	23
F	6	3.5	77	45	46	27	44	25
G	23	6.0	161	42	115	30	86	22
H	8	2.9	110	37	87	31	73	26

Table 3.7

The distribution of 2 year survival percentages corresponding to Table 3.6

Hospital	A	B	C	D
A	92.9	84.8	68.8	48.5
B	100.0	83.3	50.0	28.6
C	100.0	76.1	70.7	34.0
D	85.7	82.4	53.0	36.6
F	100.0	84.4	67.4	31.8
G	91.3	80.7	74.8	48.8
H	87.5	82.6	72.4	32.9

Hospital E is excluded from the above table (see later). Considering all of the tables above we can see that there are differences in proportions of cases falling into many broad classifications of clinical and audit interest. The survival rates are also highly variable within the various classifications. This is demonstrably evident at the hospital level and as expected the variability in case mix is even more pronounced at the surgeon level. As we saw with the earlier set of data from the Glasgow Royal Infirmary there appears to be a clear requirement to analyse surgical audit data after having adjusted for case mix. This of course is a statement made within the context of this particular study of colorectal cancer patients, although the inference is capable of generalisation to other surgical disciplines.

3.5 Missing Data – A Reduction in Case Numbers

The next stage was to consider the extent of missing data in the file. An analysis of complete cases alone discards much useful information, reducing statistical power, but excessive imputation is also not without risks, particularly in an audit study where the last thing the scientist wishes to do is incorrectly estimate measures of variability.

In the main the records were fairly complete and, as in Chapter 2, imputation was often possible by inferring values of particular missing variables from other fields in the records. Stepwise regression methods selected the same subset of variables as discussed in Chapter 2, Section 2.5 and this variable selection was validated by considering the East of Scotland data in addition.

Deleting unoperated cases reduced the file size to 1852 cases. The variables used to allow for case mix were as follows:-

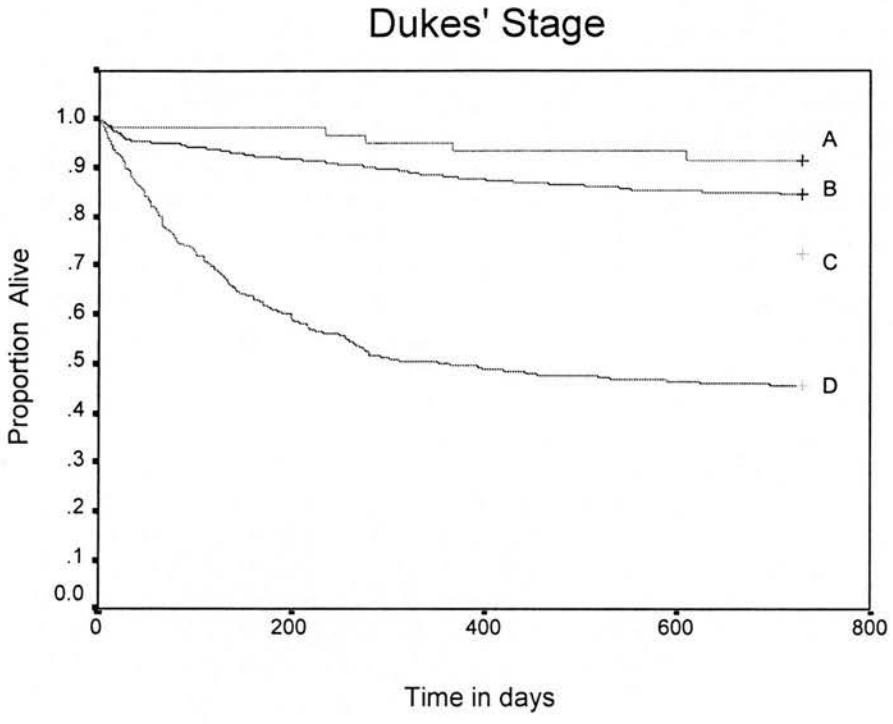
- Age (a binary variable, over or under 65)
- Sex (a binary variable, male or female)
- Presentation (a binary variable, elective or emergency)
- Procedure (curative resection, palliative resection, other)
- Dukes' Stage (A,B,C,D an ordered categorical variable)

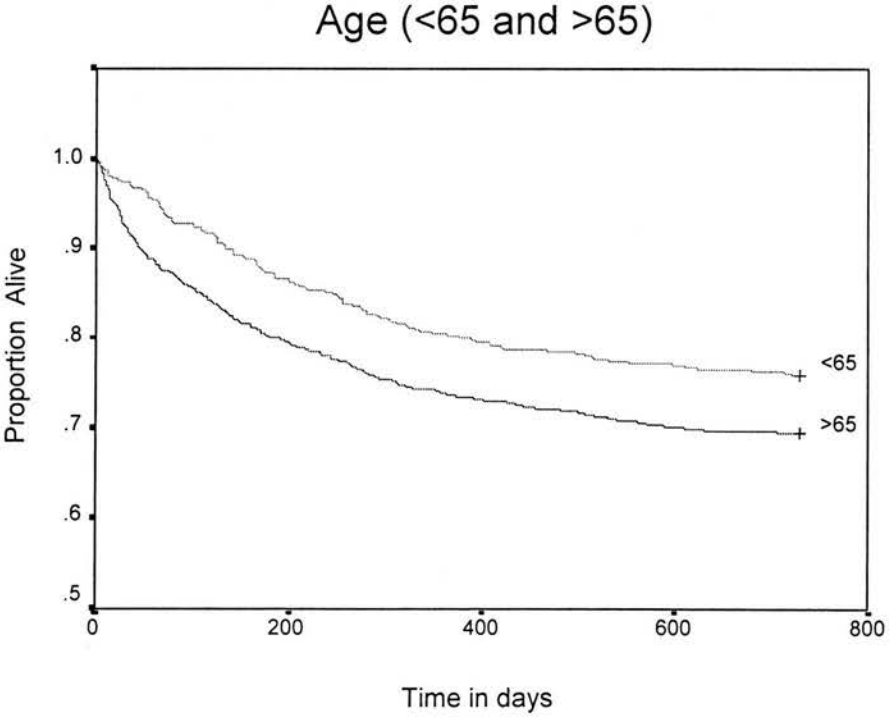
To enable other issues to be explored in due course we also recorded the status of the surgeon and assistant and the site of the tumour. This enabled a preliminary examination of whether seniority is a predictor of performance and whether certain subsets of procedures (e.g. rectal surgery) offer better discrimination between surgeons (for a given amount of data).

The following graphs illustrate the effects of the main covariates in the model for case mix.

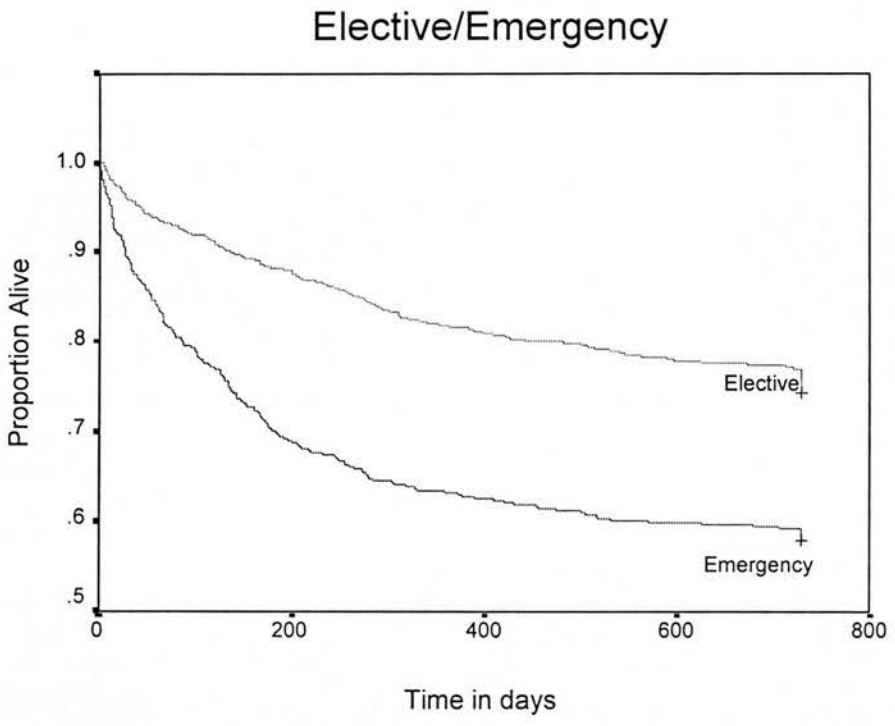
Figures 3.1 , 3.2 and 3.3

Kaplan-Meier Curves for 3 Covariates (Dukes' Stage, Age and Presentation)





Figures 3.1 , 3.2 and 3.3 (continued)



The pronounced case mix effects in evidence provide strong support to the contention that surgical audit data should be analysed using methods which capture the broad effects which are present. If this is not possible analyses should be carried out on smaller homogeneous groups of patients, albeit with a consequent reduction in statistical power. As a footnote we confirmed the continuing validity of the use of a binary variable for age. A stratification of all ages into 10-year bands enabled graphical comparison of survival curves. This suggested a split at around 70 years but 65 was used for consistency with the analysis in Chapter 2 and was also supported from wider consideration of the East of Scotland data.

3.6 Initial estimates of relative risks for surgeons and hospitals

Using the SPSS statistics software package we calculated the relative risks for a selection of surgeons and hospitals to gain additional insight into the fundamental variability present in these data. The addition of a binary covariate to the linear component of the regression model specification identifying the particular surgeon or hospital of interest (in addition to the other covariates allowing for case mix, where appropriate) enabled efficient computation of the relative risk (or hazard for the Cox proportional hazard models).

The following tables illustrate a small sample of the first estimates of hospital and surgeon level relative hazards ratios also detailing the relative performance of a surgeon within his own institution and also against all surgeons in the entire study. The upper and lower limits are for the 95% confidence interval.

Table 3.8

Initial estimates of hospital performance (without case mix adjustments)

Hospital	Hazard Ratio	Lower 2.5% Limit	Upper 2.5% Limit
A	0.86	0.68	1.09
B	1.32	0.93	1.87
C	1.12	0.89	1.42
D	1.13	0.92	1.40
F	1.12	0.87	1.46
G	0.83	0.67	1.02
H	1.06	0.84	1.32

Table 3.9

Initial estimates of hospital performance (with case mix adjustments)

Hospital	Hazard Ratio	Lower 2.5% Limit	Upper 2.5% Limit
A	0.79	0.61	1.01
B	1.37	0.91	2.06
C	1.12	0.87	1.44
D	1.23	0.98	1.54
F	1.02	0.77	1.37
G	0.79	0.63	0.99
H	1.06	0.83	1.35

Table 3.10

Estimates of surgeon performance without case mix adjustment – within Hospital A

	Ratio	Limit	Limit
1	1.13	0.61	2.09
2	0.84	0.43	1.64
3	2.13	0.78	5.83
4	0.81	0.25	2.57
5	1.03	0.65	1.63
6	0.84	0.36	1.93
7	1.37	0.76	2.49
8	4.06	1.64	10.00
45	0.37	0.15	0.93
47	insufficient patients		

Table 3.11

Estimates of surgeon performance with case mix adjustments – within Hospital A

Surgeon	Hazard Ratio	Lower 2.5% Limit	Upper 2.5% Limit
1	1.23	0.65	2.32
2	0.84	0.40	1.78
3	1.03	0.24	4.30
4	1.59	0.48	5.30
5	1.11	0.65	1.90
6	0.90	0.38	2.00
7	0.94	0.44	2.00
8	2.80	1.04	7.53
45	0.43	0.15	1.21

Table 3.12

Estimates of Hospital A surgeon performance, with case mix adjustments

(with reference to all surgeons in the study)

Surgeon	Hazard Ratio	Lower 2.5% Limit	Upper 2.5% Limit
1	0.92	0.51	1.63
2	0.69	0.34	1.40
3	0.92	0.22	3.71
4	1.13	0.36	3.54
5	0.94	0.62	1.44
6	0.68	0.30	1.52
7	0.74	0.36	1.50
8	2.21	0.90	5.39
45	0.38	0.14	1.02

Table 3.13

Estimates of surgeon performance with case mix adjustments – within Hospital D

Surgeon	Hazard Ratio	Lower 2.5% Limit	Upper 2.5% Limit
18	1.58	1.02	2.43
19	1.75	1.01	3.01
20	1.02	0.61	1.69
21	0.69	0.43	1.12
46	0.21	0.06	0.67

Table 3.14

Estimates of Hospital D surgeon performance, with case mix adjustments

(with reference to all surgeons in the study)

Surgeon	Hazard Ratio	Lower 2.5% Limit	Upper 2.5% Limit
18	1.72	1.20	2.46
19	1.90	1.15	3.14
20	1.22	0.78	1.91
21	0.86	0.57	1.31
46	0.32	0.10	1.02

At the hospital level given the large samples involved we see little movement in the hazard ratios after adjusting for case mix as the larger numbers of cases at the hospital level leads to less variability in case mix. Again hospital E is deleted from this reporting of relative performance as certain issues regarding this centre subsequently came to light.

At the surgeon level the sample tables for just two hospitals show some features of interest. When considering the results for the surgeons in hospital A we see the strong influence of case mix. The relative hazard (within the hospital) of surgeon 3 falls from 2.13 to 1.03 after we allow for case mix in the modelling process. Within the hospital surgeon 8 is apparently a poor performer with a confidence interval for the hazard ratio that excludes unity. When measured against all surgeons however a significant result is not observed. When considering the results for hospital D other features of the data are apparent. Within a single institution with 5 surgeons we obtain significant divergence from average with surgeons 18 and 19 apparently performing poorly counterbalanced by a better performance by surgeon 46.

Overall this brief extract of the many tables produced demonstrates some of the sizes of effects which are under consideration and, perhaps more interestingly, the size of confidence intervals around the point estimates of relative performance. Of course, as will be discussed later in the Chapter, the complex structure of the data is not properly allowed for in such a basic analysis but preliminary analyses such as these do give the audit specialist a basis from which to consider more sophisticated models. As discussed in Chapter 1, Section 1.10 the main weakness of any single level analysis is that by comparing one surgeon or institution against their peer group it pools the characteristics of the peer group, essentially treating them as a whole. It fails to directly address the variation within or between hospitals.

3.7 *An initial comparison of results using alternative outcome measures*

Limiting the number of surgeons to those with 20 or more cases we produced the following tables of relative hazards. The logistic regression procedures excluded those cases lost to follow up over the period under consideration.

- Logistic regression on 6 month outcomes with and without case mix adjustment.
- Logistic regression on 1 year outcomes with and without case mix adjustment.
- Logistic regression on 2 year outcomes with and without case mix adjustment.
- Cox regression on survival times censored at the two year point.

Table 3.15

Relative risk for surgeons and associated 95% confidence interval

(Logistic regression 6 month outcomes – no case mix adjustments)

Surgeon	exp(beta)	lower 2.5%	upper 2.5%
1	1.031	0.451	2.359
2	0.519	0.184	1.468
5	0.677	0.365	1.257
6	0.896	0.305	2.629
7	1.883	0.927	3.825
10	2.300	1.109	4.771
13	1.222	0.556	2.685
14	0.711	0.318	1.589
15	1.065	0.465	2.442
16	1.083	0.556	2.110
17	2.086	0.851	5.116
18	1.287	0.742	2.233
19	1.798	0.915	3.533
20	1.091	0.575	2.069
21	0.644	0.348	1.194
25	0.222	0.030	1.655
26	0.924	0.491	1.768
27	0.782	0.301	2.032
28	1.664	0.697	3.972
29	0.203	0.027	1.503
30	2.651	1.257	5.593
31	1.499	0.882	2.548
32	1.207	0.687	2.121
33	0.688	0.267	1.774
34	0.595	0.252	1.403
35	0.906	0.469	1.749
36	0.632	0.246	1.619
37	0.521	0.157	1.725
38	0.660	0.356	1.225
39	1.650	0.765	3.558
40	2.793	1.315	5.931
41	0.563	0.169	1.873
42	0.896	0.305	2.629
43	0.398	0.122	1.302
44	1.732	0.764	3.929
45	0.519	0.184	1.468
46	0.640	0.190	2.150

Table 3.16

Relative risk for surgeons and associated 95% confidence interval

(Logistic regression 1 year outcomes – no case mix adjustments)

Surgeon	exp(beta)	lower 2.5%	upper 2.5%
1	0.948	0.446	2.013
2	0.665	0.292	1.515
5	1.089	0.676	1.757
6	0.788	0.294	2.133
7	0.600	0.815	3.143
10	1.526	0.738	3.159
13	0.948	0.446	2.013
14	0.923	0.481	1.774
15	0.981	0.461	2.090
16	0.717	0.369	1.394
17	2.056	0.883	4.785
18	1.211	0.732	2.004
19	2.243	1.217	4.135
20	1.248	0.716	2.178
21	0.749	0.448	1.252
25	0.149	0.200	1.108
26	0.980	0.550	1.668
27	0.936	0.422	2.076
28	2.206	1.016	4.792
29	0.139	0.018	1.006
30	2.661	1.300	5.445
31	1.377	0.842	2.251
32	1.153	0.691	1.923
33	0.948	0.446	2.013
34	0.473	0.212	1.055
35	0.814	0.447	1.480
36	0.520	0.218	1.244
37	0.628	0.239	1.651
38	0.770	0.460	1.288
39	1.462	0.710	3.010
40	2.464	1.187	5.188
41	1.270	0.555	2.906
42	0.788	0.294	2.113
43	0.473	0.184	1.220
44	1.854	0.875	3.928
45	0.347	0.123	0.978
46	0.428	0.128	1.435

Table 3.17

Relative risk for surgeons and associated 95% confidence interval

(Logistic regression 2 year outcomes – no case mix adjustments)

Surgeon	exp(beta)	lower 2.5%	upper 2.5%
1	1.007	0.506	2.004
2	0.751	0.364	1.547
5	0.909	0.574	1.438
6	0.712	0.283	1.794
7	1.137	0.579	2.229
10	1.413	0.702	2.844
13	1.276	0.658	2.476
14	1.388	0.789	2.443
15	0.922	0.454	1.872
16	0.995	0.566	1.748
17	2.100	0.921	4.791
18	1.722	1.092	2.716
19	1.924	1.053	3.514
20	1.195	0.706	2.023
21	0.854	0.542	1.347
25	0.115	0.016	0.798
26	1.189	0.725	1.950
27	0.781	0.363	1.678
28	1.570	0.724	3.406
29	0.593	0.220	1.595
30	1.890	0.925	3.863
31	1.163	0.723	1.870
32	1.118	0.691	1.808
33	0.888	0.439	1.797
34	0.454	0.220	0.937
35	0.747	0.427	1.307
36	0.703	0.343	1.440
37	0.562	0.228	1.382
38	0.996	0.635	1.560
39	1.187	0.086	2.403
40	2.640	1.279	5.449
41	1.264	0.580	2.757
42	0.712	0.283	1.794
43	0.337	0.131	0.869
44	1.319	0.623	2.791
45	0.318	0.124	0.815
46	0.306	0.911	1.025

Table 3.18

Relative risk for surgeons and associated 95% confidence interval

(Cox regression 2 year censoring – no case mix adjustments)

Surgeon	exp(beta)	lower 2.5%	upper 2.5%
1	1.004	0.567	1.780
2	0.772	0.413	1.443
5	0.920	0.624	1.354
6	0.762	0.341	1.706
7	1.186	0.686	2.062
10	1.427	0.823	2.474
13	1.196	0.703	2.033
14	1.210	0.774	1.892
15	0.931	0.512	1.691
16	0.997	0.623	1.595
17	1.720	0.947	3.126
18	1.488	1.051	2.106
19	1.969	1.132	2.766
20	1.161	0.758	1.779
21	0.861	0.585	1.269
25	0.129	0.018	0.916
26	1.121	0.750	1.675
27	0.815	0.422	1.575
28	1.518	0.835	2.755
29	0.595	0.247	1.436
30	1.771	1.042	3.013
31	1.193	0.810	1.757
32	1.087	0.733	1.613
33	0.893	0.491	1.622
34	0.450	0.224	0.904
35	0.787	0.486	1.277
36	0.738	0.394	1.379
37	0.621	0.278	1.389
38	0.964	0.663	1.401
39	1.156	0.652	2.049
40	2.142	1.302	3.023
41	1.146	0.613	2.143
42	0.787	0.352	1.761
43	0.385	0.160	0.928
44	1.328	0.731	2.414
45	0.365	0.151	0.880
46	0.352	0.113	1.095

Table 3.19

Relative risk for surgeons and associated 95% confidence interval

(Logistic regression 6 month outcomes – with case mix adjustments)

Surgeon	exp(beta)	lower 2.5%	upper 2,5%
1	0.931	0.406	2.323
2	0.569	0.192	1.686
5	0.841	0.437	1.619
6	0.657	0.211	2.044
7	1.537	0.705	3.352
10	1.895	0.853	4.212
13	1.369	0.586	3.200
14	0.748	0.321	1.745
15	1.060	0.436	2.576
16	1.376	0.677	2.798
17	1.766	0.660	4.725
18	1.587	0.880	2.863
19	2.308	1.117	4.770
20	1.246	0.625	2.486
21	0.626	0.328	1.198
25	0.074	0.010	0.556
26	0.855	0.435	1.681
27	0.845	0.308	2.319
28	1.388	0.543	3.546
29	0.138	0.018	1.065
30	2.179	0.965	4.922
31	1.960	1.103	3.484
32	1.226	0.671	2.237
33	0.732	0.269	1.990
34	0.643	0.262	1.581
35	0.994	0.494	2.005
36	0.868	0.211	1.530
37	0.467	0.132	1.652
38	0.670	0.350	1.281
39	1.336	0.681	3.044
40	2.548	1.127	5.761
41	0.618	0.175	2.183
42	0.695	0.221	2.180
43	0.425	0.125	1.448
44	1.562	0.635	3.841
45	0.622	0.208	1.855
46	0.603	0.172	2.109

Table 3.20

Relative risk for surgeons and associated 95% confidence interval

(Logistic regression 1 year outcomes – with case mix adjustments)

Surgeon	exp(beta)	lower 2.5%	upper 2,5%
1	0.851	0.382	1.896
2	0.735	0.307	1.762
5	1.431	0.854	2.397
6	0.566	0.199	1.615
7	1.310	0.622	2.761
10	1.187	0.537	2.623
13	1.039	0.462	2.339
14	0.983	0.489	1.978
15	0.969	0.429	2.190
16	0.857	0.424	1.731
17	1.804	0.709	4.589
18	1.462	0.850	2.512
19	2.983	1.543	5.765
20	1.486	0.810	2.725
21	0.738	0.426	1.276
25	0.047	0.006	0.349
26	0.911	0.500	1.664
27	1.039	0.442	2.442
28	1.992	0.852	4.660
29	0.089	0.117	0.681
30	2.170	0.989	4.759
31	1.765	1.037	3.006
32	1.166	0.671	2.023
33	1.039	0.462	2.337
34	0.486	0.210	1.129
35	0.877	0.463	1.664
36	0.462	0.184	1.159
37	0.581	0.205	1.647
38	0.785	0.454	1.357
39	1.172	0.538	2.551
40	2.224	1.007	4.910
41	1.557	0.638	3.802
42	0.598	0.208	1.702
43	0.488	0.180	1.317
44	1.673	0.726	3.853
45	0.391	0.132	1.161
46	0.385	0.110	1.345

Table 3.21

Relative risk for surgeons and associated 95% confidence interval

(Logistic regression 2 year outcomes – with case mix adjustments)

Surgeon	exp(beta)	lower 2.5%	upper 2,5%
1	0.921	0.446	1.900
2	0.825	0.385	1.765
5	1.108	0.681	1.801
6	0.538	0.204	1.420
7	0.925	0.449	1.908
10	1.142	0.540	2.418
13	1.422	0.704	2.872
14	1.531	0.840	2.388
15	0.916	0.432	1.943
16	1.173	0.648	2.124
17	1.911	0.786	4.649
18	2.083	1.283	3.381
19	2.344	1.241	4.430
20	1.380	0.787	2.419
21	0.859	0.530	1.391
25	0.038	0.005	0.284
26	1.177	0.694	1.997
27	0.827	0.369	1.855
28	1.379	0.602	3.158
29	0.474	0.166	1.355
30	1.509	0.702	3.242
31	1.370	0.826	2.271
32	1.125	0.673	1.875
33	0.941	0.446	1.982
34	0.455	0.215	0.966
35	0.785	0.436	1.413
36	0.653	0.307	1.390
37	0.539	0.207	1.404
38	1.018	0.634	1.636
39	0.948	0.450	1.998
40	2.385	1.107	5.136
41	1.466	0.642	3.349
42	0.568	0.214	1.510
43	0.338	0.127	0.901
44	1.165	0.518	2.620
45	0.356	0.134	0.947
46	0.263	0.076	0.904

Table 3.22

Relative risk for surgeons and associated 95% confidence interval

(Cox regression 2 year censoring – with case mix adjustments)

Surgeon	exp(beta)	lower 2.5%	upper 2,5%
1	0.937	0.528	1.664
2	0.833	0.446	1.558
5	1.080	0.733	1.592
6	0.651	0.291	1.457
7	1.051	0.606	1.824
10	1.161	0.669	2.015
13	1.340	0.787	2.281
14	1.260	0.806	1.970
15	0.948	0.521	1.722
16	1.115	0.696	1.785
17	1.588	0.873	2.887
18	1.693	1.195	2.394
19	2.235	1.426	3.501
20	1.294	0.844	1.985
21	0.844	0.567	1.230
25	0.056	0.008	0.401
26	1.117	0.747	1.669
27	0.921	0.476	1.780
28	1.334	0.733	2.427
29	0.488	0.202	1.178
30	1.147	0.862	2.498
31	1.362	0.923	2.013
32	1.094	0.738	1.623
33	0.863	0.475	1.568
34	0.476	0.236	0.096
35	0.833	0.514	1.352
36	0.699	0.374	1.309
37	0.035	0.284	1.421
38	0.986	0.678	1.434
39	1.004	0.568	1.780
40	2.120	1.288	3.490
41	1.271	0.679	2.378
42	0.657	0.294	1.471
43	0.396	0.164	0.955
44	1.313	0.721	2.393
45	0.425	0.176	1.025
46	0.310	0.995	0.964

The tables highlighted a number of features of general interest:-

- Those surgeons identified as being significantly different from average at 6 months and 1 year are not necessarily still identified as being significantly different from average at the two-year point. In addition certain surgeons identified as being significantly different from the others at two years duration were 'undetected' at shorter follow up durations.

- The results from Logistic regression and Cox regression using 2 year outcomes (or survival times) are very similar.

- The results allowing for case mix variation can differ from those without such an allowance. We would expect this to be more noticeable at shorter terms and for surgeons or hospitals with lower case numbers.

- The widths of the confidence intervals increase as the power of our analysis falls through the progression of different statistical procedures. As an example the average widths of case mix adjusted 95% confidence interviews were as follows:-

Cox regression	(2 year survival term)	1.127
Logistic regression	(2 year outcomes)	1.569
Logistic regression	(1 year outcomes)	1.811
Logistic regression	(6 month outcomes)	2.011

- Consultant 25 is a clear outlier and so extreme as to alert one to possible data errors.

To investigate the data quality further a table of case mix variables was produced (see below)

Table 3.23

The distribution of case mix between surgeons (by Dukes' Stage and Age)

Surgeon	Dukes' A	Dukes' B	Dukes' C	Dukes' D	Age <65	Age >65
1	0.00	41.00	23.10	35.90	43.60	56.40
2	2.50	50.00	25.00	22.50	37.50	62.50
5	7.40	46.80	27.70	18.10	34.00	66.00
6	4.00	32.00	24.00	40.00	24.00	76.00
7	2.60	35.90	23.10	38.50	28.20	71.80
10	2.90	32.40	23.50	41.20	29.40	70.60
13	0.00	51.30	28.20	20.50	28.20	71.80
14	1.90	47.20	24.50	26.40	37.70	62.30
15	2.60	50.00	18.40	28.90	26.30	73.70
16	3.40	49.20	30.50	16.90	30.50	69.50
17	0.00	43.50	17.40	39.10	26.10	73.90
18	5.00	46.30	27.50	21.30	31.30	68.70
19	2.30	52.30	25.00	20.50	27.30	72.70
20	3.10	53.10	21.90	21.90	28.10	71.90
21	0.00	50.00	23.50	26.50	34.70	65.30
25	0.00	0.00	0.00	100.00	36.40	63.60
26	1.40	42.50	30.10	26.00	31.50	68.50
27	5.70	42.90	31.40	20.00	28.60	71.40
28	3.70	40.70	18.50	37.00	14.80	85.20
29	4.20	50.00	4.20	41.70	25.00	75.00
30	3.20	25.80	22.60	48.40	41.90	58.10
31	10.00	37.50	36.30	16.30	23.80	76.20
32	3.80	45.60	22.80	27.80	25.30	74.70
33	2.60	41.00	33.30	23.10	43.60	56.40
34	3.80	37.70	37.70	20.80	37.70	62.30
35	2.90	47.10	27.90	22.10	30.90	69.10
36	2.40	40.50	33.30	23.80	23.80	76.20
37	6.70	50.00	13.30	30.00	30.00	70.00
38	3.10	39.60	34.40	22.90	33.30	66.70
39	0.00	31.40	31.40	37.10	28.60	71.40
40	0.00	30.00	30.00	36.70	30.00	70.00
41	3.60	35.70	35.70	17.90	35.70	64.30
42	0.00	20.00	20.00	40.00	40.00	60.00
43	7.90	21.10	21.10	26.30	34.20	65.80
44	10.00	20.00	20.00	43.30	43.30	56.70
45	5.00	15.00	15.00	20.00	30.00	70.00
46	0.00	48.00	48.00	20.00	20.00	80.00

Consultant 25 is clearly incorrectly coded with 100% Dukes' Stage D patients yet exhibiting a high survival rate. A decision was made to exclude this surgeon from further analysis since it proved impossible to reconstruct the missing data from other items of information in the file.

More usefully, these case mix proportions do also give a feel for the extent of case mix variability, a feature of real data which needs to be allowed for in our modelling process and which provides useful base information upon which one can build a simulation exercise (see Chapter 4).

As an example, excluding surgeon 25, the range of proportions of patients with Dukes' Stage D tumours is 16% to 48%. The range of proportions of patients over 65 is 56% to 85%.

3.8 Site of the tumour

A brief examination of the site of the tumour was undertaken. We collapsed the extensive 'site' variable into two main categories, colon and rectum. In the form detailed in Appendix 1 the 'colon' category includes those cases falling into sites 1 to 7. A separate field was recorded for multiple invasive tumours. Approximately two thirds of the data fall into the 'colon' category and the percentages varied from 43% to 80% across the list of surgeons under consideration.

In fact although rectal surgery is considered technically more difficult, the two year survival rates were similar for the two site categories (excluding cases with multiple invasive tumours). It would have been of interest to pursue this study further since it is to be expected that more technically difficult surgery might highlight surgeon related variability more easily but two concerns over the data prevented further analysis being informative.

Firstly the reduction in patient numbers for rectal surgery reduces the power of the calculations appreciably. Secondly, and more importantly, we had clear concerns that the site field only contained information on the main tumour. This was evident from the low numbers classified as 'multiple invasive tumours' relative to the number classified as 'Dukes' Stage D' in other fields.

3.9 Finalised fixed effect audit results

A final problem was identified with the data. There were a large number of cases that had missing Dukes' Stages. We reconstructed the Dukes' Stage classifications from the other fields where possible and then calculated the proportions surviving from the 5 classifications (i.e. Dukes' Stage A,B,C and D together with a "missing" category). The proportion of patients classified as "missing" surviving to 2 years was approximately half way between the proportions surviving for categories Dukes C and Dukes D. We decided to impute these remaining fields as Dukes' Stage D. This enables all cases to be analysed, as opposed to just complete cases although there may be some bias introduced into our parameter estimates. As ever with audit exercises there is a trade off between accuracy and case numbers. Before performing this imputation exercise we checked that 'missingness' did not vary between surgeons, which it did not do to any meaningful extent. Since we are mainly interested, from an audit perspective, on the relative merits of surgeons as opposed to actual point estimates of performance the lack of variation in proportions of 'missingness' between surgeons is a comfort, since any bias in parameter estimates will be largely uniform across the pool of surgeons of interest.

From an audit perspective the imputation procedure in itself does not then cause unnecessary concerns. A decision to make a final reduction in the number of surgeons to those with over 25 cases was also made.

To assess the fundamental variation that we might expect from case mix alone we calculated the predicted mortality rates for 8 groups of patients produced by combinations of the (strongly predictive) Age and Dukes' Stage variables. These are calculated from a fixed effect regression routine where the predicted values are saved and output. The results are tabulated below.

Table 3.24

Predicted 2 year mortality rates subdivided by Age and Dukes' Stage

		Dukes' Stage			
		A	B	C	D
Age	<65	6.5%	13.7%	25.0%	48.4%
	≥65	9.7%	19.7%	34.0%	59.1%

The table below extends this analysis to illustrate both the variation in actual mortality rates and the corresponding variation in mortality rates that would have been expected for each surgeon given his or her case mix. The variation in expected mortality rates is substantial, from 26.9% to 37.8%.

The observed mortality rates vary from 12% to 53%. We also tabulate the 'excess deaths'. Figure 3.4 below plots the actual and expected mortality rates as a co-ordinate on a graph and clearly displays the greater variation in observed mortality.

Table 3.25

A comparison of observed deaths with those expected given the case mix

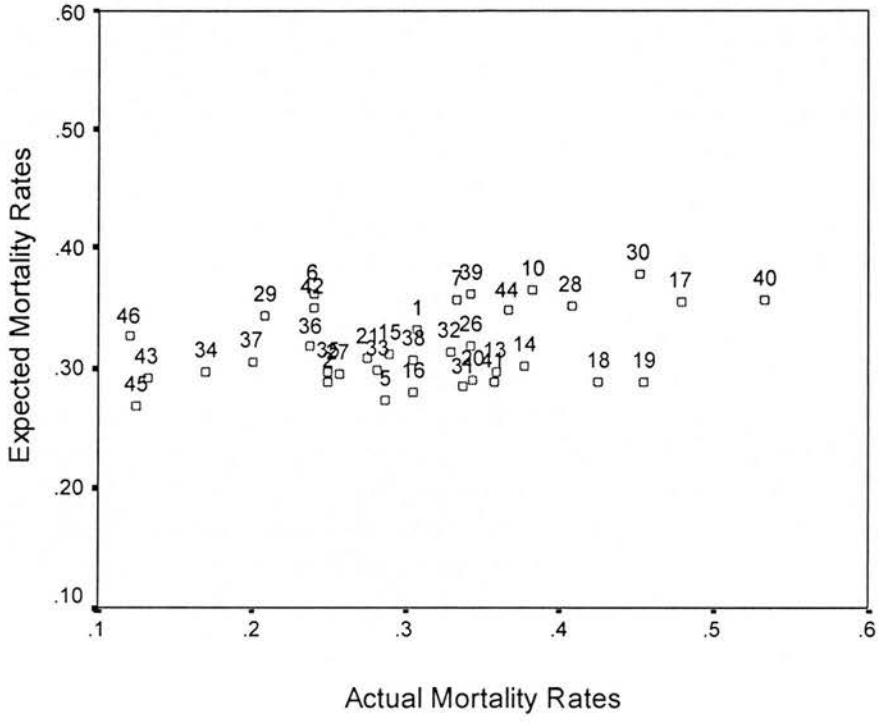
(as predicted by the regression model)

Surgeon	Cases	Deaths	Exp deaths	Excess deaths	Actual rate	Exp rate
1	39	12	12.94	-0.94	30.8%	33.2%
2	40	10	11.54	-1.54	25.0%	28.9%
5	94	27	25.77	1.23	28.7%	27.4%
6	25	6	9.06	-3.06	24.0%	36.2%
7	39	13	13.93	-0.93	33.3%	35.7%
10	34	13	12.43	0.57	38.2%	36.6%
13	39	14	11.57	2.43	35.9%	29.7%
14	53	20	16.00	4.00	37.7%	30.2%
15	38	11	11.87	-0.87	28.9%	31.2%
16	59	18	16.54	1.46	30.5%	28.0%
17	23	11	8.16	2.84	47.8%	35.5%
18	80	34	23.13	10.87	42.5%	28.9%
19	44	20	12.68	7.32	45.5%	28.8%
20	64	22	18.58	3.42	34.4%	29.0%
21	98	27	30.20	-3.20	27.6%	30.8%
26	73	25	23.25	1.75	34.2%	31.8%
27	35	9	10.33	-1.33	25.7%	29.5%
28	27	11	9.49	1.51	40.7%	35.1%
29	24	5	8.23	-3.23	20.8%	34.3%
30	31	14	11.72	2.28	45.2%	37.8%
31	80	27	22.77	4.23	33.8%	28.5%
32	79	26	24.81	1.19	32.9%	31.4%
33	39	11	11.64	-0.64	28.2%	29.8%
34	53	9	15.75	-6.75	17.0%	29.7%
35	68	17	20.17	-3.17	25.0%	29.7%
36	42	10	13.35	-3.35	23.8%	31.8%
37	30	6	9.16	-3.16	20.0%	30.5%
38	95	29	29.22	-0.22	30.5%	30.8%
39	35	12	12.67	-0.67	34.3%	36.2%
40	30	16	10.71	5.29	53.3%	35.7%
41	28	10	8.07	1.93	35.7%	28.8%
42	25	6	8.74	-2.74	24.0%	35.0%
43	38	5	11.11	-6.11	13.2%	29.2%
44	30	11	10.43	0.57	36.7%	34.8%
45	40	5	10.76	-5.76	12.5%	26.9%
46	25	3	8.15	-5.15	12.0%	32.6%
max					53.3%	37.8%
min					12.0%	26.9%

Figure 3.4

A plot of actual and expected mortality rates given case mix

(as predicted by the regression model)



The finalised tables of hazard ratios and confidence intervals are detailed below. These differ slightly from the preliminary results as data problems were resolved and the minimum number of cases increased to 25. We also restricted the explanatory variables to Age and Dukes' Stage. The tables show the case mix adjusted results for 6 month, 1 year and 2 year logistic regression models and for the Cox regression model with survival times censored at 2 years.

Table 3.26

Finalised relative risks with 95% confidence interval

Logistic regression 6 month outcomes

Surgeon	Relative Risk	95% confidence interval	
1	0.93	0.39	2.24
2	0.55	0.18	1.63
5	0.82	0.42	1.58
6	0.63	0.20	1.96
7	1.47	0.67	3.23
10	1.82	0.82	4.07
13	1.33	0.57	3.13
14	0.72	0.31	1.69
15	1.02	0.42	2.49
16	1.34	0.66	2.74
17	1.70	0.63	4.57
18	1.55	0.85	2.80
19	2.25	1.09	4.69
20	1.21	0.60	2.42
21	0.60	0.31	1.15
26	0.82	0.42	1.62
27	0.82	0.30	2.26
28	1.33	0.52	3.42
29	0.13	0.02	1.01
30	2.08	0.92	4.73
31	1.92	1.08	3.42
32	1.18	0.65	2.17
33	0.71	0.26	1.93
34	0.62	0.25	1.55
35	0.96	0.48	1.95
36	0.55	0.20	1.48
37	0.44	0.12	1.58
38	0.65	0.34	1.25
39	1.29	0.56	2.94
40	2.47	1.09	5.61
41	0.60	0.17	2.13
42	0.66	0.21	2.08
43	0.41	0.12	1.39
44	1.49	0.60	3.68
45	0.60	0.20	1.80
46	0.59	0.17	2.06

Table 3.27

Finalised relative risks with 95% confidence interval

Logistic regression 1 year outcomes

Surgeon	Relative Risk	95% confidence interval	
1	0.81	0.36	1.81
2	0.71	0.29	1.71
5	1.40	0.83	2.34
6	0.53	0.19	1.53
7	1.25	0.59	2.65
10	1.13	0.51	2.51
13	1.01	0.44	2.28
14	0.95	0.41	1.91
15	0.93	0.41	2.11
16	0.83	0.41	1.69
17	1.73	0.67	4.43
18	1.42	0.82	2.45
19	2.93	1.51	5.68
20	1.44	0.78	2.66
21	0.71	0.41	1.23
26	0.88	0.48	1.60
27	1.01	0.43	2.38
28	1.92	0.81	4.52
29	0.08	0.01	0.63
30	2.07	0.93	4.57
31	1.73	1.01	2.95
32	1.12	0.64	1.95
33	1.00	0.44	2.27
34	0.47	0.20	1.09
35	0.85	0.44	1.61
36	0.44	0.17	1.11
37	0.55	0.19	1.57
38	0.76	0.44	1.32
39	1.12	0.51	2.45
40	2.15	0.97	4.77
41	1.52	0.62	3.74
42	0.56	0.19	1.63
43	0.46	0.17	1.26
44	1.59	0.69	3.69
45	0.37	0.12	1.12
46	0.37	0.11	1.30

Table 3.28

Finalised relative risks with 95% confidence interval

Logistic regression 2 year outcomes

Surgeon	Relative Risk	95% confidence interval	
1	0.99	0.47	2.07
2	0.78	0.35	1.73
5	1.25	0.76	2.05
6	0.56	0.21	1.80
7	0.98	0.47	2.05
10	1.21	0.56	2.60
13	1.06	0.49	2.26
14	1.14	0.60	2.17
15	0.72	0.32	1.62
16	0.99	0.53	1.89
17	1.35	0.84	3.41
18	1.96	1.20	3.23
19	2.69	1.41	5.14
20	1.43	0.80	2.54
21	0.89	0.54	1.45
26	1.03	0.59	1.80
27	0.92	0.40	2.08
28	1.49	0.64	3.46
29	0.14	0.03	0.66
30	1.60	0.73	3.48
31	1.45	0.87	2.43
32	1.00	0.58	1.71
33	0.89	0.41	1.95
34	0.49	0.23	1.07
35	0.71	0.38	1.33
36	0.71	0.33	1.53
37	0.56	0.21	1.51
38	0.92	0.57	1.55
39	0.87	0.46	1.89
40	1.92	0.88	4.21
41	1.66	0.72	3.86
42	0.59	0.22	1.60
43	0.36	0.14	0.98
44	1.22	0.54	2.80
45	0.39	0.41	1.04
46	0.29	0.08	1.00

Table 3.29

Finalised relative risks with 95% confidence interval

Cox regression censored at 2 years

Surgeon	Relative Risk	95% confidence interval	
1	0.98	0.55	1.73
2	0.79	0.41	1.53
5	1.15	0.71	1.70
6	0.67	0.30	1.50
7	1.08	0.62	1.87
10	1.18	0.68	2.04
13	1.12	0.61	2.03
14	1.05	0.64	1.73
15	0.81	0.42	1.57
16	0.99	0.59	1.66
17	1.31	0.68	2.54
18	1.62	1.12	2.32
19	2.38	1.52	3.72
20	1.29	0.83	2.01
21	0.84	0.56	1.24
26	1.02	0.66	1.56
27	0.97	0.50	1.89
28	1.35	0.74	2.46
29	0.19	0.05	0.79
30	1.48	0.87	2.52
31	1.39	0.94	2.08
32	1.00	0.66	1.53
33	0.82	0.44	1.53
34	0.57	0.30	1.11
35	0.78	0.47	1.31
36	0.73	0.39	1.37
37	0.66	0.30	1.48
38	0.93	0.62	1.38
39	0.95	0.52	1.73
40	1.92	1.12	3.26
41	1.33	0.71	2.49
42	0.68	0.30	1.52
43	0.41	0.17	0.99
44	1.33	0.73	2.43
45	0.45	0.19	1.10
46	0.33	0.10	1.03

3.10 Random Effects Models

In Chapter 1, Section 1.10 we reviewed the class of statistical techniques known as random effect or multi-level modelling. We will not state again in full the advantages and disadvantages of such an approach but conceptually these methods are well suited to surgical problems since we are considering a natural hierarchical structure of patients within surgeons within hospitals. Greater insight is gained into the underlying structure of complex sets of data when we use such additional techniques (in addition to classical fixed effect models). In this study the extension to a second level (the hospital level) is limited by the low number of units at this particular hierarchy, although it can still be modelled as will be discussed below.

The key advantage of such an approach is that when comparing one surgeon against their peer group we do not make an underlying assumption that the peer group are entirely homogeneous. We can more reasonably allow for the fact that surgeons have characteristics in common and that their performance measures can be assumed to have been drawn from some underlying distribution. Several software packages are now available each with advantages and disadvantages. We chose to use the WinBUGS software as has been used before in medical studies (including the Bristol Inquiry).

One advantage of this particular software package is that it is perhaps better suited to problems where the case numbers can be small for some units (i.e. surgeons). With small data numbers the asymptotic estimates used by some software packages can lead to inaccurate estimates of confidence intervals for highly skewed distributions such as relative risk ratios (Goldstein,1995)

We again analysed the West of Scotland CRAG data but before reporting our results state the main conclusions of the fixed effect analysis.

1. No hospitals had 95% confidence intervals for the relative risk measure that excluded unity on either crude or a case mix adjusted bases.

2. Four surgeons had 95% confidence intervals for the relative risk measure that excluded unity using 2 year outcomes.

3. As one moves to shorter term outcome measures we observe in some cases that different surgeons are identified as having performance which differs from average. We observe fewer significant results as death number fall and confidence intervals widen. There was little difference between the broad inferences arrived at using 2 year logistic regression or 2 year Cox regression.

The statistical model and WinBUGS code below for a fixed effect model was used to check that our Bayesian approach produced similar inferences to our earlier analysis in SPSS.

The measures of relative performance and confidence intervals differed slightly as one would expect (since there will be some sampling error in the WinBUGS procedure and there is the additional effect of the weak prior distributional assumptions) but the surgeons identified as being significantly different from average were the same.

If p_i is the probability of death for the i^{th} patient then we model the log odds of p_i as being a linear function of the covariates of interest (age –denoted ag[i] , Dukes' stage-denoted du[i] and the surgeon identifier-denoted sur[i]). This linear function is seen in line 5 of the code below. The probabilities of death for each surgeon are independent of each other (i.e. a fixed effect model).

```
model {

for (i in 1:N) { # N = number of patients, so here we have

# y[i] = binary response for each patient

y[i] ~ dbern(p[i])

logit(p[i]) <- alpha + beta.age * ag[i] + beta.dukes[du[i]] + beta.surg[sur[i]]

}

# Vague priors on regression coefficients

alpha ~ dnorm(0, 0.000001)

beta.age ~ dnorm(0, 0.000001)

beta.dukes[1] <- 0 # set level 1 of coefficient for dukes to 0 to act as reference category

for (k in 2:4) {

beta.dukes[k] ~ dnorm(0, 0.000001)

}

# For fixed surgeon effects:

beta.surg[1] <- 0 # set level 1 of coefficient for surgeon to 0 to act as reference category

for (k in 2:36) {

beta.surg[k] ~ dnorm(0, 0.000001)

}

}

list(alpha=0.3,beta.dukes=c(NA,0,0,0),

beta.surg=c(NA,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),beta.age=0)

list(N=1636)

y[]  ag[] du[] sur[]

0        2        4        9

0        1        2        15 etc for 1636 records
```

We used this checking procedure to experiment with burn in times, run lengths and different initial values before becoming comfortable with a particular set of computing parameters.

The hospital level fixed effects can be estimated directly from the surgeon level analysis (in this case) by averaging the surgeon level beta parameters in the linear part of the model and exponentiating the results. This confirmed our results from the SPSS analysis. Interestingly one hospital contained the poorest and the second best performers and another contained the worst performer but with all other surgeons having a relative risk below 1, an example of the diversity that one can see in some cases. This is of course a concern from an audit perspective. Analysis at the hospital level gives greater patient numbers and statistical power but there is a risk that it can mask some important surgeon level variability in the data. The following WinBUGS code was then used to fit a random effects model for surgeons.

The difference between this model and the previous one is that the surgeon effects (beta.surg) are drawn from a normal distribution with variance $1/\tau$ (see line 16 of the code below). We assume therefore that there is some underlying similarity in death rates between surgeons. They are no longer independent of each other (i.e. a random effects model).

```
model {

for (i in 1:N) { # N = number of patients, so here we have

# y[i] = binary response for each patient

y[i] ~ dbern(p[i])

logit(p[i]) <- alpha + beta.age * ag[i] + beta.dukes[du[i]] + beta.surg[sur[i]]

}
```

```
# Vague priors on regression coefficients

alpha ~ dnorm(0, 0.000001)

beta.age ~ dnorm(0, 0.000001)

beta.dukes[1] <- 0 # set level 1 of coefficient for dukes to 0 to act as reference category

for (k in 2:4) {

beta.dukes[k] ~ dnorm(0, 0.000001)

}

# For random surgeon effects:

for (k in 1:36) {

beta.surg[k] ~ dnorm(0, tau)

}

tau ~ dgamma(0.001, 0.001)

}

list(alpha=0.3,beta.dukes=c(NA,0,0,0),

beta.surg=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0), tau=1, beta.age=0)

list(N=1636)

y[]  ag[] du[] sur[]

0        2        4        9

0        1        2        15 etc
```

We expected some shrinkage in the estimates and a reduction in the numbers of significant results but the conclusions were emphatic as inspection of Tables 3.28 and 3.29 illustrates. None of the surgeons were identified as being significantly different from their peer group when one moved to this more complex (but realistic) model. As an example the relative risk of surgeon 18 falls from 1.96 to 1.38.

In this particular set of data the numbers involved are relatively small both for individual surgeons and as a whole. This leads to more pronounced shrinkage of surgeon effects towards the population mean (a relative risk of 1). Other studies where data numbers are larger exhibit similar features but in a less pronounced fashion (Leyland, 1995). Given the fact that the data being analysed are representative of those for audits of specialised surgery this is an important feature of interest. Again we tested the effects of different starting values, burn in lengths and run times before being satisfied as to the accuracy of the procedure. We report only the 2 year results but the 1 year and 6 month results were also prepared and present a similar picture to what was observed with the fixed effect models.

Table 3.30

Relative risk estimates for the random effects model – 2 year logistic regression

Surgeon	Relative Risk	95% confidence interval	
1	0.98	0.60	1.54
2	0.95	0.58	1.49
5	1.07	0.76	1.56
6	0.92	0.52	1.35
7	1.00	0.64	1.55
10	1.06	0.70	1.70
13	1.00	0.65	1.58
14	1.02	0.69	1.59
15	0.92	0.54	1.37
16	0.98	0.63	1.51
17	1.09	0.71	1.96
18	1.38	0.96	2.35
19	1.40	0.94	2.68
20	1.15	0.81	1.86
21	0.96	0.66	1.36
26	1.02	0.70	1.52
27	1.00	0.64	1.59
28	1.13	0.74	2.00
29	0.77	0.35	1.16
30	1.13	0.76	1.99
31	1.19	0.83	1.87
32	1.00	0.65	1.45
33	0.98	0.61	1.53
34	0.83	0.46	1.23
35	0.91	0.55	1.32
36	0.93	0.55	1.35
37	0.90	0.48	1.35
38	1.00	0.68	1.38
39	0.98	0.61	1.50
40	1.19	0.84	2.17
41	1.12	0.75	2.04
42	0.87	0.46	1.36
43	0.87	0.45	1.29
44	1.05	0.66	1.71
45	0.82	0.41	1.22
46	0.80	0.37	1.18

We illustrate below an additional very powerful and informative facility within the WinBUGS software. As part of the Monte Carlo updating procedure one can record the rank order of a particular node at each simulation. This then enables one to consider the rank order itself as a random variable with its own statistical distribution. The table below gives the 95% confidence intervals for the rank orders in the fixed and random effect models considered earlier. As one would expect given the profile of relative risks for the random effects model reported earlier the confidence intervals for the rank orders in this case are so wide as to be meaningless in practice. This new approach to quantifying the uncertainty of institutional rankings illustrates quite clearly the risks of merely looking at confidence intervals for the relative risk. Normal methods enable one to determine divergence from average (in the context of a particular fixed or random effects model) but do not easily show how extreme a surgeon actually is in reality. Even those surgeons identified as being significantly different from average (in the context of a fixed effect model) have very wide confidence intervals for their rank orders. It is not possible to say that (if one considered rank orders alone) that many are in the top or bottom quartile (despite being significantly different from average). Figure 3.5 illustrates the distribution of rank orders for two sample surgeons in the fixed effect model.

Table 3.31

Confidence intervals for the rank order of surgeons

2 year outcomes – logistic regression – fixed effect model

Surgeon	Median Rank	95% confidence interval	
1	12	4	30
2	13	4	29
5	24	13	33
6	9	2	28
7	24	9	34
10	22	7	34
13	17	5	32
14	18	6	31
15	16	5	31
16	16	5	30
17	32	13	36
18	27	16	34
19	35	27	36
20	27	14	35
21	13	5	24
26	17	7	29
27	21	6	34
28	33	17	36
29	1	1	5
30	32	17	36
31	30	19	35
32	21	9	31
33	19	6	33
34	7	2	20
35	16	6	29
36	6	2	20
37	8	2	27
38	14	6	25
39	21	7	33
40	33	18	36
41	28	10	36
42	5	1	25
43	10	3	28
44	28	11	36
45	3	1	14
46	3	1	15

Table 3.32

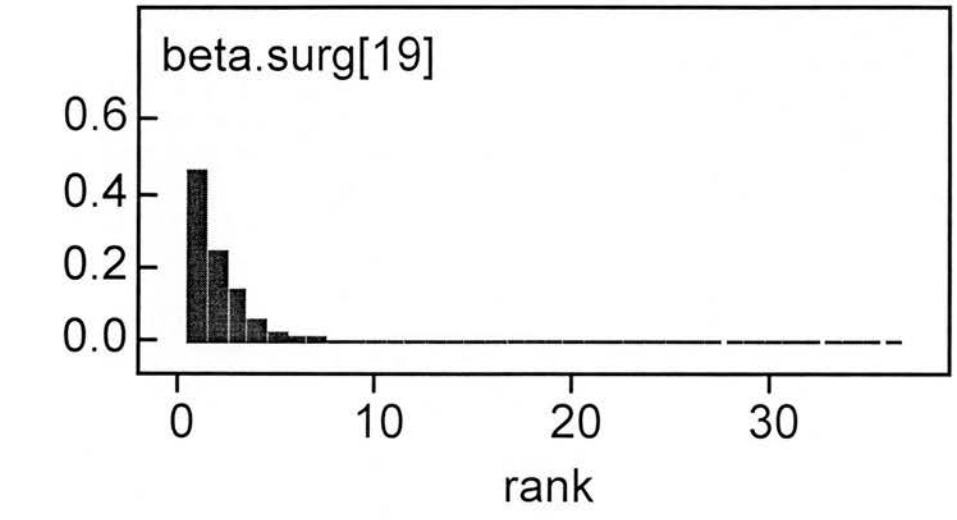
Confidence intervals for the rank order of surgeons

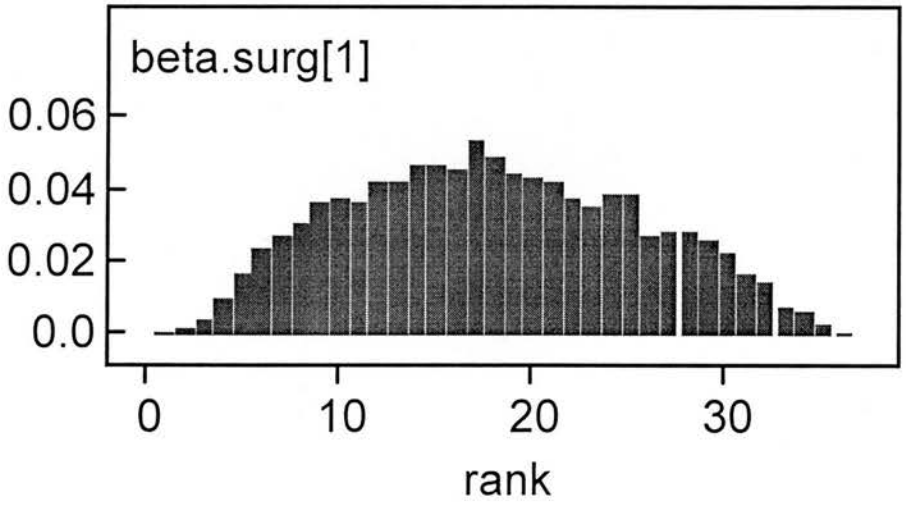
2 year outcomes – logistic regression – random effect model

Surgeon	Median Rank	95% confidence interval	
1	18	2	35
2	15	2	34
5	23	5	35
6	13	1	33
7	19	2	35
10	22	3	35
13	19	2	35
14	20	3	35
15	13	1	33
16	17	2	34
17	24	3	36
18	33	13	36
19	33	9	36
20	27	6	36
21	16	3	33
26	19	3	34
27	18	2	35
28	26	3	36
29	6	1	32
30	26	4	36
31	28	6	36
32	18	2	34
33	17	2	34
34	8	1	30
35	13	2	33
36	14	1	32
37	12	1	33
38	19	3	34
39	17	2	34
40	28	7	36
41	25	4	36
42	11	1	34
43	10	1	32
44	21	2	35
45	8	1	31
46	7	1	31

Figure 3.5

Two distributions of rank orders (surgeon 29 and surgeon 1) – Fixed Effect Model





Note for computing reasons surgeon 29 is coded as 19 in the WinBUGS analysis and output (since all surgeons must be sequentially numbered).

We extended the model to have a random effect at a second hierarchy (the hospital level). The code is detailed below and ran satisfactorily on WinBUGS but no results are reported. As expected confidence intervals widened modestly as additional prior assumptions are made but overall inferences were very similar to those observed earlier. The lack of cross classification and small data numbers weaken the effectiveness of these more complex models although they would have interesting applications for larger studies with greater hospital numbers.

```
model {
for (i in 1:N) { # N = number of patients, so here we have
# y[i] = binary response for each patient
y[i] ~ dbern(p[i])
logit(p[i]) <- alpha + beta.age * ag[i] + beta.dukes[du[i]] + beta.surg[sur[i]]
}
# Vague priors on regression coefficients
alpha ~ dnorm(0, 0.000001)
beta.age ~ dnorm(0, 0.000001)
beta.dukes[1] <- 0 # set level 1 of coefficient for dukes to 0 to act as reference category
for (k in 2:4) {
beta.dukes[k] ~ dnorm(0, 0.000001)
}
# For random hospital effects:
for (k in 1:7) {
beta.hosp[k] ~ dnorm(0, tau.hosp)
rr.hosp[k] <- exp(beta.hosp[k])
 tau.hosp ~ dgamma(0.001, 0.001)
```

```
# For random surgeon effects:

for (k in 1:36) {

beta.surg[k] ~ dnorm(beta.hosp[h[k]], tau)

rr.surg[k] <- exp(beta.surg[k])

}

 tau ~ dgamma(0.001, 0.001)

}
```

INITIAL VALUES

```
list(alpha=0.3,beta.dukes=c(NA,0,0,0),

beta.surg=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),beta.age=0,

beta.hosp=c(0,0,0,0,0,0,0), tau=1, tau.hosp=1)
```

DATA

```
list(N=1696)
```

```
y[]  ag[] du[] h[]  sur[]

0        2        4        3        9

0        1        2        4        15 etc
```

We did not extend the modelling to examine the relative merits of random effect Cox survival regression ('frailty models') as opposed to random effects logistic regression because the fixed effects Cox regression results were so similar to the logistic regression results. This would however be a topic worthy of further study, preferably in an investigation where there was greater cross classification and more significant loss to follow up.

3.11 Conclusion

In summary the CRAG West of Scotland data have provided us with some useful insights into not only the particular set of data in question but into the various classes of statistical models which one can use to analyse the data.

Our conclusions are as follows:

1. Even in a well-organised prospective study of audit data problems can arise. Considerable time has to be expended on data checking and validation before commencing formal analysis. Many items of data are not always recorded but those variables of particular interest were fairly complete. Imputation procedures can be used (with care) in some cases for particular variables.

2. There is clear variation in outcomes for different classes of patients. Case mix must be allowed for in the analysis (unless cases are stratified into homogeneous groups). We saw a wide variation in expected survival percentages for the surgeons in the study (and even greater variation in actual survival percentages). Inference bases on crude death rates alone are not reliable.

3. A small number of important covariates capture the main case mix effects. In fact Age and Dukes' Stage are satisfactory, particularly if we stratify with respect to procedure. We require two years of follow up information to be able to produce reliable inferences on surgeons or hospitals.

4. The data quality was insufficiently good to investigate whether a narrow focus group of cases might offer similar levels of discrimination between surgeons (e.g. rectal surgery). This would however be a interesting topic for further study.

5. The aggregation of surgeons into hospitals can disguise some important surgeon level effects and it would be of interest to analyse a similar (or larger) set of data where there was greater cross-classification of surgeons within hospitals.

6. It is not enough to examine data using fixed effect models alone. New software packages enable the statistician to allow for more natural descriptions of variability in the underlying model. They allow production of output which illustrates not just the size of divergence of performance relative to mean effects but also allow one to explore the distribution of rank orders of the surgeons or hospitals of interest. This is particularly useful given ongoing media interest in 'league tables' of institutional performance. There is a role for both fixed and random effects models but exclusive use of the former class of models can lead to systematic overstatement' of variability.

Chapter 4

A Simulation Exercise in Surgical Audit

4.1 Introduction

The previous chapters have highlighted several issues within the framework of an audit of mortality data following surgery for colorectal cancer.

1. There are considerable problems in using crude mortality data alone in audit exercises. Case mix effects must be allowed for in any analysis and a small number of key covariates capture the main effects. The risk adjusted performance measures can however be sensitive to the selection of certain explanatory variables which are proxies for the same clinical features. The use of ranking procedures when examining institutional effects can be seriously misleading.

2. Conventional statistical analyses using fixed effect models can produce results which highlight evidence of poor or good performance relative to other surgeons (or hospitals) but that in the same analyses when more complex and realistic models are developed these inferences are less forceful. In the context of the number of cases involved and the high mortality rates associated with colorectal cancer statistically significant results are also highly clinically significant.

3. There are problems with data quality in many circumstances. Dealing with missing data, normally using imputation methods, can also lead to artificially small estimates of statistical variability. This could lead to inappropriate inferences being made about institutional performance

The exploratory analysis of the CRAG West of Scotland colorectal data provided us with a basis for developing a simulation exercise the objective of which was to explore some wider statistical issues as they relate to audit exercises. In particular these real data provide us with a basis from which to build parametric models for survival times where the model parameters vary depending on the characteristics of the individual patient.

It also gives us a basis from which to examine the pace at which caseload is accumulated by surgeons and the variability of case mix between surgeons and between hospitals, as appropriate. If a simulation study is to be helpful it has to be a reasonable, if idealised, approximation to what is observed in practice.

4.2 The motivation behind simulation studies

Given that we were satisfied that the CRAG West of Scotland colorectal cancer data are a sound basis for developing a simulation model there is a requirement to state clearly in advance exactly what the motivation behind such a simulation study generally is and further what specific research questions such a study enables us to answer. The reason for such precise focus being required is that there are innumerable scenarios that one could consider, even for a fairly modest simulation based exercise. In this thesis only a limited number of scenarios and problems are explored although this does enable other questions to be addressed by implication. An example of this point is that some of the conclusions from a simulated surgical audit exercise based on, say, equal numbers of patients can be generalised to smaller or more variable groups of patients per surgeon. In effect certain results from studies based on simulated data give us bounds on the results for other studies which have not been explicitly explored.

The motivation behind most simulation studies is that they facilitate the exploration of the evolution of very many future scenarios for a random process of interest. We can simulate data and examine the effects on outcomes or functions of outcomes. They enable the researcher to examine the sensitivity of results and inferences to changes in the underlying model specification and parameters. For many statistical problems simulation methods do indeed provide the only possible way of exploring aspects of random processes. Closed form solutions to statistical problems are not always available. Applications of simulation methods range from industrial and financial problems (e.g. the study of the evolution of economic time series) to medical problems (Morgan, 1984).

In the context of this thesis the key advantage of simulation exercises is that they enable us to specifically examine the ability of prevailing statistical methods to detect differences in performance when we know *a priori* that they do in fact exist. We have the ability to alter the model parameters in such a way that a particular surgeon or hospital has a hazard ratio of, say, twice his peer group. Knowing this we can perform many simulations and observe how many times we actually detect this known effect. The power will depend on the size of the effect we are trying to detect and on the underlying model specification (the parameters of the model, the distributions of numbers of patients between surgeons and the fundamental variability of the process we are examining).

Another application of a simulation study is that we can constrain all our surgeons to have identical performance, for a given patient type, and examine how many times we falsely infer that an individual actually differs significantly from his or her peer group. The key point is that we know the answer to our problem in advance, and we are examining the effectiveness of the statistical tools at our disposal to correctly identify what we know to be true. This then enables us to appraise the effectiveness of individual audit studies that can be thought of as being a single realisation of one particular batch of random outcomes.

The two main questions specified above (effectively the determination of Type 1 and Type 2 statistical errors) are addressed later in this chapter but other areas of study can emerge from simulation exercises. One potential interest for audit specialists would be an examination of the variation in power for different samples sizes (and for a given effect size). This could take the form of progressive overall increases in sample sizes with patient numbers equal within the overall sample, the variation of surgeon numbers within a fixed population of cases or a variation in the distribution of numbers of patients per surgeon. As discussed above there is no limit to the scenarios which can be explored but the results for some main studies can be used to place upper or lower bounds on the likely results for subsidiary analyses.

In a simulation study we can also examine aspects of parameter uncertainty and in particular examine the sensitivity of outcomes and inferences to alterations in our pre-specified model parameters. Ideally we would wish our results to be relatively robust with respect to modest alterations in our pre-specified parameters. These final points are areas of possible future research activity.

4.3 Preliminary Data Analysis

As reported in Chapter 3, Sections 3.5 and 3.9 the CRAG West of Scotland Colorectal data were reduced for the purposes of analysis to exclude certain surgeons with very low case numbers and a degree of (reasoned) imputation was carried out. Table 1 below gives the proportion of cases that fall into the 8 potential categories for combination of Dukes' Stage and Age (as represented by the binary variable used in Chapters 2 and 3).

Table 4.1

Proportions of cases in the CRAG West of Scotland data falling to groups stratified by Age and Dukes' Stage

	Dukes' Stage			
Age	**A**	**B**	**C**	**D**
Age < 65	1%	12%	9%	10%
Age ≥ 65	2%	31%	17%	18%

The observed proportions surviving two years were as follows

Table 4.2

2 Year Survival proportions from the CRAG study –

Stratified by Age and Dukes' Stage

	Dukes' Stage			
Age	**A**	**B**	**C**	**D**
Age < 65	90.9%	91.0%	71.7%	50.6%
Age ≥ 65	92.5%	78.7%	65.5%	46.0%

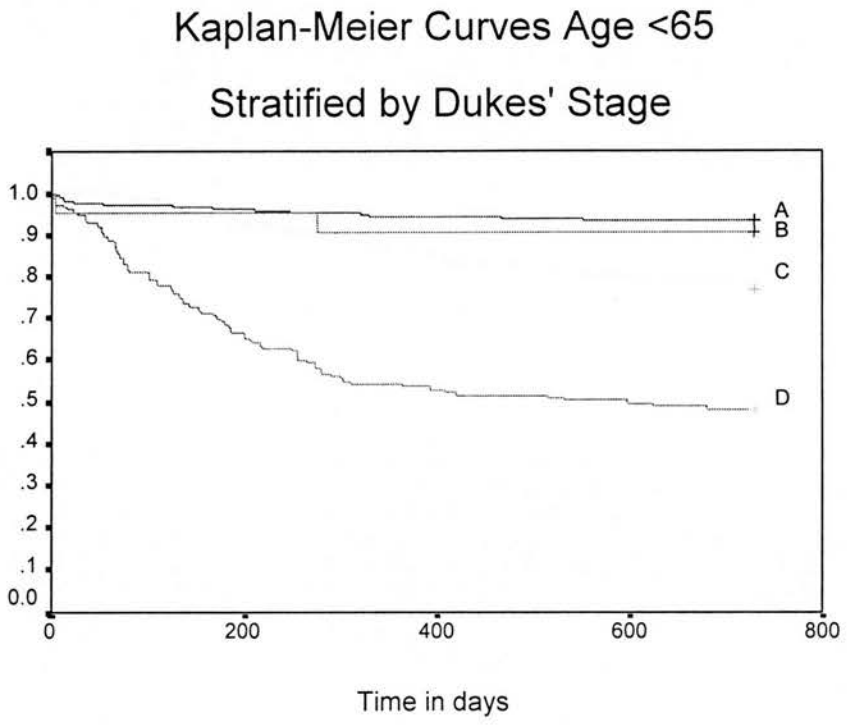
It can be readily seen by inspection that the proportion of cases which represent the least advanced cases of disease is small (i.e. Dukes' Stage A). Of course there are a number of other explanatory variables which could be used either as a proxy for Dukes' Stage or in addition to this clearly important covariate. The objective of a simulation example is not however to develop a large number of potential categories into which a patient might fall but to develop a framework which captures the broad personal and diagnostic effects which influence survival following surgery for colorectal cancer. The conclusions drawn in Chapter 2 support this approach.

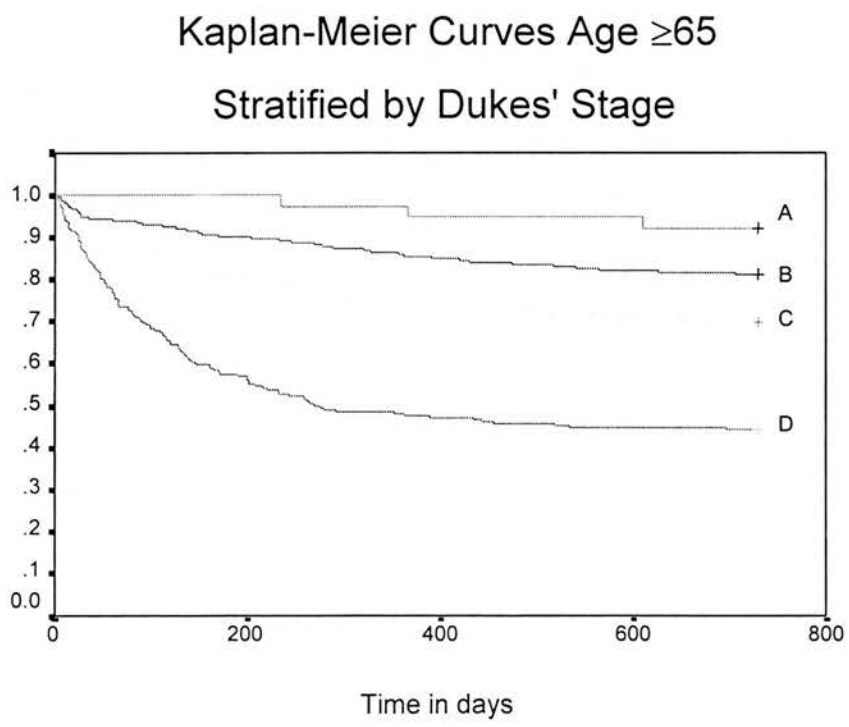
On balance, having examined the dominant effects evident from stepwise regression procedures and paying due regard to the well established clinical methods of grading severity of disease, we decided to base the simulation exercise where patients could fall into 6 potential categories, Dukes' Stage B,C or D and Age < 65 or Age ≥ 65.

The survival curves from the CRAG West of Scotland data stratified by Dukes' Stage and Age are detailed in the figures below.

Figures 4.1 and 4.2 – CRAG West of Scotland Kaplan-Meier Curves

(censored at two years)





As reported earlier the analysis of mortality data is perhaps best achieved, for colorectal cancer survival data, with a logistic regression model based on survival (or not) to two years following (admission for) surgery. Shorter-term binary outcome measures have the problem of insufficient deaths and can reduce the effectiveness of the modelling of major case mix effects.

For a substantial set of data on colorectal cancer surgery use of Cox regression models (with all observations censored at 2 years) gave very similar inferences to basic logistic regression models using 2 year outcomes. This of course is dependant on there being minimal loss to follow up. If patients lost to follow up were large in numbers then logistic regression would be a poor alternative to analyses based on survival times, discarding as it would much useful data.

Much of the analysis that follows (and in particular the use of random effects models) has been based around logistic regression models. We did for completeness however rework the power calculations for the fixed effect Cox models (including and excluding allowance for case mix) and these were seen to be extremely similar to the results obtained from the logistic model. The notion of survival time is the most natural framework upon which to build surgeon and hospital effects but in this particular context logistic analyses provide a useful and practical alternative.

4.4 Parametric Models for Survival Times

A mechanism was therefore required to simulate the individual case survival time conditional on occupancy of a particular state (a permutation of Dukes' B, C, D and age <65 and > 65). The empirical survival functions graphically illustrated above give a framework for the selection of parametric models of choice. An exact representation is not required, only a model that captures the broad case mix effects observed in practice.

Non-linear regression routines written for the software package MINITAB 11.0 were used to examine the goodness-of-fit of the range of models for survival times following surgery for colorectal cancer detailed below.

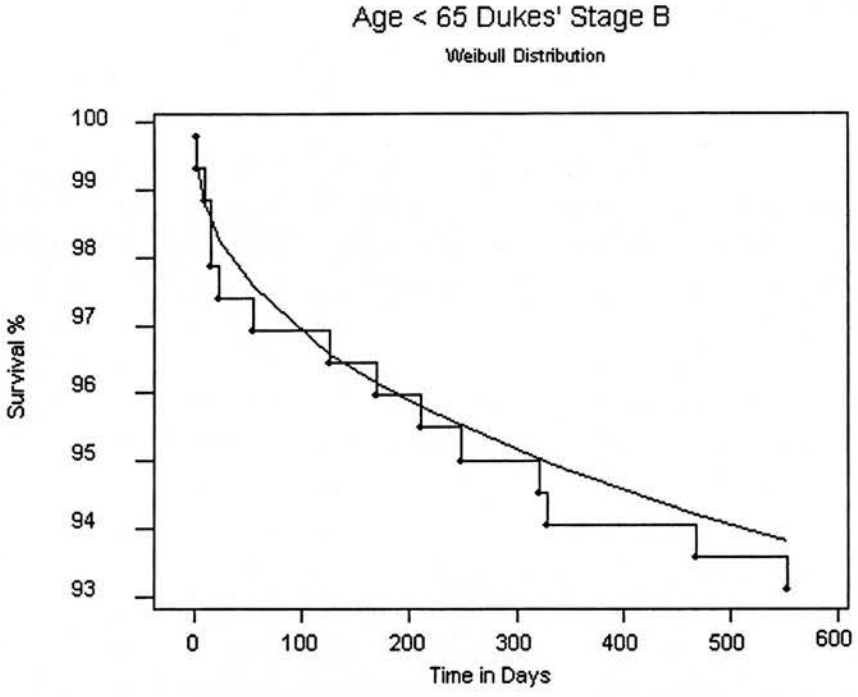
- Exponential distributions
- Weibull distributions
- Lognormal distributions
- Log logistic distributions

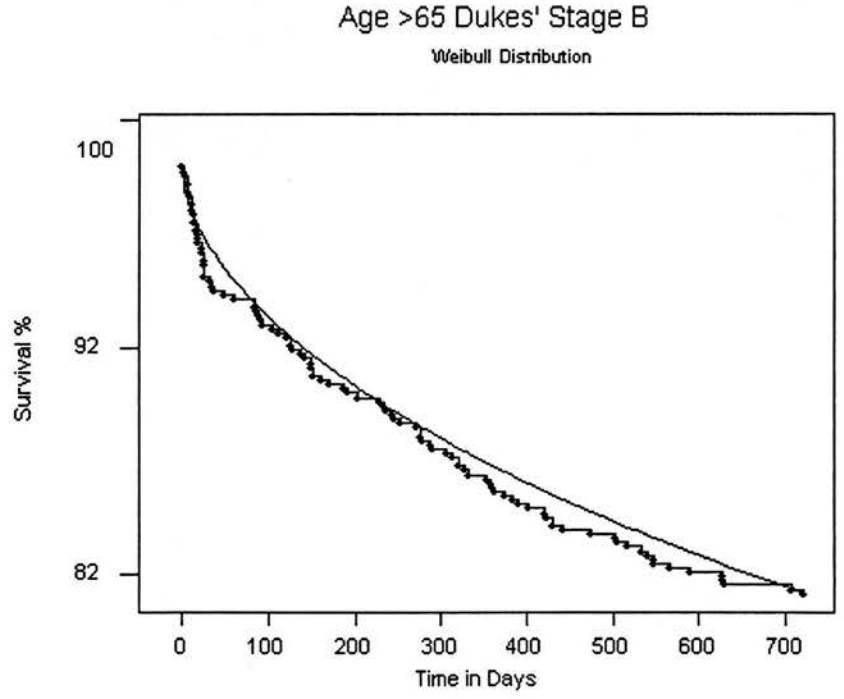
In fact the Weibull distribution offers the best overall fit to the data across the six possible categories, Details of the exploratory analysis involved in fitting a variety of different parametric models are not reported but the conclusion is in accordance with general thoughts about the flexibility of the Weibull distribution to model many types of survival and reliability data in both medical and industrial fields of application.

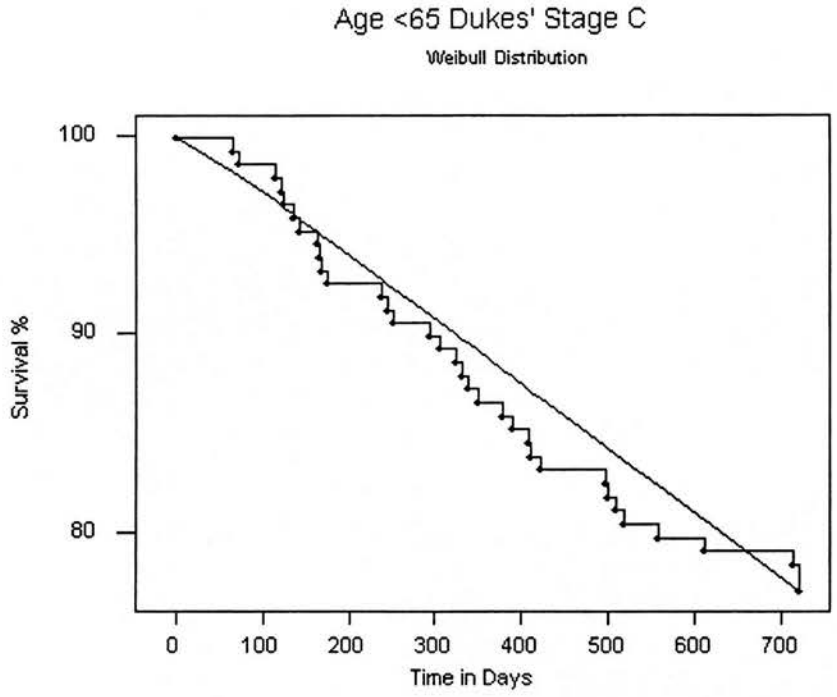
Figures 4.3 to 4.8 illustrate the general adequacy of the fit across the range of categories. Improved adherence to data could have been achieved by using alternative models for the two Dukes' Stage D groupings which better cope with the rapid initial decline in the survival function but, given the fact that it was 2 year outcomes which were mainly of interest as opposed to the distribution of times of death before 2 years, the Weibull distributions were considered to be a satisfactory basis for the simulation exercise.

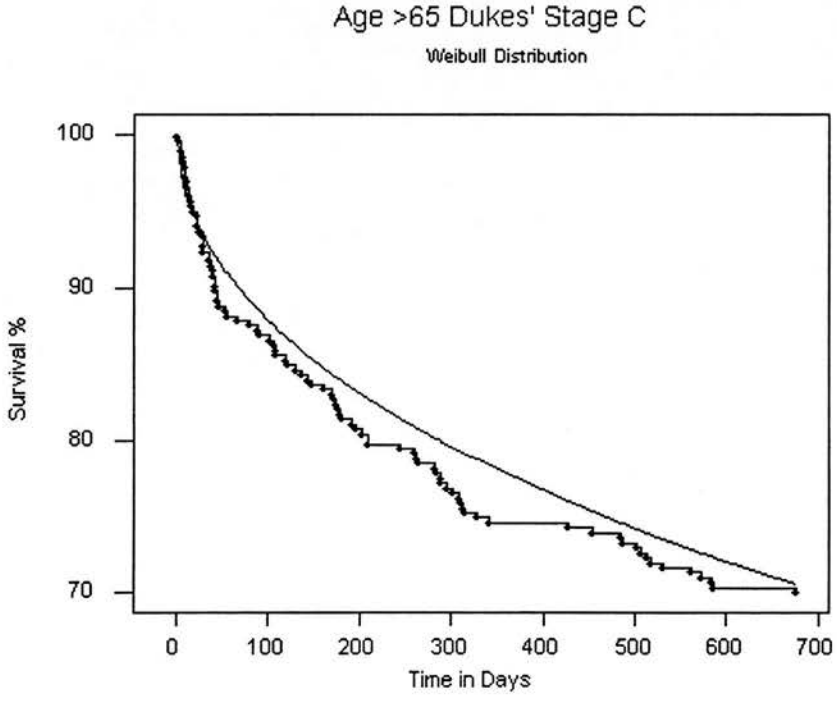
Figures 4.3 to 4.8

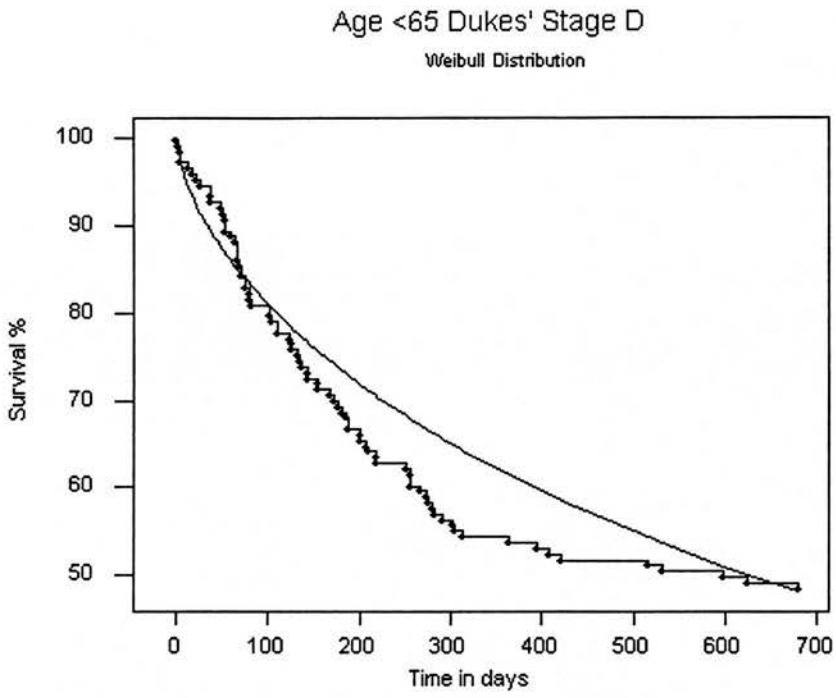
Graphical Comparison of Kaplan-Meier and Weibull Models

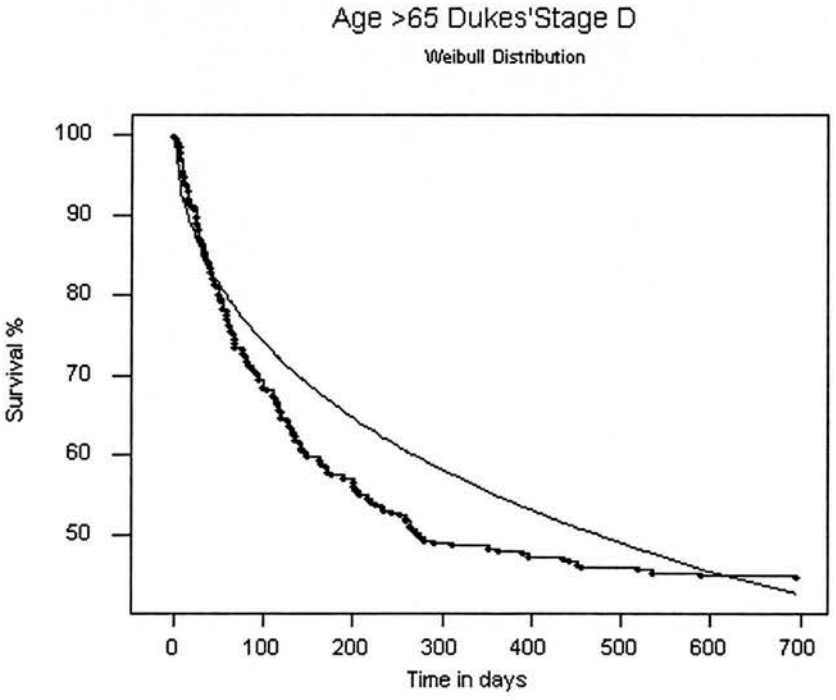





Age <65 Dukes' Stage C
Weibull Distribution
100
90
80
Survival %
0
100
200
300
400
500
600
700
Time in Days


Age >65 Dukes' Stage C
Weibull Distribution
100
90
80
70
Survival %
0
100
200
300
400
500
600
700
Time in Days


Age <65 Dukes' Stage D
Weibull Distribution
Survival %
100
90
80
70
60
50
0
100
200
300
400
500
600
700
Time in days


Age >65 Dukes'Stage D
Weibull Distribution
Survival %
100
90
80
70
60
50
0
100
200
300
400
500
600
700
Time in days

A further advantage of the Weibull distribution is that the parameters have a natural interpretation when one considers the a typical outcome measure used in audit studies, the relative risk measure. The Weibull model has two parameters, a shape parameter and a scale parameter. It can be demonstrated that a fixed proportionate adjustment to the scale parameter in a Weibull distribution (with a fixed shape parameter) is equivalent to the same proportionate adjustment in the relative risk. That is to say, if the scale parameter is doubled the relative risk is doubled. In some statistical texts and software packages the scale parameter is in fact inverted so a relative risk of 2 would be given if the scale parameter is halved. Having chosen the Weibull distribution we estimated the 12 parameters of the six distributions using the non-linear regression routine in MINITAB 11.0. The results are detailed in the table below.

Table 4.3

Parameters of the Weibull Models Derived from the CRAG Study

	Shape	Scale
Dukes' B & Age <65	0.43	360411
Dukes' B & Age ≥65	0.56	11831
Dukes' C & Age <65	1.12	5078
Dukes' C & Age ≥65	0.52	2386
Dukes' D & Age <65	0.65	1094
Dukes' D & Age ≥65	0.54	922

Algebraic manipulation of the formula for the Weibull distribution also gives a convenient expression for the expected proportion alive two years after surgery. These have been calculated and are detailed in the table below. The figures in parenthesis are the actual observed survival proportions from the analysis of the CRAG West of Scotland set of data.

Table 4.4

A comparison of observed and predicted survival rates at 2 Years (Weibull Model)

	Dukes' Stage			
Age	**A**	**B**	**C**	**D**
Age < 65	N/a	93% (91%)	77% (72%)	46% (50%)
Age ≥ 65	N/a	81% (79%)	69% (66%)	41% (46%)

Given the broad objectives of this simulation exercise this is an encouraging fit at the end point of the study (two years). The earlier graphs show by eye that adherence to the survival curve is adequate up to two years and probability plots were reviewed to look at measures of fit in a more formal context as well.

4.5 Allowance for Case Mix Variation

A decision also had to be made as to the basic case mix to be assigned to an individual surgeon. This was centred around rounded proportions from the actual CRAG data analysis and is summarised in the table below.

Table 4.5 – Case Mix Allocation Chosen For Simulation Purposes

	Dukes' Stage			
Age	**A**	**B**	**C**	**D**
Age < 65	nil	10%	10%	10%
Age ≥ 65	nil	30%	20%	20%

Of course in reality surgeons have varying case mix. We needed then to develop a mechanism for varying case mix between surgeons as part of the simulation exercise. This could have been done using probabilistic methods randomly allocating cases to categories in accordance with a prespecified distribution with the above asymptotic proportions in mind but this ran the risk of producing results which would have diverged meaningfully from the actual experience we observed when analysing the CRAG West of Scotland data. As a compromise we selected 20 surgeons at random from the 36 analysed formally before and used the observed case mix of these 20 surgeons as a basis for imposing case mix variability on our simulation study. This imposition of case mix is important since we need to comment on the necessity of using case mix adjustments in analysing data.

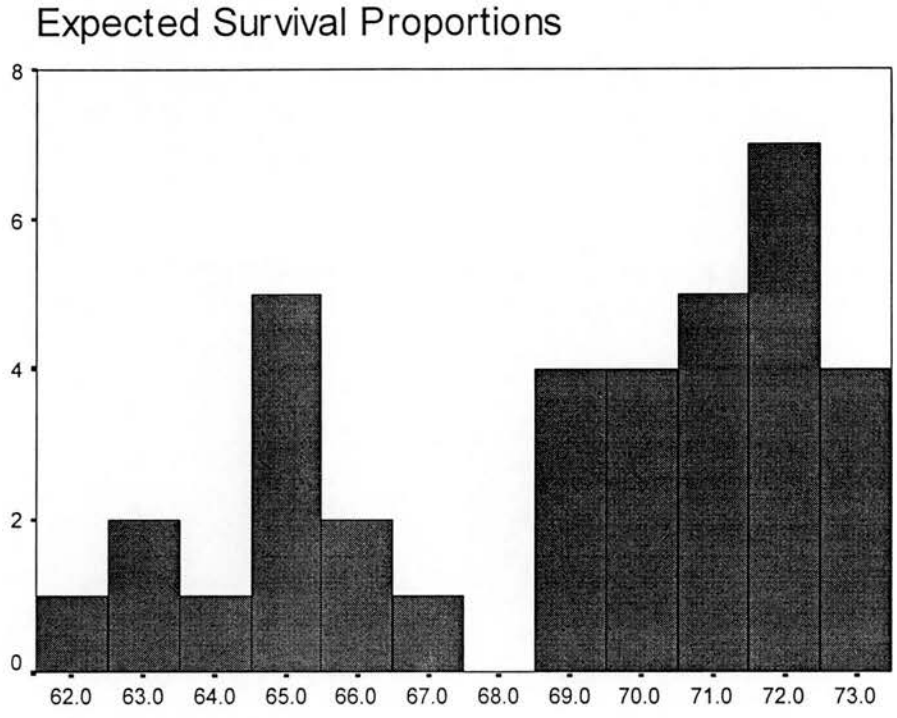
To gauge the effect of the importance of case mix we can generate a list of expected survival proportions for the actual surgeons analysed. This was reported in Chapter 3, Section 3.9 when logistic regression models were used to give expected survival proportions but we can repeat the exercise by using the predicted survival proportions from our six Weibull models derived above.

For the 36 surgeons analysed the expected survival proportions ranged from 61.6% to 73.4% based on the Weibull distributions. The corresponding range from the logistic regression analysis (using all 4 Dukes' Stages) was 62.2% to 73.1%. Again this gives comfort as to the applicability of our simulation procedures. A histogram of proportions is detailed below.

Figure 4.9

Distribution of expected survival proportions

(as predicted by Weibull model given known case mix)



The table below gives the numbers of patients per category for the restricted number of surgeons used in the simulation exercises. It illustrates the variation in case mix between surgeons and the impact this has on expected survival rates.

Table 4.6

Case mix imposed on data for simulation exercise (100 cases per surgeon)

Surgeon	Age <65 Dukes' B	Age >65 Dukes' B	Age <65 Dukes' C	Age >65 Dukes' C	Age <65 Dukes' D	Age >65 Dukes' D	Expected Surv %
1	13	29	8	15	23	12	67.8%
2	15	38	12	12	10	13	72.4%
3	4	32	8	16	12	28	64.1%
4	15	23	8	16	12	28	67.0%
5	15	21	3	20	11	30	64.7%
6	16	36	10	18	2	18	72.7%
7	10	42	5	13	10	20	69.1%
8	15	37	9	22	7	10	73.6%
9	13	30	4	13	9	31	65.5%
10	12	39	9	18	10	12	71.8%
11	15	40	8	14	5	18	72.0%
12	20	30	4	20	10	16	71.2%
13	20	28	3	29	6	14	72.4%
14	4	41	0	18	11	26	65.3%
15	10	39	4	19	12	16	69.4%
16	12	32	9	24	2	21	70.3%
17	6	26	10	20	14	24	64.7%
18	17	28	15	21	4	15	72.7%
19	15	51	7	8	8	11	74.5%
20	4	28	8	40	8	12	69.0%

4.6 The number of cases to incorporate

The principal problem with simulation exercises is determining the range of potential issues that one wishes to explore. There are however a number of practical considerations which frame the limit of any simulation based project.

Firstly, the number of surgeons, hospitals and cases should be reasonable having regard to the actual circumstances of potential audit studies. In particular an audit study which takes many years to accumulate data is essentially worthless since there is no ability to close the audit 'loop', feeding back information on performance to the surgical practitioners and the institutions within which they work.

Regard must be paid to the time it takes a practised surgeon to accumulate a reasonable case load. The table below illustrates the time taken to accumulate 100 patients undergoing surgery for colorectal cancer based on the five most active surgeons in the CRAG West of Scotland set of data.

Table 4.7

Period of Time Taken to Accumulate 100 Cases

(The 5 most active surgeons in the CRAG study)

Surgeon	Cases	Start	Finish	Years for 100 Cases
5	94	11/12/90	13/12/94	4.26
18	80	14/1/91	4/12/94	4.86
21	98	14/1/91	30/12/94	4.03
31	80	3/1/91	30/11/94	4.88
38	96	14/1/91	2/1/95	4.13

The emphasis in the above paragraph is appropriate. We are primarily interested in a simulation study in detecting Type 1 errors (the situation when, using the statistical methods at our disposal, we detect a surgeon (or hospital) as having divergent performance when in fact this does not exist *a priori)* and Type 2 errors (the inverse of the power to detect divergent performances when we know, *a priori*, that they do exist).

If the conclusion of a simulation study is that there is a considerable risk of Type 1 errors where case load numbers are high and constant over surgeons in the simulation then this conclusion can reasonably be generalized to the more realistic situation where some surgeons operate on quite low numbers of patients.

We base our simulation study on 100 cases per surgeon. Since for specialist surgeons it takes 4 years or more to accumulate such a number of (operable) cases it would be inappropriate to use higher numbers solely because they would increase the statistical power of our study. Realism must be a key feature of our case study.

We should also try to build in other reasonable features present in actual data. In particular it would be unreasonable to assume all surgeons are identical in performance. That said, for a full exploration of the underlying complexity of the simulated data it is a valuable exercise to look at a variety of scenarios which gradually build up the realism of the basic data setup when viewed in terms of variability of case mix and performance. We should try to incorporate some measured and justifiable variation in performance between surgeons. This is discussed further below.

4.7 The data and modelling scenarios considered in the simulation

With the general objective of examining Type 1 errors and power calculations in mind we reviewed the following 16 blocks of simulation exercises. In effect we are attempting to progress to a realistic 'null' situation, in statistical terminology. When we impose case mix and performance variability we are not doing so to the extent that differences are introduced of the orders of magnitude that audit scientists would be interested in detecting. Instead we attempt to introduce a realistic amount of 'background variation', modest levels of surgeon to surgeon variability that we might reasonably expect to observe in practice.

Table 4.8

The data set up and modelling scenarios considered in the simulation exercise

Scenario	Casemix	Performance	Model Type	Allowance for Casemix
Band 1				
1	same	same	fixed effect	ignored
2	same	same	fixed effect	allowed for
3	same	same	random effect	ignored
4	same	same	random effect	allowed for
Band 2				
1	same	differs	fixed effect	ignored
2	same	differs	fixed effect	allowed for
3	same	differs	random effect	ignored
4	same	differs	random effect	allowed for
Band 3				
1	differs	same	fixed effect	ignored
2	differs	same	fixed effect	allowed for
3	differs	same	random effect	ignored
4	differs	same	random effect	allowed for
Band 4				
1	differs	differs	fixed effect	ignored
2	differs	differs	fixed effect	allowed for
3	differs	differs	random effect	ignored
4	differs	differs	random effect	allowed for

This provides a progression through over simplistic scenarios such as, " all surgeons have the same performance for a given case, all have the same case mix and we will analyse the data using a conventional fixed effect model" to scenarios where case mix varies and random effect models are used in the analysis.

We proceed to examine both Type 1 errors and Power calculations for a given effect, namely that a relative risk of a particular surgeon or institution is a fixed multiple of the mean effect of his or her peer group.

The determination of the extent of case mix variability depended on the 20 sampled sets of case mix from the CRAG data and the variation in performance was calculated by adjusting the base scale parameter by a multiple of 0.85 rising uniformly in a deterministic fashion to 1.15 (i.e. in steps of 0.015). The objective of this exercise was to impose a measure of reasonable 'background' variation in performance, since in observed studies not all comparator groups of surgeons perform identically. Given the skewed distribution of the relative risk measure a non-uniform distribution might have been felt more appropriate but for the range involved it was felt that the progression of relative risks described above was adequate (for a simulation study).

Before documenting and commenting on the results of our analysis the following general point is worthy of emphasis.

The conclusions of the simulation study are likely to be relatively robust with respect to minor modelling issues (e.g. the selection of covariates) and general modelling issues (e.g. the choice of fixed or random effects models). It would however be unwise to generalise the numeric inferences about errors and power calculations since these are specific to the particular survival experience of patients undergoing surgery for colorectal cancer. These inferences are however likely to be similar for other types of surgery where mortality in the two years following surgery is comparable to that where patients are undergoing surgery for colorectal cancer (and where caseload accumulates at a similar rate).

4.8 The mechanics of the simulation procedure

An SPSS .SAV file was created with 2000 records subdivided into 20 surgeons each with 100 cases. The twenty surgeons were nested within 4 hospitals. It is to be noted that this accords with the raw data analysed. Surgeons, with only small exceptions, tended to operate within one institution for the 4 year period it took to accumulate their caseload. It would of course have been be possible to analyse cross classified data as well where surgeons had moved between hospitals. The fields were as follows

Case	: 1 to 2000
Hospital	: 1 to 4
Surgeon	: 1 to 20
Dukes'	: 2, 3 or 4 (corresponding to Stages B,C and D)
Age	: 1 or 2 (corresponding to age<65 or age ≥65)
Shape	: The shape parameter of the appropriate Weibull model
Scale	: The scale parameter of the appropriate Weibull model
Scaledum	: A variable enabling us to alter the scale parameter if required
Time	: The survival time generated from a sample from the Weibull distribution with parameters Shape and Scaledum
Indic	: A binary variable representing survival (or not) to 730 days.
Cons1..20	: 20 indicator variables to separate surgeon 1..20 from the others
Hosp1..4	: 20 indicator variables to separate hospital 1..4 from the others

The scenarios above fall into 4 bands (Band 1 to Band 4). Each of these bands has a basic file format which is unchanged within the band. After setting the random number seed in SPSS to enable replication of results additional SPSS .SAV files can be produced successively using the following syntax commands.

```
COMPUTE time = min(730,rv.weibull(scaledum,shape)) .
EXECUTE .
COMPUTE indic = 0 .
EXECUTE .
IF (time=730) indic = 1 .
EXECUTE .
```

These files are then saved for analysis. Although laborious this procedure does enable a complete record to be accumulated rather than output alone. In particular it makes it easy to produce graphical output on survival curves since we retain the survival function and covariate information as well as the outcomes themselves.

Bands 1 and 3 differ from Bands 2 and 4 in that the scale parameters for a given age and Dukes' combination are different. With Bands 2 and 4 we require to modify the 'scaledum' parameters to allow for variations in the basic relative risk from 0.85 to 1.15 as discussed above. Again this is achieved in SPSS with the following syntax (using a relative risk of 0.9 across all categories as an example)

```
IF (surg=1) scaledum = scale*(1/0.9).
EXECUTE .
COMPUTE time = min(730,rv.weibull(scaledum,shape)) .
EXECUTE .
COMPUTE indic = 0 .
EXECUTE .
IF (time=730) indic = 1 .
```

Note : The scale parameter actually works inversely as discussed earlier, that is you divide by 0.9 as opposed to multiplying by 0.9 in SPSS.

There are then two types of statistical methods to use in the analysis. We can use a fixed effect model implemented in SPSS or we can use a random effects (multilevel) model using WinBUGS. These two methods were contrasted in Chapter 3. It would have been possible to use WinBUGS to fit the fixed effect models as well but this is a time consuming exercise (an individual analysis of one simulation can take 10 minutes on a fast PC) and also ignores the fact that in practice much data is analysed using standard packages such as SPPS and MINITAB and we are after all interested in making a practical comment on the likelihood of getting errors in a practical setting. In fact we would not expect a large difference in the results from fixed effect models fitted using SPSS or WinBUGS. In other simple examples in the thesis where case mix is ignored (Chapter 5, Section 5.7) we see that estimates from WinBUGS for mortality rates per hospital were effectively identical to those from a traditional analysis).

Considering then the fixed effect analyses these are obtained from the simulated data files using the following syntax

```
For z =1 to 20

LOGISTIC REGRESSION VAR=indic
 /METHOD=ENTER consz
 /CONTRAST (consz)=Indicator
 /PRINT=CI(95)
 /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

LOGISTIC REGRESSION VAR=indic
 /METHOD=ENTER age duke consz
 /CONTRAST (age)=Indicator  /CONTRAST (duke)=Indicator  /CONTRAST
 (consz)=Indicator
 /PRINT=CI(95)
 /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

The first block of commands calculates the relative risk and the 95% confidence interval for the relative risk for surgeon z ignoring case mix. The second block of commands repeats the procedure but allows for the effect of Age and Dukes' Stage. The syntax for the simulation of the survival times and the syntax for the associated analysis can be combined to produce output for relative risks alone without the intermediate step of documenting and saving the individual survival times. This method was used in some circumstances. This does however prevent specific examination of individual sets of simulated data when required. In addition as we need to extract the data to use in other software package (e.g. WinBUGS) this is clearly unhelpful.

4.9 Type 1 Errors - Fixed Effect Models

For each of the sixteen scenarios detailed above we determined the extent to which we observed a significant event, namely a confidence interval for the relative risk for any of the 20 surgeons which excluded unity. For fixed effect models the logistic regression syntax above computes the relative risk and the associated 95% confidence interval. For the more basic scenarios the significant results can almost be arrived at by inspection of tables of frequencies of deaths, and this helps check the accuracy of the regression routines.

As an illustration simulation 1 of scenario 1 within the Band 1 data setup produced a confidence interval for the relative risk of surgeon 20 which excluded unity. The corresponding frequency data are detailed below.

Table 4.9

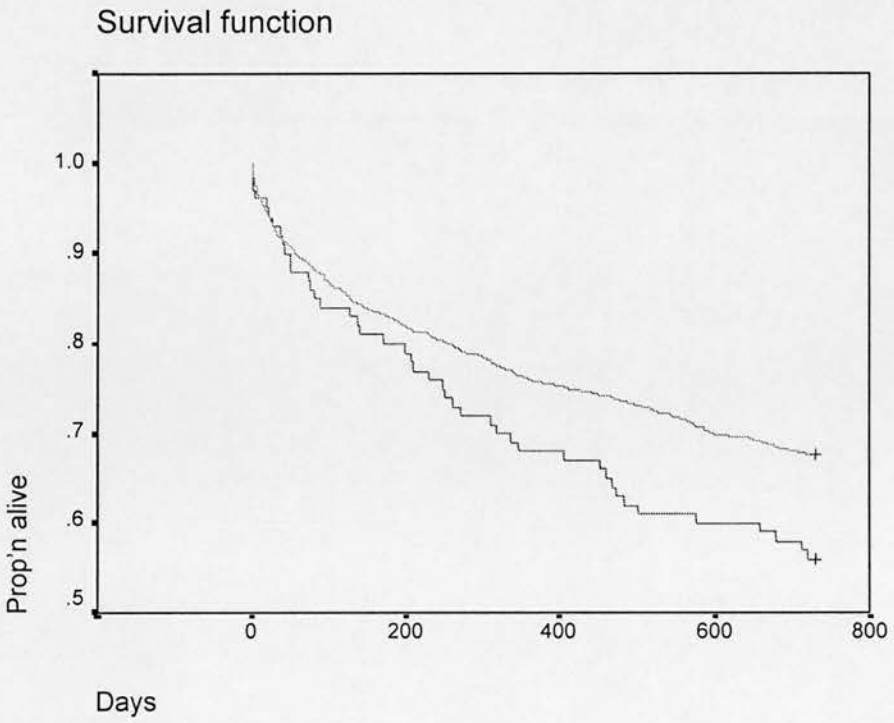
A sample realisation of outcomes from an individual simulation

Surgeon	Dead	Alive
1.00	28	72
2.00	29	71
3.00	27	73
4.00	27	73
5.00	30	70
6.00	36	64
7.00	36	64
8.00	31	69
9.00	39	61
10.00	29	71
11.00	33	67
12.00	38	62
13.00	39	61
14.00	28	72
15.00	40	60
16.00	25	75
17.00	32	68
18.00	36	64
19.00	33	67
20.00	**44**	**56**

The graph below shows the survival function to illustrate the extent of the divergence between surgeon 20 and the others in aggregate (for this single simulation).

Figure 4.10

The survival curves for a single simulation which gave a significant result



As further emphasis the log rank test gave a p-value of 0.018 based on the null hypothesis that the two survival distributions were the same (i.e. consultant 20 and 'the rest'). In simulation 1 of scenario1 we have detected a divergent performance for surgeon 20 when we know that *a priori* that he is no different from his peers. For this extreme example he has the same case mix and the same performance for a given case (that is the survival times for a given case are drawn from the same statistical distribution). We now illustrate the results of our investigation into Type 1 errors in the following table (based on 400 cases per scenario).

Table 4.10

Type 1 Errors for a number of data and modelling scenarios

Casemix	Performance	Model Type	Allowance for Casemix	PercentSignificant
same	same	fixed effect	ignored	3.5%
same	same	fixed effect	allowed for	3.5%
same	differs	fixed effect	ignored	6.8%
same	differs	fixed effect	allowed for	6.2%
differs	same	fixed effect	ignored	3.5%
differs	same	fixed effect	allowed for	3.2%
differs	differs	fixed effect	ignored	7.0%
differs	differs	fixed effect	allowed for	6.5%

There are several points of interest.

Certain of the scenarios are largely redundant – as an example the results for scenarios 1 and 2 will be almost identical since the allowance for case mix is of no practical benefit (since the case mix is identical). There may however be small amounts of additional variability introduced through the estimation of the case mix effects.

The number of Type 1 errors increases as complexity (i.e. realism) is introduced into the models. Most of the conclusions accord with prior expectations. As an example we would expect more Type 1 errors when case mix is different and when we are analysing the data without allowing for casemix. It should be noted that the standard error of the proportions involved is around 1%.

4.10 Type 1 Errors – Random effects models

The data on outcomes (in SPSS) were pasted into the data file used by WinBUGS via both Microsoft Excel and a text editor and the models developed in Chapter 3, Section 3.10 were run having satisfied ourselves as to the appropriate level of 'burn-in' and length of the Monte Carlo routine which gives stable results. This is a very time consuming process but it was thought unnecessary to examine all individual simulations since we know that the use of the random effects models is only likely to reduce the number of significant results. A reasonable compromise process is to only look at those simulations that produce significant results when using fixed effect models. The results were as follows.

1. None of the simulations in the Band 1 scenarios produced significant results.

2. The most extreme example of significance in the Band 2 scenarios came from simulation 1. Here surgeon 8 had only 8 deaths and surgeon 14 had 44 deaths. When performing the random effects modelling exercise we failed to achieve a level of significance in either case. When adjusting for case mix the confidence intervals for the relative risk for the surgeons were 0.92 to 1.70 and 0.64 to 1.09 respectively. Without adjusting for case mix the confidence intervals become 0.97 to 2.14 and 0.53 to 1.03 respectively. With WinBUGS, as described more fully before, we can also obtain a confidence interval for the rank orders of the surgeon and in neither case did we manage to place the surgeon in one half of the notional league table (with 95% confidence).

3. None of the simulations in the Band 3 scenarios produced significant results.

4. The most extreme example of significance in the Band 4 scenarios came from simulation 20. When performing the random effects modelling exercise we do note an isolated case with a level of significance when we do not adjust for case mix but this is lost when we do allow for casemix. When adjusting for case mix the confidence intervals for the relative risk for surgeon 3 was 0.49 to 1.04. Without adjusting for case mix the confidence interval was 0.45 to 0.99 The confidence interval for the rank order of the surgeon in question (developed using the methodology previously discussed) was 1 to 9, placing the rank order just in the top half of a notional league table. There were only two other instances of Type 1 errors, again only when looking at the performance measures excluding allowance for case mix.

In summary the conclusion of this part of our simulation exercise is that the use of random effects models as opposed to fixed effect models makes it much less likely that we will incorrectly infer that a surgeon's performance is significantly different from his or her peer group when we know *a priori* that the surgeon is in fact not importantly different. We also gain further evidence that it is crucial to adjust for case mix when analysing surgical audit data of a type where case mix effects are pronounced. The analysis could be made more complete in several respects

(a) We could increase the number of simulations examined

(b) We could vary the number of cases as well as the case load.

We feel however that there is no reason to expect the general conclusions to differ with these analyses and so have not performed them, although this could be an area for further work in future. In particular smaller case numbers per surgeon will increase the chance of more extreme results but this will be counterbalanced by an increase in the width of the confidence interval. The extent of shrinkage towards the mean effect in the random effects models will also be more pronounced for smaller case load surgeons. On balance we are satisfied as to the general reasonableness of our conclusions.

4.11 Power Calculations –Introduction and Fixed Effect Results

In this section we look at the converse of the Type 1 error analysis described earlier. We seek to assess the power of various statistical techniques to detect abnormal performance when we know *a priori* that it does exist. This is achieved by altering the scale parameter of the Weibull distribution to give a relative risk for a particular surgeon of a fixed multiple of his or her peer group. Clearly the power will depend on the magnitude of the effect we are trying to analyse and on the statistical distribution of survival times we impose on the data structure.

In contrast to the work on Type 1 errors we confined ourselves to the scenario most likely to be considered an approximation to a genuine audit study. The data take the form of 2000 cases allocated to 20 surgeons within 4 hospitals where there is a degree of actual variation in performance and a clear case mix difference between the surgeons. As before we looked at fixed and random effect models. Surgeon 10 (near the centre of the distribution of performance figures with a slightly easier case mix than average, see page 138) was chosen to have the Weibull scale parameter scaled to give a relative risk of 1.5, 2.0, 2.5 and 3.0. A similar exercise was performed for a group of 5 randomly chosen surgeons to look at similar hospital level effects.

The SPSS code below can be used to both simulate the survival times for the 2000 cases and perform the necessary logistic regression to determine whether the confidence interval for the relative risk for surgeon 10 excludes unity (both including and excluding allowance for case mix)

```
IF (surg=10) scaledum = scale/2 (for the relative risk of 2.0 example).
EXECUTE .
COMPUTE time = min(730,rv.weibull(scaledum,shape)) .
EXECUTE .
COMPUTE indic = 0 .
EXECUTE .
IF (time=730) indic = 1 .
```

```
LOGISTIC REGRESSION VAR=indic
 /METHOD=ENTER cons10
 /CONTRAST (cons10)=Indicator
 /PRINT=CI(95)
 /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

LOGISTIC REGRESSION VAR=indic
 /METHOD=ENTER age duke cons10
 /CONTRAST (age)=Indicator  /CONTRAST (duke)=Indicator  /CONTRAST
 (cons10)=Indicator
 /PRINT=CI(95)
 /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

The conclusions of the analysis for a run of 100 separate sets of simulations of 2000 cases are as follows.

Table 4.11

Power to observe surgeon effects of various magnitudes

(Fixed Effect Models)

Relative Risk	Power(no casemix)	Power(casemix allowance)
1.5	12%	38%
2.0	40%	72%
2.5	86%	94%
3.0	100%	100%

As an aside there was one particular simulation where the relative risk for surgeon 10 was set up to be 1.5 where surgeon 10 had a relative risk which excluded unity, but from below ! This was an example where we did not allow for case mix and highlights the low power in these circumstances very clearly.

As discussed in the introduction to this chapter the power calculations were repeated using Cox regression (with survival times censored at 2 years) as opposed to logistic regression with an outcome of two year survival. The results were very close to those for the logistic analyses reported above. For the purposes of answering the questions being specifically posed logistic regression is adequate (and extends more easily to random effects models) but clearly in a practical study with real data we may have a need to use survival analysis techniques to complement or replace binary logistic regression methods depending on the individual data issues involved.

The similar calculations for an idealised scenario where all the surgeons in a hospital have survival times which are realisations from Weibull distributions with similarly scaled parameters was as follows. Other scenarios could be considered where a hospital had a mix of risks with a similar net aggregate effect. We looked only at case mix adjusted figures since the pooling of case mix in a hospital level analysis gives very similar results, effectively the case mix variations are evened out in an aggregated study.

Table 4.12

Power to observe hospital effects of various magnitudes

(Fixed effect models)

Relative Risk	Power (casemix allowance)
1.5	78%
2.0	94%
2.5	100%

Given the results above there was no need to produce figures for the Relative Risk of 3 scenario considered at the surgeon level.

4.12 Power calculations – Random Effects Results

A further analysis was made of the surgeon level models using a random effects routine within WinBUGS. This was exactly the same approach which was followed for the Type 1 error investigation but more time consuming as unlike the previous scenario there are no obvious cases which do not need to be re-run. Using a case mix adjusted random effects model (previously documented) the power fell as follows :-

Table 4.13

Power to observe surgeon effects of various magnitudes

(Random Effect and Fixed Models)

Relative Risk	Power(random effects analysis)	Power(fixed effects)
1.5	14%	38%
2.0	52%	72%
2.5	82%	94%

4.13 Conclusions

Some care has to be taken when interpreting the above tables since there is dependence between the results of the 20 surgeons. Most of the effects observed accord with prior expectations although the scale of some of the results is perhaps a little surprising. In particular the power to detect known excess risks can be low.

1. We can see, as would be expected, that for a realistic data set up our power to detect a known effect increases with effect size and increases when we allow for case mix.

2. Introducing realistic case mix and modest surgeon to surgeon variability increases the Type 1 error rate

3. Random effects models are very conservative. Almost all the results highlighted as being significant (in the Type 1 error study) are removed when analysing the data using these models.

4. The additional complexity and prior assumptions (and a small amount of Monte Carlo simulation error) appear to lead to a reduction in the power when looking at random effects models. In effect this is the price we pay for running a reduced risk of inferring excess or underperfomence when it does not exist.

The main conclusions are that the risk of failing to detect poor performance (or good performance) is worryingly high at the level of risk ratio where we might have thought (in this example) from a clinical viewpoint that such underperformance or out performance should be detectable, say where the relative risk is 1.5 to 2.0. This cannot be generalised easily beyond the circumstances of the colorectal example we have reviewed in this thesis but the general conclusions seem soundly based.

There is no unique definition of what might constitute underperformance or outperformance. Ideally these should be set by clinicians before an audit commences and the magnitudes chosen will reflect the characteristics of the particular disease classification or operative procedure under consideration.

To achieve a reasonable power to detect excess performance it seems necessary to increase numbers of cases beyond the level which would be considered practicable for an audit study. This simulation was based around a four year accumulation of patients for those surgeons who dealt with the largest number of patients with these particular characteristics. It may then be a more realistic objective to detect aggregate effects such as hospital or regional effects in a larger study. It certainly would caution against early 'whistle blowing' at the surgeon level on the basis of mortality data alone. Both fixed and random effect models have a role to play in analysing audit data but the dominant concern is on the size of data sets. Patient numbers are just too small in many practical audits for reliable inferences to be

made. There is a clear danger that policy makers may use unreliable audit information to make important decisions in a range of areas. The limitations on the power of many audit studies should be fully appreciated before firm conclusions are drawn from such investigations.

Chapter 5 -A study of outcomes following paediatric cardiac surgery for the Bristol Royal Infirmary Inquiry

5.1 Introduction

The background to what has become known simply as the Bristol Inquiry into paediatric cardiac surgery was detailed in Chapter 1, Section 1.5 and a brief overview was given in that Chapter of the substantial problems which had to be overcome when analyzing the available surgical audit data from a statistical perspective. There were a number of key sources of data and this Chapter concerns itself with one particular set of data, that provided by the UK Cardiac Surgical Register (UKCSR). We were invited by the Inquiry to examine certain features of this particular set of data amongst others. We were to comment on data issues as well as to perform comparative analyses of surgical performance. In analyzing these particular data we did however make reference to other sources of data since a major difficulty faced by the statistical team looking at the evidence was the comparability of different sources of data on outcomes following paediatric cardiac surgery. We also contributed to the development of a consensus on the overall statistical approach which would take place in the more comprehensive analysis of data reported to the Inquiry (Spiegelhalter, 1999, Murray *et al*, 1999, Aylin *et al*, 1999). More recently part of our results were published in abbreviated form in the Lancet (Aylin *et al*, 2001).

Although not reported, considerable time was expended on reproducing elements of the many tables of excess deaths (and associated confidence intervals) in Dr Spiegelhalter's main synthesis both using his own Bayesian approach to the problem using the WinBUGS software and approximating these accurate results with more conventional statistical techniques. This was all part of an overall work programme aimed at maximsing the robustness of any conclusions drawn by the statistical team.

5.2 Objectives

Although other issues were considered and are reported elsewhere, with full details on the Bristol Royal Infirmary Web Site, we focus in this Chapter on the following main statistical issues.

- We investigate and comment on the basic structure and quality of the data provided by the UK Cardiac Surgical Register.
- We comment on the comparability of the UKCSR data (for certain time periods) with other sources, notably that provided from the Hospital Episode Statistics.
- We comment on the difficulties in mapping diagnostic and procedural codes.
- We compare the performance of Bristol with other English centres performing paediatric cardiac surgery. This is stratified by age, time and procedure.
- We attempt to identify whether the performance at Bristol was outlying in a statistical sense.

5.3 A basic description of the data

The UK Cardiac Surgical Register was an essentially voluntary (and anonymous) system run by the Society of Cardiothoracic Surgeons of Great Britain and Ireland. It was established in 1977. Each year cardiac units submitted summaries of their activities and the outcomes associated with this activity. A copy of one submitted form is included as Appendix 2. The data are subdivided by age for children under 1 year and for those over 1 year. As we shall see when discussing comparability with other sources, there was no upper age limit in the second classification as there was in some other sources of data examined by the Inquiry. Another major subdivision of the data was into 'open' versus 'closed' cardiac surgery. Open heart surgery is in this context defined as being a procedure where the heart is stopped and cardio-pulmonary bypass is required. Closed surgery is defined as being procedures where such a bypass is not required.

Within these four categories are numerous subgroups relating to the diagnostic and clinical features under consideration. Not all procedures were necessarily curative.

Importantly the data were not supplied on a case by case individual patient basis but were in fact aggregated over the various categories of interest. We have then to analyse the totals per category and the numbers of deaths associated with these individual numbers of cases. This is not ideal from a statistical perspective but was a fundamental feature of the data which could not be avoided.

5.4 The impact of data quality on results

Before considering our preliminary study of these data it is important to make some comments regarding the possible impact on results (and choice of statistical methods) of data quality. It is intuitively obvious that if the inferences on data analysed using statistical techniques are to be valid, and unbiased, then there has to be confidence in the underlying data. Poor data will lead to not only inaccurate point estimates of relative performance but, almost as importantly, inaccurate and (probably) understated estimates of statistical variability.

Where data quality is poor an attempt must be made to investigate the robustness of conclusions with respect to data issues. Since the analysis for the Inquiry was to be in the form of separate studies for pre-defined subgroups there was not a particular concern over omitted variables. It was a greater concern that cases were completely missing as opposed to elements of data being missing for a particular case. The statistics for centres were aggregated over patients and this could disguise a number of underlying features of the data. A greater concern was the comparability of different sources of information. In this case as opposed to using data from one source to improve or validate a second source two separate analyses were performed and the results compared.

Poor data can arise for many reasons. The forms in the Appendices to this thesis show the extent of the data collection exercise which can be involved in an audit of surgical practice. We cannot however infer from the forms themselves anything about the quality of the data as reported. We cannot, for example, say that the data were ever reviewed or checked. The individuals involved in their preparation may have been very differently trained and motivated in different institutions or geographical areas. Further than this the data as reported could have mistakes arising purely as a result of the coding process. There could be scope for multiple recording of deaths or procedures particularly relating to multiple admissions.

These problems can be particularly serious when data are reported in aggregate rather than on a patient by patient basis. There may also have been inconsistencies or changes in definitions or codes particularly when data have been collected over a large number of years. This is an ever present problem with any retrospective analysis of data.

These last two issues (the aggregation of data and the problems of coding) are particularly relevant for the Bristol Inquiry since these were particular problems which we encountered when analyzing the UK Cardiac Surgical Register data source.

A final important point is that whilst a very thorough attempt was made to validate the Bristol data by comparing it closely with many other sources of local data and by looking at individual case notes and records this could not feasibly be done to any extent with the comparator group of institutions.

There is a risk therefore that we are not always comparing like with like (i.e. the Bristol data have clearly been more extensively checked and cross referenced than was the case with the comparator group). A detailed review of data quality in the Bristol Inquiry is presented in Lawrence *et al* (2000)

5.5 Preliminary data analysis

On balance the data from the UK Cardiac Surgical Register were less than ideal. Some of the paper based returns had in fact been inadequately copied and stored and had physically degraded to such an extent as to render them useless. That said, the inferences from the UKCSR data for the Inquiry period did broadly concur with those from other sources. The absolute numbers diverged on occasions but the ratios of interest were more stable. The data were supplied to us in paper form and were then input into files compatible with Excel and SPSS. This coding was done directly without making even obvious modifications and subsequent examination of the files highlighted some immediate concerns. There were considerable reservations about data quality and this is discussed in some detail below as we outline the main steps of our data analysis.

We had to aggregate aspects of the data since some paediatric centres had more than one centre reference number. Bristol for example had three centre numbers that had to be pooled before analysis could proceed. The complete files contained both designated paediatric centres and non paediatric centres.

The initial approach we adopted was to analyse all data for all individual years which had been made available. The Inquiry did however require analysis for three main 'epochs' 1984-87, 1988-90 and 1991-95. These periods were defined by the Inquiry and were not chosen after reference had been made to any data or preliminary results. Between different data sources and within the UKCSR data source itself there were some problems with consistency in the period definitions since records for some periods dealt with calendar years and others with financial years (i.e. those ending on 31 March). There was concern about duplication or that some 'years' would contain less than 12 months data.

The data from the UKCSR were also much longer than other sources in terms of the number of years covered. As discussed earlier surgery was broadly classified into open/closed and under 1 year/over one year as well as by much more focused diagnostic groups.

Considerable effort was expended on finding and deleting duplicate records and in closely examining the missing data in the file. There were also records which we determined were in fact subsets of other records in the file and these were adjusted as appropriate. An example of this would be two records for the same unit one containing 3 months data and the other a twelve month period including the aforementioned 3 months. We were careful not to assume that a blank entry for deaths meant a zero return from that category. Where doubt existed the record would be deleted from the analysis, particularly in the open category where mortality is so much higher. By contrast blank returns for deaths in the closed over one group could more safely be imputed as being a zero entry. As ever with missing data problems we have to reach a compromise between introducing bias into estimates and deleting large quantities of potentially useful data.

Several files were created which covered various restrictions to the data (say to restrict to the Inquiry period alone).

Even at this very preliminary stage we were able to quickly explore the broad profile of mortality rates for Bristol and the other centres in aggregate. The figures below were periodically revised as further data problems were discovered and resolved. That said, this first early information on mortality rates is interesting in itself and also as a comparison with the more polished results that follow later in the Chapter. In the tables in the remainder of this Chapter any figures for the comparator group exclude the data for Bristol itself.

Table 5.1

The first calculation of comparative mortality rates on an annual basis

Year	Over 1 Year Open Surgery		Under 1 year Open Surgery	
	Bristol	Average	Bristol	Average
1984	0.083	0.051	0.149	0.209
1985	0.059	0.069	0.214	0.241
1986	0.112	0.075	0.250	0.221
1987	0.082	0.075	0.280	0.280
1988	0.081	0.067	0.379	0.216
1989	0.118	0.070	0.375	0.208
1990	0.168	0.040	0.304	0.162
1991	0.140	0.049	0.304	0.162
1992	0.032	0.057	0.151	0.126
1993	0.043	0.036	0.280	0.113
1994	0.078	0.016	0.219	0.130
1995	0.015	0.024	0.060	0.110

By inspection, and without even examining intervals of confidence around the point estimates detailed above there seemed to be preliminary concerns about the performance of Bristol in the late 1980's and early 1990's in the under one year of age category.

The mortality rates for Bristol appear to remain relatively stable against a background of clearly falling mortality rates in other English centres (for open heart surgery).

Another stage in our data cleaning exercise was to adjust the files to enable analysis by Inquiry epoch classification. We wished ultimately to be able to consider measures of excess mortality such as the relative risk (or excess deaths) and to determine confidence intervals around such estimates. We would wish to look at the trends in mortality rates over time and within Inquiry epochs. We wished to explore the distribution of rank orders of institutions in the Inquiry.

5.6 The performance of 'focus' and 'non focus' groups of institutions

The data provided by the UKCSR were much more extensive than was considered necessary for analysis by the Inquiry (and more extensive than could be provided by other sources such as the Hospital Episode Statistics). In particular it included many more institutions performing surgery than were to be considered in the final comparative analysis. As an example the raw data included hospitals in Scotland. In what follows those institutions falling into the group to be finally analysed by the Inquiry we will refer to as the 'focus group' (excluding Bristol) with the remainder being the 'non focus group'. The Inquiry specified, from general knowledge of the comparability of the institutions, exactly which institutions were to be compared with Bristol. Non specialist centres were excluded from the comparison.

The initial comparative analysis reported above then was refined by removing institutions not in the focus group. The tables below give the numbers in the two groups and some basic features of the data are apparent.

Table 5.2

A Comparison of the Focus and Non Focus Groups (over 1 year)

By Epoch and in Total

	Over 1 Closed Cases	Over 1 Closed Deaths	Over 1 Open Cases	Over 1 Open Death	Over 1 Open Mortality
Focus Group					
1984-1987	1584	28	3552	268	**0.075**
1988-1990	1087	25	3590	234	**0.065**
1991-1994	809	21	4737	212	**0.045**
Subtotal	3480	74	11879	714	**0.060**
1995	218	4	1208	35	**0.029**
Total	3698	78	13087	749	**0.057**
Non Focus Group					
1984-1987	1254	48	2778	200	**0.072**
1988-1990	405	1	1505	80	**0.053**
1991-1994	376	10	1571	101	**0.064**
Subtotal	2035	59	5854	381	**0.065**
1995	117	1	568	17	**0.030**
Total	2152	60	6422	398	**0.062**

Table 5.3

A Comparison of the Focus and Non Focus Groups (under 1 year)

By Epoch and in Total

	Under 1 Closed Cases	Under 1 Closed Deaths	Under 1 Open Cases	Under 1 Open Death	Under 1 Open Mortality
Focus Group					
1984-1987	2161	154	1389	286	**0.206**
1988-1990	1855	107	1881	333	**0.177**
1991-1994	1941	57	3233	407	**0.126**
Subtotal	5957	318	6503	1026	**0.158**
1995	627	12	1017	111	**0.109**
Total	6584	330	7520	1137	**0.151**
Non Focus Group					
1984-1987	1098	97	708	184	**0.260**
1988-1990	422	21	358	86	**0.240**
1991-1994	401	21	533	93	**0.174**
Subtotal	1921	139	1599	363	**0.227**
1995	145	8	187	30	**0.160**
Total	2066	147	1786	393	**0.220**

Table 5.4

A Comparison of Bristol and the Total Group (over 1 year)

By Epoch and in Total

	Over 1 Closed Cases	Over 1 Closed Deaths	Over 1 Open Cases	Over 1 Open Death	Over 1 Open Mortality
Totals					
1984-1987	2838	76	6330	468	**0.074**
1988-1990	1492	26	5095	314	**0.062**
1991-1994	1185	31	6308	313	**0.050**
Subtotal	5515	133	17733	1095	**0.062**
1995	335	5	1776	52	**0.029**
Total	5850	138	19509	1147	**0.059**
Bristol					
1984-1987	191	4	429	36	**0.084**
1988-1990	127	4	304	37	**0.122**
1991-1994	88	3	382	28	**0.073**
Subtotal	406	11	1115	101	**0.091**
1995	24	1	136	2	**0.015**
Total	430	12	1251	103	**0.082**

Table 5.5

A Comparison of Bristol and the Total Group (under 1 year)

By Epoch and in Total

	Under 1 Closed Cases	Under 1 Closed Deaths	Under 1 Open Cases	Under 1 Open Death	Under 1 Open Mortality
Totals					
1984-1987	3259	251	2097	470	**0.224**
1988-1990	2277	128	2239	419	**0.187**
1991-1994	2342	78	3766	500	**0.133**
Subtotal	7878	457	8102	1389	**0.171**
1995	772	20	1204	141	**0.117**
Total	8650	477	9306	1530	**0.164**
Bristol					
1984-1987	221	24	110	23	**0.209**
1988-1990	152	12	108	31	**0.287**
1991-1994	179	5	181	43	**0.238**
Subtotal	552	41	399	97	**0.243**
1995	54	0	50	3	**0.060**
Total	606	41	449	100	**0.223**

The percentage of patients falling into the focus group is higher in general and is higher still for open heart surgery in under one year old patients. Mortality rates are higher for the non focus group dealing with open heart surgery on over one year old patients and substantially higher in the corresponding under one category. The trends in surgical practice are towards use of specialized centres of excellence in a number of clinical and surgical disciplines and these data suggest the strong logic in such a move. Interestingly the observed performance was poorer in Bristol than in the non focus group in Epoch's 2 and 3.

As identified earlier there appeared to be some general descriptive evidence of trends in reported mortality rates. In open heart surgery for under one year old patients we summarise the previous tables with the following abbreviated table of rates of mortality by group and epoch.

Table 5.6

A summary comparison of Bristol, The Focus Group and the Non Focus Group

(Mortality rates for open surgery in children aged under one year)

	Epoch 1	Epoch 2	Epoch3
Focus Group	20.6%	17.7%	12.6%
Non Focus Group	26.0%	24.0%	17.4%
Bristol	20.9%	28.7%	23.8%

Although it was too early to tell, this early examination of data, faults included, did point towards the possible conclusion that Bristol performed poorly in this category in Epochs 2 and 3 showing none of the improvement in mortality rates seen elsewhere in other designated centres for paediatric cardiac surgery.

5.7 A preliminary analysis based on ranks

These types of statistical procedures were reviewed in earlier Chapters but for completeness we will outline briefly the background to this relatively recently developed method.

The problem with many institutional comparisons is that the confidence intervals for the relative performance measures overlap substantially. It can be inferred from this that the true, but unknown, rank order of institutions could vary quite substantially. This would be particularly the case where sample sizes were small or where the institution is more 'average'. The rank order itself can then be considered to be a random variable and the powerful and increasingly widely used WinBUGS software has enabled statisticians to consider the rank order problem in a more rigorous fashion. (Speigelhalter , 1995, 1996 (a), (b) and Marshall & Speigelhalter, 1998)

The underlying mortality rates for institutions can either be assumed to be completely independent of each other or they can be assumed to have an underlying similarity in some way. In this latter situation they can be thought of as being drawn from a population of some type with a particular mean effect. These two models are termed fixed and random effect models respectively and were considered in another context in previous chapters when we discussed surgical audit within the context of colorectal cancer data. In many cases, if patient numbers are high, these two methods do in fact produce similar results.

The WinBUGS software proceeds to estimate model parameters or functions of the parameters in a Markov Chain Monte Carlo simulation procedure (Gilks *et al,* 1996). At each iteration the sampled value of interest, say the probability of death, is noted for each institution. The rank order is then produced and stored. As the number of iterations increases we obtain an approximation to the distribution of the true but unknown rank.

What this adds to the analysis is as follows. In many applications comparisons purely based on divergence from the average can be misleading. An institution identified as being significantly different from average (in a statistical sense) may well not even be placed in the top or bottom quartile of the rank order (or 'league table').

The influence of sample size is also very significant since it greatly increases the power of the analysis (reducing width of confidence intervals). An institution with a superficially very good performance based on low case numbers might have a very high rank but associated with this may well be an extremely wide confidence interval.

As we demonstrated in Chapter 2, Section 2.10 substantial changes in rank may not necessarily be associated with changing performance. Institutional ranks can therefore be highly unreliable as a performance measure but the corollary of this is that these methods do assist greatly in identifying genuinely outlying or extreme performance. Evidence of outlying performance could be an institution which was ranked very highly or lowly and who, in addition, displayed a narrow confidence interval around the (extreme) mean rank.

At this stage of our analysis of the UKCSR data we made a preliminary examination of the distributions of rank orders using the methodology and WinBUGS software described in Chapters 1, 3 and 4 of this thesis. The following tables are for the main classifications of interest, namely open heart surgery for the three Inquiry epochs.

Table 5.7

The Rank Order of Bristol out of 12 Centres - Over 1 Year Open Heart Surgery

Time Period	Lower 2.5%	Median	Upper 2.5%	Number
1984-87	5	8	11	12
1988-90	9	12	12	12
1991-94	7	10	12	12
All times (+ 1995)	9	11	12	12

Table 5.8

The Rank Order of Bristol out of 12 Centres - Under 1 Year Open Heart Surgery

Time Period	Lower 2.5%	Median	Upper 2.5%	Number
1984-87	2	6	10	12
1988-90	7	10	11	11
1991-94	10	11	11	11
All times (+ 1995)	9	11	12	12

These tables exhibit some interesting features. Firstly Bristol performs averagely for both age groups in the period 1984-87. In later epochs however performance is poor with Bristol being ranked bottom for the under 1 group and 1991-1994 epoch. For this period the confidence interval is (10,11) with 11 institutions considered. In both 1988-90 and 1991-94 the confidence intervals include the extreme rank attainable from of the number of institutions involved.

The following tables and figures illustrate the mortality rates, confidence intervals and rank order distributions for all centres for the under 1 open surgery 1991-94 category. The 'average institution' shown in Figure 5.2 was the one having a median rank in the centre of the list and a symmetrical distribution about that median.

Table 5.9

Preliminary Mortality Estimates from WinBUGS Software

(Open Heart Surgery Under One 1991-94 Category)

Centre	2.5%	median	97.5%
1	0.1175	0.1684	0.2322
2	0.0927	0.1331	0.1825
3	0.1220	0.1678	0.2200
4	0.0741	0.1086	0.1509
5	0.0863	0.1211	0.1642
Bristol	0.1817	0.2398	0.3052
7	0.0897	0.1189	0.1523
8	0.0425	0.0717	0.1116
9	0.0947	0.1372	0.1894
10	0.1217	0.1557	0.1950
11	0.0719	0.0915	0.1173

Table 5.10

The rank orders based on early data (with 95% confidence intervals)

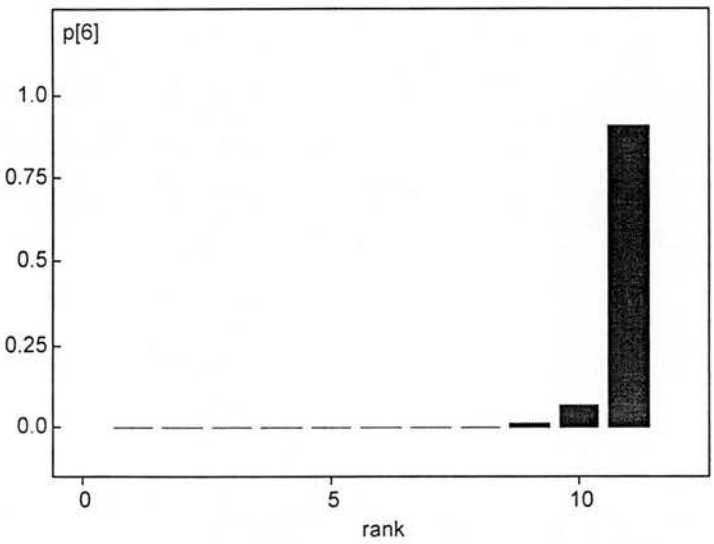
Open Heart Surgery Under One 1991-94 Category.

Centre	2.5%	median	97.5%
1	4	9	11
2	2	6	10
3	5	9	11
4	1	4	8
5	2	5	9
Bristol	10	11	11
7	2	5	8
8	1	1	4
9	3	7	10
10	5	8	10
11	1	2	4

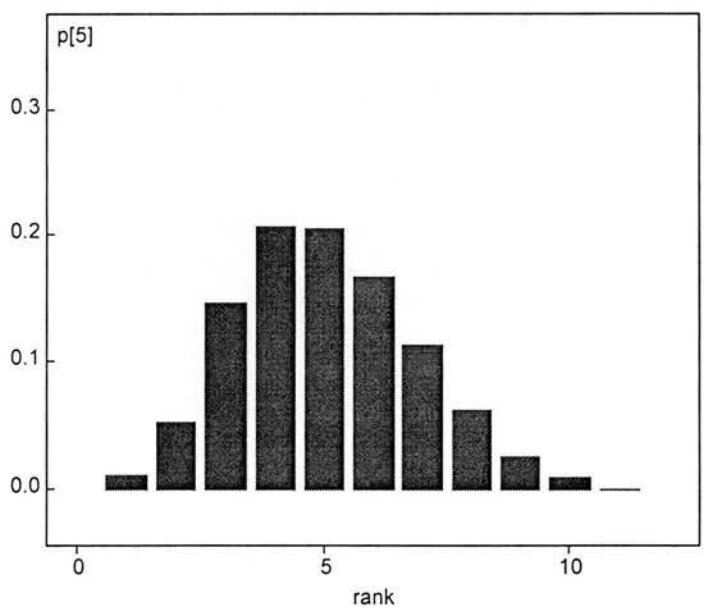
Figures 5.1 and 5.2

The rank order distribution of Bristol compared with an average institution

Bristol



An 'average' institution



It is worthy of note that the median estimate for the probability of death in the 1991-94 period from the WinBUGS analysis is 23.8%, which was identical to the arithmetic average obtained from the original crude data (see earlier). The confidence interval around this estimate was 18.2% to 30.5%.The lower 2.5% limit for Bristol is higher than the upper 2.5% limit for 5 of the other 10 institutions available at the time this preliminary calculation was made. This feature is of course what makes the confidence interval for the rank order so very narrow for Bristol. It is not just divergent from average but it is apparently outlying compared with the other institutions in a statistical sense. The histogram of rank orders created from the WinBUGS estimation procedure highlights this observation quite clearly.

At this stage in the analysis a major problem was identified with the data, namely that there appeared to be a substantial problem with the anonymous hospital codes for 1984. We identified some features of the data which led us to believe that there had been a renumbering of centres in the Register files between 1984 and 1985. This clearly casts doubt on the earlier preliminary inferences which we had made.

5.8 A discussion of coding problems and the case mix of Bristol

The next important data issue in the analysis was the mapping of the UK Cardiac Surgical Register diagnostic codes to the 13 procedure groups specified by the Inquiry. These groupings were selected to divide the cases into relatively homogeneous categories. The mapping exercise was necessary because the UKCSR data were based on diagnoses and not procedures. This is in contrast to other sources of data such as the Hospital Episode Statistics where clear procedures are identified (Aylin *et al,* 1999). The mapping exercise took place with the assistance of expert input from paediatric cardiac surgeons and cardiologists and was effected by the mapping of OPCS4 Procedure codes to the Inquiry Groupings and the associated mapping of UKCSR diagnoses to these consensus groups. Full details of the mapping procedures are contained in Murray *et al (1999).*

The coding was also intended to offer a grading of risks for the groupings under consideration. As the GMC Inquiry and earlier non statistical evidence to the Bristol Inquiry demonstrated, a particular class of operation, the 'Switch' was a prime focus of attention. One problem which then had to be considered was that although the basic diagnostic category remains consistent throughout the period the surgical procedure used to treat the patient could vary over time as new surgical techniques were developed and implemented across the specialized paediatric centres in the UK. Other sources of data were easier in this respect to categorize and for the UK Cardiac Surgical Register we will see that a particular procedure grouping could not be created from this source. In other areas the correspondence of the different groupings between different data sources varied from situations giving very close accord to those exhibiting quite wide differences. After checking data quality with procedures designed to highlight obvious coding and data entry errors we finally arrived at the stage where we had two definitive files. The first included the totals for the four broad categories of under and over 1 year old patients and open or closed procedures. The second subdivided these broader classifications into the 13 Inquiry specified groupings, where possible.

Although clinical opinion could differ the broad profile of risk in the procedure groupings was felt to be as follows.

Table 5.11

Synthesis of Statistical Sources: Primary Procedure Ranking

Rank	Group	Description
1	G 8	Truncus Arteriosus
2	G 9	Fontan type operations
3	G4	TAPVD
4	G 3	Other TGA
5	G 2	Interatrial TGA
6	G 5	AVSD
7	G 11	Mitral valve procedures
8	G 10	Aortic and pulmonary valve procedures
9	G 1	Tetralogy of Fallot
10	G 7	Closure of VSD
11	G 6	Closure of ASD
12	G 12	Closed Shunts
13	G 13	Simple Coarctation

As discussed above the UKCSR data did present a major difficulty in that it is not always easy to map the diagnosis to the procedure. Considerable effort was expended on this important issue and even then one category (12, the closed shunt) could not be constructed from the basic source information.

Table 5.12 below shows an initial profile of risks across the above groupings illustrating not just the variation in mortality rates by procedure but the pronounced age effect which is seen in the death rates for the very young. A particular source of interest to the Inquiry was Group 4 ' the switch'.

Table 5.12

The total numbers of patients and deaths in descending order of procedural risk

(1991-95)

Group	Over 1 Cases	Over 1 Deaths	Under 1 cases	Under1 Deaths	Over 1 Mortality	Under 1 Mortality
8	29	6	214	94	0.207	0.439
9	367	53	62	19	0.144	0.306
4	82	5	446	76	0.061	0.170
3	356	31	1225	186	0.087	0.152
2	35	2	74	19	0.057	0.257
5	435	57	572	96	0.131	0.168
11	237	30	65	11	0.127	0.169
10	1413	58	451	84	0.041	0.186
1	1751	110	419	40	0.063	0.095
7	1234	37	1518	79	0.030	0.052
6	3523	22	276	20	0.006	0.072
13	26	3	18	2	0.115	0.111

Table 5.13

The numbers of patients and deaths in descending order of procedural risk (Bristol)

Group	Over 1 Cases	Over 1 Deaths	Under 1 cases	Under 1 Deaths	Over 1 Mortality	Under 1 Mortality
8	5	0	15	9	0.000	0.600
9	59	16	1	1	0.271	1.000
4	5	1	49	16	0.200	0.327
3	50	6	93	18	0.120	0.194
2	0	0	0	0		
5	30	10	49	15	0.333	0.306
11	16	2	2	0	0.125	0.000
10	108	0	16	2	0.000	0.125
1	182	20	10	2	0.110	0.200
7	119	2	106	6	0.017	0.057
6	331	1	10	5	0.003	0.500
13	5	1	0	0	0.200	n/a

Table 5.14

The distribution of cases in the various categories of surgery

Group	Total Over 1	Bristol Over 1	Total Under 1	Bristol Under 1
8	0.3%	0.5%	4.0%	4.3%
9	3.9%	6.5%	1.2%	0.3%
4	0.9%	0.5%	8.4%	14.0%
3	3.8%	5.5%	22.9%	26.5%
2	0.4%	0.0%	1.4%	0.0%
5	4.6%	3.3%	10.7%	14.0%
11	2.5%	1.8%	1.2%	0.6%
10	14.9%	11.9%	8.4%	4.6%
1	18.5%	20.0%	7.8%	2.8%
7	13.0%	13.1%	28.4%	30.2%
6	37.1%	36.4%	5.2%	2.8%
13	0.3%	0.5%	0.3%	0.0%

As we can see the case mix for Bristol was not untypical of the country as a whole. There was some statistical evidence (an analysis of proportions) which showed certain divergences from average, most notably for groups 2 and 4 in the under one category. If however one groups the data into two broad classifications, namely 8+9 (the two high risk categories) and 'the rest' we see actual numbers of procedures for the under one's at Bristol which are very close to those expected on the basis of national statistics

Table 5.15

A comparison of case mix for Bristol

	Under 1 Actual	Under 1 Expected
8+9	16	18
The Rest	335	332

Examination of the data has therefore shown that the broad case mix experienced by Bristol was comparable with those seen in other designated centres of surgery. This is particularly helpful background knowledge when one examines the relative performance of Bristol in broad classes of operations which conceivably could mask quite substantial case mix related variability in expected survival percentages. As we have seen in earlier Chapters data either have to be stratified into clearly homogeneous groups prior to analysis or case mix has to be taken into account in the modelling process. The ordering of procedures by severity drew on expert clinical advice on the most common combinations of procedures and mortality rates. When one examines the observed aggregate mortality rates by procedure they are consistent with this clinical ranking.

5.9 A comparison of the UKCSR data with Hospital Episode Statistics

We were also provided with the Inquiry source data based on Health Episode Statistics. The responsibility for the analysis of these data lay with other statistical experts invited by the Inquiry but our function was to assess how closely the HES data corresponded with what we had been provided with by the UK Cardiac Surgical Register. The data needed to be recoded to map the codes correctly between the two sets of data and there were some inconsistent definitions of time periods to overcome, namely that the HES data consistently used financial years whereas the UK Cardiac Surgical Register data mainly used calendar years. The HES data also offered more subdivision with respect to age than was available from the UK Cardiac Surgical Register. In particular the HES data discriminated between those children over and under 3 months of age as well as over or under 1 year of age. To compare the two sets of data we then had to merge certain files in the HES data.

A preliminary analysis of the revised groupings for the UK Cardiac Surgical Register showed very good correspondence with the local Bristol sources, notably the South West Congenital Heart Register. The comparisons between the UKCSR and the HES data are given in Table 5.16 but fuller descriptions of the more extensive data comparisons are contained in our report to the Inquiry (Murray *et al,* 1999)

5.10 Finalised results and commentary

We have described above the broad process of data validation and analysis that we followed encountering and addressing many data inconsistencies and errors as we worked. Eventually having discarded the 1984 data since we were not confident that we could identify the correct centre from each form we were left with certain definitive files for the remainder of our analysis. These subdivided the data into Inquiry epochs, by open/closed and by procedure (where possible).

The remainder of this chapter is in the form of a series of tables of results based on the analyses of these 'definitive' files. The basic strategy involved (given that the data issues had been largely addressed at this stage) was as follows.

1. To tabulate the basic data being analysed and to compare these data between different sources both in aggregate and for Bristol alone. This exercise to be carried out for the various subgroups of interest (i.e. open/closed, by age, by year and by procedure group, where possible)

2. To calculate measures of relative performance between Bristol and Other Centres on a pooled basis and to place these estimates within an appropriate confidence interval. This exercise to be carried out for the various subgroups of interest (e.g. open/closed, by age, by year and by procedure group, where possible)

3. To examine the theoretical distribution of rank orders of the Centres involved in the analysis with a view to assessing whether Bristol was truly outlying in statistical terms as opposed to being 'merely' worse than average. This exercise to be carried out using different modelling approaches for a number of important subgroups of the data.

4. To relate our specific findings to those of other statistical experts involved with the Inquiry.

We have used below a measure of relative performance commonly used in reported mortality studies, the odds ratio (Altman, 1991). Associated with these point estimates of relative performance are 95% confidence intervals. It is to be noted that many of these estimates do not take account of the centre to centre variability, although this was examined by ourselves and other statisticians as part of the overall synthesis of data analysis (Spiegelhalter, 1999). We present our results below in the form of a succession of tables with associated discussions. These tables were included, in whole or part, in Murray *et al* (1999) and Aylin *et al* (2001). It is to be noted that in these tables Bristol Royal Infirmary is coded as Centre 1, and is occasionally abbreviated to BRI.

Table 5.16

Comparison of UKCSR returns with HES data for 1991-1994.

	Number of Cases			Number of Deaths			Ratio of Death Rates
	UKCSR	HES	Ratio	UKCSR	HES	Ratio	
Surgery							
Open	8227	7116	1.16	698	563	1.24	1.07
Closed	2898	2768	1.05	86	98	0.88	0.84
Total	11125	9884	1.13	784	661	1.19	1.05
Age							
Under 1	5360	4896	1.09	500	454	1.10	1.01
Over 1	5765	4988	1.16	284	207	1.37	1.19
Centre							
1	830	691	1.20	79	68	1.16	0.97
2	758	601	1.26	43	37	1.16	0.92
3	556	1049	0.53	50	53	0.94	1.78
4	295	359	0.82	27	27	1.00	1.22
5	664	544	1.22	61	39	1.56	1.28
6	1372	1306	1.05	96	80	1.20	1.14
7	819	633	1.29	40	32	1.25	0.97
8	1187	955	1.24	82	64	1.28	1.03
9	805	603	1.33	49	46	1.07	0.80
10	709	569	1.25	87	70	1.24	1.00
11	1921	1446	1.33	95	85	1.12	0.84
12	1209	1128	1.07	75	60	1.25	1.17
Group							
G1	921	810	1.14	57	46	1.24	1.09
G2	76	152	0.50	15	17	0.88	1.76
G3	685	561	1.22	89	70	1.27	1.04
G4	203	195	1.04	28	26	1.08	1.03
G5	553	758	0.73	65	73	0.89	1.22
G6	1525	1099	1.39	11	18	0.61	0.44
G7	1141	1249	0.91	26	59	0.44	0.48
G8	123	101	1.22	30	32	0.94	0.77
G9	340	616	0.55	42	67	0.63	1.14
G10	827	866	0.95	42	44	0.95	1.00
G11	160	224	0.71	15	22	0.68	0.95
G13	757	618	1.22	12	18	0.67	0.54
Year							
1991	3255	2576	1.26	254	184	1.38	1.09
1992	3403	2912	1.17	245	202	1.21	1.04
1993	2352	2270	1.04	142	144	0.99	0.95
1994	2115	2126	0.99	143	131	1.09	1.10
1995	3509	1982	1.77	195	134	1.46	0.82

The ratios above are the figures for the UKCSR source divided by the HES source.

Table 5.16 shows some over-reporting of both activity and deaths in the UKCSR data relative to the HES data. Since the HES data were only available for the period after 1991 the comparison is restricted to this particular epoch. This can largely be explained by the inclusion of adults in the former source and indeed for two years this extra subdivision was available and did very substantially explain the differences for those periods.

The ratios of interest are however more stable, particularly for the broader classifications. As an example for all cases UKCSR reports 13% more activity relative to HES but the ratio of death rates is 1.01. This is encouraging since despite some data discrepancies between UKCSR and HES (mainly in the over one categories) we are broadly satisfied as to the reporting of rates of mortality and ultimately it is the rates of mortality that are of particular interest for mortality comparisons.

There is considerably more variability at the centre level as one would expect with some centres giving quite divergent information. As an example Centre 3 reports only 43% of HES activity but records 94% of the deaths giving a ratio of death rates of 1.78. The final analyses performed for the Inquiry included sensitivity analyses addressing the issue of whether the centres where the HES and UKCSR data were discrepant were having a major impact on the conclusions (Aylin, 2001). It was not felt that data issues were likely to invalidate the conclusions drawn by the statistical team, including ourselves.

When one looks at the Inquiry groupings agreement between UKCSR and HES can again be seen to be poor and there appears to be a problem in the classifications for groups 2 and 3 (the latter being the controversial 'switch').

Table 5.17

Comparison of UKCSR returns with HES data for 1991-1994 ***for Bristol alone***.

	Number of Cases			Number of Deaths			Ratio of Death Rates
	UKCSR	HES	Ratio	UKCSR	HES	Ratio	
Surgery							
Open	563	451	1.25	71	61	1.16	0.93
Closed	267	240	1.11	8	7	1.14	1.03
Total	830	691	1.20	79	68	1.16	0.97
Age							
Under 1	360	295	1.22	48	47	1.02	0.84
Over 1	470	396	1.19	31	21	1.48	1.24
Group							
G1	58	47	1.23	6	5	1.20	0.97
G2	4	18	0.22	0	3	0.00	0.00
G3	45	19	2.37	10	11	0.91	0.38
G4	19	14	1.36	6	5	1.20	0.88
G5	41	34	1.21	11	12	0.92	0.76
G6	126	89	1.42	2	5	0.40	0.28
G7	90	93	0.97	0	1	0.00	0.00
G8	8	5	1.60	2	3	0.67	0.42
G9	39	37	1.05	7	5	1.40	1.33
G10	34	48	0.71	1	4	0.25	0.35
G11	9	21	0.43	0	3	0.00	0.00
G13	61	91	0.67	0	2	0.00	0.00
Year							
1991	215	185	1.16	30	22	1.36	1.17
1992	231	176	1.31	12	11	1.09	0.83
1993	202	169	1.20	20	18	1.11	0.93
1994	182	161	1.13	17	17	1.00	0.88
1995	264	170	1.55	6	3	2.00	1.29

Table 5.17 concerns itself with Bristol alone. It is to be noted that we were more confident about the Bristol data than for other centres (since additional validation was possible in an exercise not extended to other centres). Again the over-reporting of activity and deaths is present. The coding inconsistencies between G2 and G3 discussed earlier in this Chapter are apparent and there is the possibility that this problem could extend more widely. For the broader categories the ratios are more stable (0.97 for all surgery combined). At the individual group level we would expect greater variability given the small data numbers.

Table 5.18

Comparison of UKCSR returns with HES data for 1991-1995 ***for Bristol alone***.

		Number of Cases			Number of Deaths			Ratio of Death Rates
		UKCSR	HES	Ratio	UKCSR	HES	Ratio	
Category	Year							
Under 1, Open	1991	46	36	1.28	14	10	1.40	1.10
	1992	53	36	1.47	8	9	0.89	0.60
	1993	50	37	1.35	14	12	1.17	0.86
	1994	32	33	0.97	7	9	0.78	0.80
	1995	50	24	2.08	3	2	1.50	0.72
Under 1, Closed	1991	53	42	1.26	2	3	0.67	0.53
	1992	43	38	1.13	0	0		
	1993	49	40	1.23	2	2	1.00	0.82
	1994	34	33	1.03	1	2	0.50	0.49
	1995	54	31	1.74	0	0		
Over 1, Open	1991	93	84	1.11	13	9	1.44	1.30
	1992	94	72	1.31	3	2	1.50	1.15
	1993	93	74	1.26	4	4	1.00	0.80
	1994	102	79	1.29	8	6	1.33	1.03
	1995	136	87	1.56	2	0		
Over 1, Closed	1991	23	23	1.00	1	0		
	1992	41	30	1.37	1	0		
	1993	10	18	0.56	0	0		
	1994	14	16	0.88	1	0		
	1995	24	28	0.86	1	1	1.00	1.17

Again concerned with Bristol alone Table 5.18 subdivides the data into the individual years within the third epoch. This presents the analyst with a picture of the consistency of data issues over time, albeit within a single epoch. The same comments as above regarding over-reporting in the UKCSR apply but the ratios clearly become more volatile with smaller numbers being involved.

Table 5.19

Total UKCSR Congenital Activity 1985-1994 Split Open/Closed and by Consensus Group, for Under and Over 1's

	Aged Under 1 Year			Aged Over 1 Year		
	Cases	**Deaths**	**Death Rate (%)**	**Cases**	**Deaths**	**Death Rate (%)**
Surgery						
Open	6666	1088	16.3	11696	782	6.7
Closed	5878	297	5.1	3333	73	2.2
Total	12544	1385	11.0	15029	855	5.7
Group						
G1	455	45	9.9	1729	110	6.4
G2	199	34	17.1	79	4	5.1
G3	1303	203	15.6	355	31	8.7
G4	452	78	17.3	85	5	5.9
G5	587	103	17.5	427	56	13.1
G6	265	21	7.9	3276	19	0.6
G7	1552	80	5.2	1176	34	2.9
G8	239	101	42.3	32	7	21.9
G9	31	11	35.5	517	80	15.5
G10	538	105	19.5	1375	60	4.4
G11	67	12	17.9	251	33	13.1
G13	1177	30	2.5	814	3	0.4

Table 5.19 gives the aggregated mortality rates for all periods and for all centres in the focus group. It is split between those patients above and below the one year of age point. It highlights the features of main interest regarding the relative levels of risk in surgery on the very young and for open as opposed to closed heart surgery. The table gives a useful scale against which one can consider the numbers of cases in individual periods and the distribution of cases between the main diagnostic and procedural groupings. Importantly it offers a framework within which to look at the relative risk of different procedures. The aggregation of the data does however mask some important trends in both the types of procedure undertaken and the improvement that was seen nationally in performance over this long period.

Table 5.20

BRI versus All Other Centres Pooled by Epoch, Age and Surgery (Death rates, Odds ratios, 95% Confidence Intervals)

		BRISTOL			NON BRISTOL			ODDS RATIO		
Group	Epoch	cases	deaths	rate(%)	cases	deaths	rate(%)	estimate	95% CI	
Open <1	1985-1987	63	16	25.4	1308	275	21.0	1.28	0.67 -	2.34
Open <1	1988-1990	108	31	28.7	1863	336	18.0	1.83	1.14 -	2.86
Open <1	1991-1994	181	43	23.8	3161	395	12.5	2.18	1.49 -	3.15
Open <1	1995	50	3	6.0	1049	126	12.0	0.47	0.09 -	1.49
Closed < 1	1985-1987	154	18	11.7	1851	112	6.1	2.06	1.14 -	3.52
Closed < 1	1988-1990	152	12	7.9	1750	96	5.5	1.48	0.72 -	2.79
Closed < 1	1991-1994	179	5	2.8	1839	57	3.1	0.90	0.28 -	2.26
Closed < 1	1995	54	0	0.0	658	18	2.7	0.00	0.00 -	2.79
Open >1	1985-1987	284	24	8.5	2989	242	8.1	1.05	0.65 -	1.63
Open >1	1988-1990	304	37	12.2	3333	225	6.8	1.91	1.28 -	2.79
Open >1	1991-1994	382	28	7.3	4508	232	5.1	1.46	0.93 -	2.20
Open >1	1995	136	2	1.5	1305	42	3.2	0.45	0.05 -	1.76
Closed > 1	1985-1987	120	3	2.5	1293	21	1.6	1.55	0.29 -	5.32
Closed >1	1988-1990	127	4	3.1	1002	21	2.1	1.52	0.37 -	4.60
Closed >1	1991-1994	88	3	3.4	792	21	2.7	1.30	0.24 -	4.47
Closed >1	1995	24	1	4.2	233	3	1.3	3.33	0.06 -	43.17

Table 5.20 is of particular interest since it looks at the relative performance of Bristol and places a measure of uncertainty around the estimate. The finalised figures confirm the earlier conclusions in our exploratory phase of analysis namely that Bristol exhibited none of the improvement in mortality rates in the under one year of age category for open heart surgery that was more generally evident. Nationally, excluding Bristol, the death rate in the under one category for open heart surgery falls from 21.0% to 12.5% from 1985 to 1994. By contrast in Bristol the mortality rate remains relatively constant being 25.4%, 28.7% and 23.8% respectively in the three epochs. The odds ratios are 1.83 and 2.18 in epochs 2 and 3, both significantly above 1. Bristol is worse than average for these two periods and in this class of surgery. In other categories and periods the performance of Bristol is more average. The only other statistically significant results were in the closed/under 1/epoch 1 category and the open/over 1/epoch 2 category. In both cases Bristol was significantly worse than average.

Table 5.21

BRI versus All Other Centres Pooled, 1985-1994 (Death rates, Odds ratios, 95% Confidence Intervals) for Under 1's

	Bristol			Non-Bristol			Odds Ratio		
	Cases	Deaths	Death Rate (%)	Cases	Deaths	Death Rate (%)	Estimate	95% CI	
Surgery									
Open	352	90	25.6	6332	1006	15.9	1.82	1.40 -	2.34
Closed	485	35	7.2	5440	265	4.9	1.52	1.02 -	2.20
Group									
G1	4	2	50.0	452	44	9.7	9.27	0.65 -	129.5
G2	11	0	0.0	189	34	18.0	0.00	0.00 -	1.92
G3	82	15	18.3	1222	188	15.4	1.23	0.64 -	2.24
G4	45	17	37.8	411	65	15.8	3.23	1.56 -	6.50
G5	49	15	30.6	538	88	16.4	2.26	1.09 -	4.46
G6	10	5	50.0	255	16	6.3	14.94	3.02 -	70.7
G7	91	4	4.4	1462	77	5.3	0.83	0.21 -	2.28
G8	16	9	56.3	223	92	41.3	1.83	0.58 -	5.99
G9	1	1	100.0	30	10	33.3	∞	0.05 -	∞
G10	18	4	22.2	527	102	19.4	1.19	0.28 -	3.90
G11	2	0	0.0	66	12	18.2	0.00	0.00 -	25.5
G13	72	1	1.4	1115	29	2.6	0.53	0.01 -	3.28

Table 5.21 deals solely with the under one year of age grouping. It shows data grouped by years from 1985 to 1994. Without this grouping the numbers in the diagnostic groupings are so small in many cases as to make the confidence intervals uninformatively wide. Again there is evidence that Bristol is a poor performer in the open heart surgery classification and that certain procedure groups in particular contribute to this performance. We would wish to express some caution with the Group based results since the data had to be constructed as opposed to being recorded by definition (as it was with other data sources).

Table 5.22

BRI versus All Other Centres Pooled, 1985-1994 (Death rates, Odds ratios, 95% Confidence Intervals) for Over 1's

	Bristol			Non-Bristol			Odds Ratio		
	Cases	Deaths	Death Rate (%)	Cases	Deaths	Death Rate (%)	Estimate	95% CI	
Surgery									
Open	970	89	9.2	10830	699	6.5	1.46	1.15 -	1.85
Closed	335	10	3.0	3087	63	2.0	1.48	0.67 -	2.94
Group									
G1	157	20	12.7	1575	91	5.8	2.38	1.34 -	4.04
G2	1	0	0.0	78	4	5.1	0.00	0.00 -	731.0
G3	43	5	11.6	315	26	8.3	1.46	0.41 -	4.19
G4	4	1	25.0	82	4	4.9	6.50	0.10 -	102.7
G5	29	10	34.5	399	46	11.5	4.04	1.57 -	9.76
G6	292	1	0.3	3032	18	0.6	0.58	0.01 -	3.67
G7	108	2	1.9	1069	32	3.0	0.61	0.07 -	2.46
G8	3	0	0.0	29	7	24.1	0.00	0.00 -	9.14
G9	53	14	26.4	464	66	14.2	2.16	1.03 -	4.34
G10	89	0	0.0	1305	61	4.7	0.00	0.00 -	0.89
G11	16	2	12.5	237	31	13.1	0.95	0.10 -	4.46
G13	109	0	0.0	731	4	0.5	0.00	0.00 -	10.2

Table 5.22 presents a similar picture as was seen earlier for the younger age group. The death rates are of course much lower for this older age group but Bristol again emerges as a poor performer in open heart surgery with some procedure groupings giving significant results (despite the reduction in data). One note of caution on the interpretation of the confidence intervals is that between centre variability has not been allowed for in this comparison. Any such allowance tends to make it less likely you will observe a statistically significant result. This is of particular importance if one is commenting on particular subgroups with low case numbers or where the level of significance was marginal.

At this stage for completeness we repeated elements of our analysis using random effect models as opposed to fixed effect models. This extended our analysis to (arguably) more realistic models by modelling between centre variability. It also enabled us to verify a sample of the results produced by other expert statisticians working for the Bristol Inquiry who adopted this modelling approach when calculating measures of relative performance, which were reported in the form of 'excess deaths' and an associated confidence interval. The WinBUGS code used to implement fixed and random effects models is undernoted together with a brief verbal description of the underlying statistical process being modelled.

Fixed effects model

```
{
  for( i in 1 : N ) {
    p[i] ~ dbeta(1.0, 1.0)
        r[i] ~ dbin(p[i], n[i])
  }
}
```

In addition to the above model specification there are further lines of code to specify the data being analysed (the number of records ('N' in the above code), the number of cases per record, the number of deaths per record and an identifier for the Centre) together with initial values for the stochastic elements being simulated.

The other parameters needed for the procedure to run are specified in the WinBUGS operating system itself, specifically the number of iterations involved in the MCMC procedure and the number of values to discard before preparing any estimates of probabilities, ranks or posterior distributions. You also have to specify the elements which you wish to examine from the perspective of rank orders (Spiegelhalter *et al,* 1999 (b)).

The prior assumption made in the WinBUGS code above is that the mortality rates are independent for each Centre (that is 'fixed effects') and we use an uninformative prior for the mortality rates. The numbers of deaths are modelled as a binary response through use of the Binomial Distribution.

Random effects model

```
{
  for( i in 1 : N ) {
    b[i] ~ dnorm(mu,tau)
    r[i] ~ dbin(p[i],n[i])
    logit(p[i]) <- b[i}
  }
popmean<- exp(mu)/(1+exp(mu))
mu ~ dnorm(0,1.0E-6)
sigma <- 1/sqrt(tau)
tau ~ dgamma(0.001,0.001)
}
```

The code for the fixed effect model was then extended to the model above. In this specification we assume that the mortality rates are not independent but have characteristics in common. This is known as a 'random effects' model. We then have to specify a non informative prior for the population mean (on a transformed scale) and the precision (the inverse of the variance). The mortality rates (on a logit scale) are assumed to be drawn from a normal distribution. This is the same model specification used to calculate the 'excess deaths' figures reported in the final synthesis. The effect of this modelling change, in general terms, is to introduce 'shrinkage' in the mortality rates towards the estimated mean effect. The shrinkage can be pronounced if data numbers are small for some centres as was evident when we considered similar types of models in Chapter 3.

As discussed earlier in this thesis the examination of rank orders and their distribution can complement the production of tables showing the divergence of Bristol from average. Rank order analyses potentially highlight the outlying nature of institutional performance more clearly. In general terms such 'league tables' demand considerable critical examination because of the extreme sensitivity of rank orders to fairly small alterations in mortality rates relative to average but the corollary of this is that if an analysis of ranks shows an institution to be at the 'bottom of the leaguers' with a very narrow confidence interval for the rank order then this could lend support to the contention that the institution was genuinely extreme.

Table 5.23 below (based on fixed effect models) shows the clear poor performance of Bristol. It is ranked bottom of a notional league table for open surgery in the under one's in 1991-94. It does not however place Bristol completely beyond the distribution of performance of other centres.

For a number of possible groups we illustrate in Table 5.24 the effect of changing to a random effects specification. In most cases the results are very robust with respect to the alteration in model specification (from a fixed to a random effects model) but where patient numbers are smaller in some procedural groupings the observed widths of the confidence intervals widen sufficiently to perhaps lessen the conclusions we may be tempted to draw from the analysis.

Table 5.23

The rank order of Bristol for various categories and the associated 95% confidence interval (for the fixed effects model).

			Rank of Bristol		
				95% CI	
Group	Period	Number of Centres	Estimate		
Under 1 open	**1985-1994**	12	12	10	12
	1985-1987	12	8	3	11
	1988-1990	12	10	7	12
	1991-1994	12	12	10	12
Over 1 open	**1985-1994**	12	11	8	11
	1985-1987	13	8	4	12
	1988-1990	13	11	8	13
	1991-1994	13	9	6	11
Under 1 G3	**1985-1994**	12	8	2	12
	1985-1987	11	2	1	9
	1988-1990	12	4	1	9
	1991-1994	12	11	6	12
Under 1 G4	**1985-1994**	12	11	9	12
	1985-1987	11	9	5	11
	1988-1990	12	10	7	12
	1991-1994	12	10	6	12
Under 1 G5	**1985-1994**	12	10	6	12
	1985-1987	10	8	4	10
	1988-1990	12	8	3	11
	1991-1994	12	11	6	12
Under 1 G8	**1985-1994**	12	10	3	12
	1985-1987	11	10	4	11
	1988-1990	11	7	1	11
	1991-1994	12	7	1	12
Over 1 G1	**1985-1994**	12	11	8	12
	1985-1987	11	9	5	11
	1988-1990	12	9	4	12
	1991-1994	12	10	5	12
Over 1 G9	**1985-1994**	12	10	6	12
	1985-1987	10	7	1	10
	1988-1990	12	10	7	12
	1991-1994	12	9	5	11

Table 5.24

A comparison of rank orders and associated confidence intervals using fixed and random effect models -1985 to 1994 inclusive.

group	Fixed or random effect	2.5%	median	97.5%
under 1 open	fixed	10	12	12
	random	9	12	12
over 1 open	fixed	8	11	11
	random	8	11	11
under 1 G3	fixed	2	8	12
	random	2	8	12
under 1 G4	fixed	9	11	12
	random	9	11	12
under 1 G5	fixed	6	11	12
	random	6	10	12
Under 1 G8	fixed	3	10	12
	random	1	8	12
over 1 G1	fixed	8	11	12
	random	6	11	12
over 1 G9	fixed	6	10	12
	random	5	10	12

Our final table 5.25 below shows for comparative purposes the results from the Inquiry analysis based around excess deaths (using a random effects model specification). These results mirror our own findings, expressed in terms of confidence intervals for the odds ratio and the rank order. Bristol is significantly poorer than average in the under 1/1991-95/open category which was much in focus at the Inquiry. The mortality rate in this period was approximately twice as high as for other centres with excess deaths of 19 reported for this period (confidence interval 2 to 32).

Table 5.25

Relative performance for Bristol – expressed in terms of 'excess deaths'

Source	Epoch	Children aged under 1 year Mortality in Bristol Deaths/cases (%)	 Mortality elsewhere Deaths/cases (%)	Exp	Excess
Open					
Cardiac Surgical	1: 1984-1987	16/63 (25)	275/1308 (21)	14.0	2.0
Register	2: 1988-1990	31/108 (29)	336/1863 (18)	22.3	8.7
	3: 1991-1995	43/181 (24)	395/3161 (12)	24.0	19.0*
	4: 1995-1996	*3/50 (6)*	*126/1049 (12)*	*6.0*	*-3.0*
Total	1984-1996	93/402 (23)	1132/7381 (15)	66.3	26.7*
Hospital Episode	3: 1991-Mar 1995	41/143 (29)	356/3176 (11)	16.9	24.1*
Statistics	*4: 1995*	*2/24 (8)*	*68/563 (12)*	*2.8*	*-1.8*
Total	1991-1995	43/167 (26)	424/3739 (11)	19.7	22.3*
Closed					
Cardiac Surgical	1: 1984-1987	18/154 (12)	112/1851 (6)	9.4	8.6
Register	2: 1988-1990	12/152 (8)	96/1750 (5)	7.9	4.1
	3: 1991-1995	5/179 (3)	57/1839 (3)	6.2	-1.2
	4: 1995-1996	*0/54 (0)*	*18/658 (3)*	*1.5*	*-1.5*
Total	1984-1996	35/539 (6)	283/6098 (5)	25.0	10.0
Hospital Episode	3: 1991-Mar 1995	7/153 (5)	78/1784 (4)	6.9	0.1
Statistics	*4: 1995*	*0/31 (0)*	*25/357 (9)*	*2.8*	*-2.8*
Total	1991-1995	7/184 (4)	103/2141 (5)	9.7	2.7

Source	Epoch	Children aged 1 to 15 years Mortality in Bristol Deaths/cases (%)	 Mortality elsewhere Deaths/cases (%)	Exp	Excess
open					
Cardiac Surgical	1: 1984-1987	24/284 (8)	242/2989 (8)	23.3	0.7
Register	2: 1988-1990	37/304 (12)	225/3333 (7)	22.4	14.6
	3: 1991-1995	28/382 (7)	232/4508 (5)	22.8	5.2
	4: 1995-1996	*2/136 (1)*	*42/1305 (3)*	*4.4*	*-2.4*
Total	1984-1996	91/1106 (8)	741/12135 (6)	72.9	18.1
Hospital Episode	3: 1991-Mar 1995	21/314 (7)	194/4211 (5)	15.0	6.0
Statistics	*4: 1995*	*0/87 (0)*	*31/695 (4)*	*3.7*	*-3.7*
Total	1991-1995	21/401 (5)	225/4906 (5)	18.7	2.3
Closed					
Cardiac Surgical	1: 1984-1987	3/120 (2)	21/1293 (2)	2.0	1.0
Register	2: 1988-1990	4/127 (3)	21/1002 (2)	2.6	1.4
	3: 1991-1995	3/88 (3)	21/792 (3)	2.5	0.5
	4: 1995-1996	*1/24 (4)*	*3/233 (1)*	*.3*	*0.7*
Total	1984-1996	11/359 (3)	66/3320 (2)	7.4	3.6
Hospital Episode	3: 1991-Mar 1995	0/89 (0)	15/893 (2)	1.7	-1.7
Statistics	*4: 1995*	*1/28 (4)*	*0/111 (0)*	*0.0*	*1.0*
Total	1991-1995	1/117 (1)	15/1004 (1)	1.7	-0.7

Note : '*' denotes a statistically significant result.

5.11 Summary

We will summarize this Chapter briefly focusing on two main topics of interest, data quality issues and conclusions on comparative performance.

Data Quality Issues :-

1. Due to certain prior data considerations it was never going to be possible to map the UKCSR congenital activity data onto other sources in detail. As an example operative categories for the UKCSR had to be implied from the data as opposed to being observed. The UKCSR also covered a much longer time period than other sources and included patients over 16 year of age.
2. Despite the above comments, and particularly when one considers Bristol alone, the overall inferences drawn about relative performance were considered to be relatively robust with respect to data quality issues. Despite over reporting of activity and deaths by the UKCSR the mortality rates observed from different sources were stable in the larger categories of interest.
3. A satisfactory level of correspondence between data sources was evident in open heart surgery in under one year old patients and the data were further validated by comparisons with several other sources of information (not reported in this thesis but fully described elsewhere).
4. A limitation in the analysis was the inability to segregate cases into procedure Group 12, the closed shunt.
5. Data analysed for Bristol, not just as a consequence of the Inquiry itself, were much more comprehensive than for other institutions. Whilst we would hope that this does not introduce bias into the results there have been anecdotal concerns expressed that 'the greater the record keeping the poorer the results'. Bristol in fact did keep better records than most and although our conclusions, and those of the other statistical groups involved are emphatic in stating that Bristol was a poor performer and close to being an 'outlier' in some periods for some categories of surgery there is a risk that our conclusions are at the more pessimistic end of a range of possible actual scenarios.

6. There are some concerns about the use of the standard 30 day mortality definition for surgical outcomes. Indeed there have been vocal condemnations of this definition by parents of children who survived beyond 30 days but who then either died shortly afterwards or who continue to suffer severe long term impairments. One particular problem was that in-hospital mortality was reported as opposed to 30 day mortality (in or out of hospital) in some instances. Looking forward more extensive definitions of negative outcomes and comparisons of morbidity as well as mortality would be appropriate.
7. If surgical audit is to become more formalized, and this seems inevitable following the Government's immediate response to the findings of the Bristol Inquiry, more comprehensive information and data systems need to be established and maintained. Statistical analysis will have to be comprehensive and make use of the most recent developments in statistical knowledge and software if inappropriate inferences are to be avoided. We would support development of local independent sources of data that can be linked to national databases.

Results :-

1. The early data analysis we performed highlighted concerns over Bristol's performance that were subsequently verified by more extensive investigations when the very many data issues had been addressed. Progressive developments in our statistical techniques, more properly describing variability and interpreting rank orders, did nothing other than reaffirm our initial impressions.
2. Performance was seen to be better in the 'focus' group of institutions. This supports the general trend towards dealing with complex surgery in 'centres of excellence' and in particular supports the specific recommendations on such issues as are to be found in the Inquiry Report.
3. Nationally mortality rates for open heart surgery in under one year old patients improved substantially from 1984 to 1994 but this improvement was not seen at Bristol where no improvement was in evidence.

4. We were satisfied that data issues would not unduly invalidate any conclusions we would draw from our analysis.
5. Bristol performed poorly relative to other institutions in Inquiry Epochs 2 and 3 and was clearly ranked at or close to the bottom of the 'league table' of hospitals involved with surgery of this type. Performance was particularly poor in open heart surgery on under one year old patients and in specific procedure groups. These results were largely confirmed by other statistical experts working for the Inquiry using other sets of data.

Chapter 6 - Summary and Recommendations for Further Research

6.1 Summary of Previous Chapters

This thesis has been primarily concerned with the analysis of data although we do review and apply recent developments in statistical theory. In particular we have examined data from the following studies:-

1. Outcomes following surgery for colorectal cancer in the Glasgow Royal Infirmary in the period 1974 to 1984.

2. Outcomes following surgery for colorectal cancer in several hospitals covering a larger number of surgeons in the West of Scotland between 1991 to 1995.

3. Numerous sets of simulated data on outcomes following surgery for colorectal cancer based on models derived from the actual experience observed in the above studies.

4. Outcomes from paediatric cardiac surgery in various specialist centres in England and Wales from 1984 to 1995. These data formed a meaningful part of the basic information used for the statistical input to the Bristol Royal Infirmary Inquiry.

Several issues have arisen from our research. It has become clear that data quality is often a concern in the audit of surgical outcomes. If the audit process is to be of more than purely historical value it should be based around studies which produce feedback to practitioners, and the institutions within which they work, in a timely fashion. The data collected must be extensive enough to provide sufficient power in the statistical analysis. This inevitably takes time. If however the period of data acquisition is very long then the data may lose relevance as surgical practice alters over time. In a practical setting a pragmatic compromise may often be required.

Given the importance of inferences made on surgical performance the data analysed should have been accurately and consistently recorded. Regrettably this is not always the case.

Some of the data we analysed were well documented with considerable time having been invested in the basic design aspect and in collecting, validating and recording the data. The Glasgow Royal Infirmary data are notable in this respect but there was a clear research interest in the data gathering exercise in this example with specific funding being available to employ an individual whose sole responsibility was to collect reliable audit data.

Another issue when looking at institutional effects is that quality of data recording is not necessarily uniform across institutions. In the Bristol Inquiry it appears that the Bristol Royal Infirmary itself kept better records than other institutions. Missing data is a major statistical concern when analysing surgical audit data. An analysis of only complete cases discards useful information and, more importantly, is liable to introduce bias. Moreover imputation methods tend to underestimate statistical variability leading to spuriously precise inferences being made.

The effects of the recommendations that have emerged from the Bristol Inquiry have yet to be felt but it seems clear that closer examination of surgical performance will be inevitable. It is to be hoped that more structured audit programmes and methodical data collection do not create an unnecessarily risk averse culture amongst surgeons. A problem is of course that many audit studies consider operated cases alone. Care has to be taken not to penalise the 'adventurous' surgeon although adequate allowance for case mix should alleviate some of these concerns. The data concerns discussed in this thesis may well be lessened over time as more high quality information becomes routinely collected with a view to it being scrutinised by clinical and statistical practitioners.

The CRAG data were also of good quality and reflect the substantial commitment of the sponsoring institution and funding agency to audit exercises. That said, considerable effort had to be expended in obtaining a working file of data to analyse. The data recorded should contain enough information to enable statistical analysis to be undertaken using a subset of the broad covariates of interest but should also allow detailed examination of more minor (but often interesting) clinical features. Inevitably this may lead to 'over-recording' of data, a time consuming exercise. This demands allocation of resources, not something that has always been a high priority in the face of heavy financial demands on all aspects of health care in the UK.

On occasions the data presented to the statistician can be daunting. The principal reason for this is that a host of patient and institutional effects are recorded many of which are of little statistical interest in a general analysis examining large clinical or surgical effects. It is debatable however how much of this information could be ignored in a data recording exercise. Whilst the location of a tumour might not be of direct interest to the statistician modelling institutional performance (being swamped by more powerful predictors of outcome) this additional information is useful for subsidiary analyses looking at, for example, whether more complex surgery offers a better chance of discrimination between surgeons. There is evidence that experience increases performance and records of the seniority of the principal surgeon, the assisting surgeon and the anaesthetist are all useful items of data to record. In summary it would still be advisable to record data as fully as possible. Audit exercises can then focus on a subset of important covariates to produce rapid output and the more comprehensive data can be used as a basis for a more leisurely analysis of other aspects of clinical interest.

The Bristol Royal Infirmary Inquiry has highlighted these issues very clearly. The Inquiry has shown the difficulties which arise from the analysis of different sources of data covering the same outcomes, the problems of linkage and classification and the problems of missing or incomplete data. The raw data collected for one part of the data analysis (the UK Cardiac Surgical Register) were paper based and required a very substantial effort in data cleaning and classification before being able to be analysed with any confidence. It is doubtful whether any audit exercise without the financial resources of the Bristol Inquiry would have been able to perform such a thorough exercise.

One of the residual concerns with data in the Bristol Inquiry is that the data were more comprehensive for Bristol itself. There were additional sources of data which were used to validate the main sources of data (e.g. the South West Register and Surgeon's Logs). This exercise was not extended to the other centres (and in many cases could not have been even if this had been required) and could potentially have been another source of bias in the institutional comparisons that were made. The conclusion of the analysis was however that data quality issues did not cast doubt on the principal conclusion, namely that paediatric cardiac surgeons at Bristol Royal Infirmary performed particularly poorly in one epoch and that this performance could not be explained by statistical variation or systematic bias.

The statistical conclusions of our data analyses and simulation work are as follows:

The longditudinal analysis of the GRI colorectal cancer data enabled us to conclude that observed performance does not remain stable over even quite lengthy periods of time and that statistical variation in the data can be sufficiently large that it swamps surgeon level effects. The rank orders of the same surgeons in two different five year periods varied substantially as they did when 10 years of cases were allocated at random to two equal sized sets of data. This observed conclusion accords with the theoretical expectations that can be seen from work undertaken on the distribution of rank orders when looking at institutional performance.

The same set of data also showed that the choice of explanatory variables used to model the clear case mix effects present with colorectal cancer can influence the results to a meaningful degree. This is a significant concern since the choice of explanatory variables is not always one which can be made by the statistician. Different studies might record different data. Regression procedures may also accept or reject different covariates in a manner which is sensitive to the particular set of data being analysed.

In a detailed surgical audit study it would be preferable to analyse data several times using different covariates (all with the aim of capturing the case mix effects) to examine the sensitivity of inferences to purely modelling issues. There is a responsibility on the statistician involved in audit work to report conclusions only after as thorough an analysis as possible. The objective may not necessarily be to identify poor performance (although this tends to be the feature of audit studies most often reported in the public domain) but to genuinely report in a timely fashion on all aspects of surgical performance. This process identifies good surgeons as well as those who are under-performing. It enables elements of good practice to be identified and highlights areas of concern where further training (or resources) may be required. As a by-product audit data, comprehensively recorded, can form the basis of clinical studies examining features of the disease process and the factors that influence survival.

The analysis of the CRAG data showed that many subtle features of data can be analysed in an audit study as well as the main effects. Apart from general insight into the factors influencing survival, the distributions of case mix and sample size issues, the analysis of the CRAG data enabled us to examine the appropriate duration of follow up for studies involving colorectal cancer. A cut off point of two years was chosen since there were concerns over the accuracy of data recording beyond this time but we also examined shorter term outcomes (survival to six months or one years) and the relative merits of Cox regression over logistic regression.

On balance the two year logistic and Cox regression models allowing for case mix were preferred. The main conclusion of the analysis was however that the use of random effect models extends our insight in several respects.

Available software such as WinBUGS enable one to introduce more realism into the analysis. We no longer have to assume that when comparing a surgeon with his or her peer group that they are all identical in performance. We can now assume that they have characteristics in common and are realisations from an underlying distribution with a mean effect. This introduces a measure of shrinkage towards the mean effect which can be pronounced for low case numbers. Surgeons identified as having performance which is significantly different from average when using fixed effect models can become non significant when the realism of the model is extended. An important by-product of an analysis using WinBUGS is that we can obtain information on the distribution of rank orders, which become available through the BUGS MCMC process. These provide information which extends our insight beyond point estimates of relative risks and confidence intervals. A surgeon might have marginal significance in terms of difference from the mean effect but might, with confidence, only be placed in the top (or bottom) half of the rank order.

Criteria on identification of discrepant performance can vary but as a guide we would wish to see a rank order placed in the top (or bottom) quartile. The key statistical issue arising from the Bristol Inquiry was that we do not seek to identify the institution that is 'worst' or 'best'. After all, someone has to be in these positions. In effect we seek to identify the 'outliers' in a statistical sense, those institutions whose measures of performance do not appear to have been drawn from the same distribution, let alone be in the tails of the distribution.

The work which we undertook for the Bristol Inquiry highlighted the clear divergence of Bristol's performance relative to the mean effect in many instances (procedures and epochs) but when examining the distributions of rank orders the insight is magnified. In some cases we had confidence intervals for the rank of Bristol which excluded the lowest rank. This has the natural interpretation that, for that class of procedure and epoch, we could say that Bristol appeared not to be the 'worst' institution. We have attempted to show that the use of random effects models and the examination of distributions of rank orders greatly assists in determining whether an institution is not just poor (or even the poorest) but whether it is genuinely extreme.

It is to be remembered that the Bristol Inquiry analysis proceeded by analysing institutional performance by procedure within certain epochs. It would have been possible to have performed a more all embracing analysis where the case mix was introduced as one of a number of explanatory variables. On balance however the need to give detailed comment on specific procedure groups and the ease at which the results could then be communicated to a non statistical audience outweighed the temptation to aggregate the data (increasing power) and analyse with regression methods including covariates.

The Bristol Inquiry was principally an analysis of surgical outcome data at the hospital level although comment on individual performance was of course made. It did however illustrate just how thorough the analysis must be before we can draw any specific conclusions. The majority of audit studies by contrast proceed with the benefit of fairly restricted levels of sponsorship or funding.

Our simulation work based around a model for colorectal cancer surgical outcomes was designed to explore certain issues mainly concerning the reliability of inferences made in this field. We developed parametric models which fitted well to the data analysed previously from the CRAG West of Scotland source. The models we produced had satisfactory adherence to the empirical survival curves and importantly gave theoretical survival proportions at two years duration which compared well with those actually observed.

We were able to impose a realistic amount of case mix variability in line with observed experience and to examine the theoretical likely variation in survival proportions which case mix variability imposes on the data structure. We drew the following broad conclusions from this study :-

1. Parametric models for survival can be developed which adhere satisfactorily to observed survival experiences. Suitably modified these models can be adjusted to account for case mix and thereby form a basis for a simulation study.

2. Fixed effect models do not always properly account for the complexity of the underlying process. They can increase the risks of Type 1 errors, an obvious concern in the current environment where litigation is common.

3. The power to detect surgical under-performance (or out-performance) is surprisingly low in some circumstances.

4. The use of random effect models reduces the risk of Type 1 errors but at the cost of apparent further reductions in power.

5. It is important to look at the distribution of rank orders as well as measures of relative risks.

On balance for an audit study of colorectal cancer with a realistic number of patients and two years of follow up it is unlikely that we could identify a poorly performing surgeon with a satisfactory degree of statistical confidence unless the under-performance was extreme (say a relative risk of over 2.5). A more realistic objective would be to reliably identify hospital level effects as opposed to surgeon level effects. Indeed, based on our analyses of the colo-rectal cancer data and the data from the Bristol Royal Infirmary Inquiry, this finding is likely to be true much more widely in surgical audit. This conclusion could be invalidated if hospitals consciously choose to introduce diversity in skills and practice in their surgical teams.

Audit exercises based on outcomes alone are likely to lack the sensitivity necessary to identify modest but important differences in performance between individuals, or to provide useful information on the trend over time of an individual's performance.

This thesis has shown that surgical audit is a less precise process than perhaps had been previously thought. To do the job properly can require the commitment of large financial and administrative resources. The Bristol Inquiry seems set to alter many aspects of clinical practice and audit is likely to be structured more soundly in future which should alleviate some of the data concerns we have had to deal with ourselves.

If audit results are to be communicated back to surgeons and hospitals within a meaningful and useful timescale the statistical input has to be considered very carefully in the light of our conclusions. The statistical evidence can only be one part of the audit exercise but it will remain an important aspect of the process.

In many ways the results of our work return one to some of the basic audit principles set out in the introduction to this thesis. Surgical audit by analysis of outcomes, however good the data and sophisticated the statistical methods used to perform the analysis, is not sufficient on its own to provide information and guidance to practitioners, the institutions within which they work or central NHS agencies. Given the need to provide feedback on performance within acceptable timescales the power of statistical analyses is inadequate to detect all but the genuinely outlying. If anything the work of this thesis serves as a warning against early 'whistle blowing' against surgeons, particularly where the numbers of procedures undertaken is modest.

Data quality will have to be improved and the analysis of outcomes will no doubt continue to proliferate but this should only be done as part of a general improved level of focus on all aspects of the traditional audit 'loop' (and its more recent extensions). If the statistical power to detect excess performance is as low as it has been shown to be in this thesis then there needs to be a parallel focus on process as well as outcome.

Indeed as we conclude work on this thesis a recent paper in the British Medical Journal (Mannion & Goddard, 2001) raises a number of related concerns about published outcome indicators and the relative merits of analysing process or outcome, from a clinical and administrative viewpoint.

The analysis of outcome data must then be used in a structured and informed way to assist in the appraisal and modification of the surgical process. Further than this the structure within which surgeons work may also need to be altered in a responsive way to reflect the results and recommendations of audit studies. In many ways these developments have already started.

In our research we made an attempt to focus on the practical issues facing statisticians analysing surgical outcome data. We did not attempt to review all possible statistical techniques and model extensions for analyzing audit data but we did note that many new techniques continue to be developed and refined. Given the fact that surgical audit exercises are invariably observational studies as opposed to controlled trials those methods which make an attempt to control for unobserved case mix variation and selection effects may be of particular importance in future.

The results of the work in this thesis, and the more extensively reported statistical analyses for the Bristol Inquiry, do however give some support to the contention that the main constraint on what can be achieved from an analysis of outcome data is not the fine detail of the statistical approach but the practical limit on sample sizes (and outcome events of interest) that applies in many instances. The statistical approaches used in the Bristol Inquiry ranged from preliminary descriptive analyses to highly sophisticated computer intensive methods. These methods came up with broadly the same answers. This should not discourage one from implementing complex models which fundamentally address known features of the data but an exploratory examination of data using cruder methods can, as it did with the Bristol Inquiry, help to emphasise the overall robustness of the findings which are reported. The fact that, in this particular case, the 'simple' methods came to the same conclusions certainly helped the statisticians communicate the results to a wider audience.

Had the more complex statistical approaches lessened the force of the conclusions (as they might well have done given the data concerns and low power evident from some of the work undertaken in this thesis) the exercise in communication would have been considerably more difficult.

The Bristol Inquiry was an enormously costly exercise and the extent of data checking and cross validation involved would be completely impractical to implement into ongoing audits of routine surgical practice in most disciplines. That said it seems likely that data quality will improve in future as more is collected with the knowledge that it will be required by audit specialists for comparative analyses of practice and performance.

Given these comments the current expectations of clinicians for audit exercises based on outcome measures alone may well be overoptimistic. Data concerns may well invalidate or at least lessen the strength of conclusions drawn from both large and small studies of clinical outcomes.

Looking forward it is important that audit specialists take considerable care in drawing inferences from small audit studies, those where data quality is poor, those where there are concerns about unobserved selection bias and even larger studies where account has not been taken of heterogeneity in the data structure. It is clearly important that policymakers do not place excessive reliance on studies of outcome alone in making decisions about resource allocation and the overall structure of clinical departments.

6.2 General comments on statistical modelling and topics for future research

Despite the concluding comments above on the data and sample size constraints which can exist in an audit exercise it is clear that the statistician should implement as appropriate model as possible when analyzing data. In particular when data is naturally hierarchical, as it is in an audit of comparative hospital and surgeon performance, then an attempt should be made to account for this feature and a comparison made of results accounting for this structure with those where no such attempt is made.

When implementing any statistical model regard should be paid to the underlying assumptions that are being made. A process of model checking and validation is essential. The Cox model is extremely widely used in the medical literature and one reason for this is that it makes no general assumption about the distribution of survival times, albeit constrained by the assumption that hazard rates are proportional. It should not be applied unless one can be satisfied that the base assumption of proportionality of hazards for different factors is a reasonable one. In assessing the validity of the Cox model modern software packages offer the statistician the facility to examine, for example, complementary log plots. Graphical methods such as this enable one to assess proportionality of hazards but when one is fitting a model with several covariates the process must be repeated many times stratifying with respect to the covariate of interest. In our analyses reported in Chapters 2 and 3 these model checking procedures were followed. We were satisfied as to the validity of the basic modelling assumptions made but some concerns were raised about the proportionality of hazards in analyzing colorectal cancer survival data with long periods of follow up.

Given our chosen strategy of focusing on practical issues as opposed to the fine detail of the statistical approach and the fact that the proportionality assumptions for our particular audit studies seemed reasonable we did not examine, more complex, alternative specifications of the Cox model in Chapters 2 and 3.

In other surgical disciplines where the aforementioned modelling assumptions were not adhered to sufficiently well more complex models would be worthy of further study, albeit with the knowledge that data quality and sample size considerations place a constraint on the inferences which can be made. A particular area which could be investigated would be use of stratified Cox regression models where the assumption of proportional hazards does not apply to particular groupings of cases but hazards are acceptably proportional within each identified group. It may also be of interest in future audit studies to investigate whether the incorporation of time dependent covariates in the model would help cope with departures from basic proportionality assumptions in more straightforward analyses.

The choice of covariates to include in a given model to explain case mix variability can be guided by past research in the field and a number of stepwise procedures should be compared before selecting a set of 'final' predictors of outcome. In a classical framework alternative models can be compared by examining the change in deviance which results from the addition of further covariates. This forms the basis of a test statistic which can be used examine whether there is evidence to support the rejection of a hypothesis that the fit of the two models is comparable. As with regression procedures in any field of applied statistics a compromise has to be found between goodness of fit and a parsimonious modelling structure. Future areas of study would include a more detailed review of techniques for model selection and comparison, particularly within a Bayesian framework where methods of assessing goodness of fit are developing rapidly.

Examination of residuals following fitting is important to confirm the validity of the assumptions made about the distribution of the 'error' term in, say, a logistic regression model. The influence on the derived model parameters of particularly influential cases in the analysis should also be reviewed in detail. The process of identification and examination of extreme or influential residuals may highlight particular cases or groups of cases where the model is a poor predictor and this may lead to model refinements or further clinical insight. This procedure is particularly necessary in smaller studies where a small number of observations may carry appreciable weight in the fitting process.

In Chapter 2 we examined several situations where the chosen outcome was 'death from a cancer related cause'. This raises the question of models which allow one to study the effect of excess mortality relative to a standard population suffering 'competing risks'. This may be of particular importance in audits of surgical performance in the elderly where there are accelerating mortality risks present purely from the ageing process alone.

In the analysis of actual data which was performed in Chapters 2 and 3 the information recorded for any one individual surgical episode was very extensive and, fortunately, was complete for the outcomes and covariates of particular interest. Where fields of information for these important covariates were missing case by case imputation from other items in the records was often possible. In addition the covariates used gave little scope for measurement error or recording bias. This situation may well not exist in other audit studies and in these situations particular consideration will need to be given to various methods for coping with missing data. In some circumstances data might be missing completely at random, being unrelated to the value which might have been observed. In these circumstances a 'complete cases' analysis will produce unbiased results albeit at the cost of inefficient discarding of other data. More importantly however there may be some situations where the missing data are not in fact random. In these circumstances imputation methods need to be used with considerable care. One might use mean value imputation (the missing value being replaced with the mean from the rest of the data) or one could use the conditional expected value given the data observed. In a classical context with data missing at random the EM algorithm is sometimes used to develop appropriate maximum likelihood estimates. As discussed before a major concern with imputation of missing values is underestimation of variability and this is especially relevant when considering relative assessments of institutional performance. It would be interesting to see extensions of the work in this thesis investigating audit problems where the extent of missing data was larger and the nature of 'missingness' such that explicit modelling to cope with this feature of the data was required.

Although not strictly a missing data problem it would also be of interest to review various audit scenarios past and present to consider whether certain key aspects of data have not in fact been recorded at all. There may be practical reasons for some items of data being unrecorded, say to prevent identification of patients or clinicians in a study preserving anonymity. The lack of inclusion of certain unobserved covariates may lead to bias in any resulting estimates. In some cases proxy variables may be used to account for the effect of unrecorded items of information and this would be a potential area of future study within an audit context. The use of methods involving instrumental variables to control for unobservable characteristics in the data may be applied more extensively in future where selection bias is considered likely.

In cases where the determination of final definitive outcomes take considerable time the use of shorter term proxy or 'surrogate' outcome measures may allow for more rapid assessment of performance. This might however be associated with a reduction in the overall quality of the inferences made. One of the functions of analysis of outcomes is to feed back information into the audit loop as discussed in Chapter 1 and this analysis of outcomes may well focus attention on aspects of process and structure at certain institutions. This in turn may enhance the statistician's knowledge of other aspects of the data. As an example this greater focus might give additional insight into important unobservable or qualitative characteristics of the data.

The final reported analysis of data for the Bristol Inquiry was performed for a number of stratified groups of patients with similar characteristics and was repeated for several distinct 'epochs' of interest to those involved. The longer time series of data available from the UK Cardiac Surgical Register did however enable some examination of the progression of mortality rates over time.

The problems of analysing longitudinal data using random effect models (in different terminology and specifications also known as mixed or multilevel models) have become more widely considered in medicine and the social sciences in recent years. In effect one can consider a series of outcomes for a hospital as being repeated observations on the same institution over time. There will then be positive correlation within the series. The analysis of repeated measures data is conceptually the same as the modelling of hierarchical data discussed in this thesis variously as multilevel or random effects modelling. The software MLWin and WinBUGS could both be used to analyse longitudinal medical outcomes data in appropriate circumstances. It may be of interest to consider whether the measured effects apparent in a cross sectional analysis for one year alter over a period of years. As clinical audit data quality improves in future and longer time series become available to statisticians one can expect greater attention to be focused on longitudinal methods of analysis.

Recent years have seen a rapid development in statistical techniques and software used to model situations where several aspects of dependence in the data may exist. This is a very positive development but care should be taken not to implement excessively complex models, particularly where sample sizes are small.

In summary there are a number of potential projects for further research that could follow on from the work undertaken in this thesis. The list below includes some of the topics discussed in this section together with other potential areas of study which arose as we progressed with our research.

1. There is a clinical expectation that the same level of discrimination between surgeons could be obtained from a restricted set of procedures (probably more complex surgery). This was briefly examined in Chapter 3, Section 3.8 but merits more rigorous analysis.

2. The variation in performance allowed for in the simulation study could be made more complex. As an example we could examine the power to detect poor performance if the poor performance was only confined to cases with Dukes Stage C or D.

3. The effect on results of using other parametric models could be explored.

4. The entire power calculation (and indeed the simulation process itself) could be calculated internally within the WinBUGS program where a random variable can be monitored which denotes whether a realisation of the simulation gives a relative risk in the upper 2.5% point of the density function for the relative risk.

5. As discussed above when using the Cox regression model to determine the measures of relative performance for surgeons in Chapters 2 and 3 we satisfied ourselves (using diagnostic checks and examination of complementary log plots) that the proportionality of hazards assumptions which underlie this procedure for the particular sets of covariates and durations of follow up we used were supportable. There was a concern however that at longer durations (beyond those reported in this thesis) that these assumptions might not necessarily hold true. Under such circumstances the estimates of the model regression coefficients (which are used to determine relative performance) will be incorrect as will any inferences made on the basis of the coefficients. Given the wide use of the Cox model in the medical literature a more detailed exploration of these issues (covering stratified models and time dependent covariates) would be a useful area of further research both within and outwith the audit area.

6. A greater study could be made of the choice of covariates to be used to explain case mix variability and the effect on audit results of using different combinations of factors. This could be related to a study of the impact of unobserved variables on assessments of relative performance.

7. Given the requirement to obtain results as quickly as possible it would be of interest to examine the advantages and disadvantages of using proxy variables and surrogate outcome measures within an audit context.

8. The results obtained from the random effects models using the WinBUGS software could be compared with those obtained using other hierarchical modelling software. This would be of particular interest for small sample examples.

9. A more thorough analysis of the problems of missing data in audit analyses could be undertaken, specifically how to allow for missing data in the modelling process as opposed to using imputation methods or analysing only complete cases.

10. It would be of considerable interest to analyse larger sets of data with additional hierarchies where there was greater cross-classification in the data. This would enable greater study of the components of variation in the surgical process. This could be extended to incorporate a fuller investigation of the longitudinal aspects of audit data.

In conclusion we would hope that our work provides not only some useful insight into the statistical issues involved in the specific surgical audit problems we have addressed but also provides a basis for future research efforts in the same field.

Appendix 1

A brief description of the data analysed in Chapter 2

These data are fully discussed in McArdle & Hole (1991) and McArdle *et al* (1996). The first paper addresses aspects of the data collected in respect of colo-rectal cancer surgery at the Glasgow Royal Infirmary in the period 1974-79 and the second paper addresses these same data in addition to information collected on a further set of patients undergoing surgery in 1980-84. The figures below briefly summarize some of the main features of interest.

	Period	
	1974-79	1980-84
Patient numbers	645	483
Emergency admissions	32%	30%
Age <65	45%	36%
Sex (M:F)	50%	51%
Dukes's Stage A	4%	6%
B	32%	37%
C	31%	23%
D	33%	34%
Curative resections (as a % of emergency cases)	45%	39%
Curative resections (as a % of elective cases)	58%	60%
2 year survival rate (all cases)	47%	46%
2 year survival rate (curative resections)	72%	66%
5 year survival rate (all cases)	30%	25%
5 year survival rate (curative resections)	66%	39%

Note : A curative resection is one where the surgeon considers that there is no residual tumour after completion of the procedure.

Appendix 2

A brief description of the data analysed in Chapter 3

These data were discussed and analysed in Chapter 3. The form used to collect the data is reproduced in full in Appendix 3 and illustrates the range of basic items of data which were collected and the extensive list of outcome and follow up measures which needed to be recorded after surgery. The figures below briefly summarize some of the main features of interest. Some of the figures cannot be compared directly with those in the previous Appendix which concerned data collected over a longer period in one institution as opposed to a short period of follow up (2 years) in 8 institutions. In addition the data analyzed in Chapter 2 and summarized in Appendix 1 included very severe cases where only an investigative procedure was undertaken. These would not necessarily form part of the data as analyzed in Chapter 3.

Initial Patient numbers (before reduction)	2090
Emergency admissions	33%
Age <65	32%
Sex (M:F)	52%
Dukes's Stage A	3%
B	42%
C	26%
D	28%
2 year survival rate (elective cases)	74%
2 year survival rate (emergency cases)	57%
2 year survival rate (Dukes' Stage A)	92%
2 year survival rate (Dukes' Stage B)	82%
2 year survival rate (Dukes' Stage C)	68%
2 year survival rate (Dukes' Stage D)	43%

Appendix 3

A sample audit form referred to in Chapter 3

Colorectal Cancer Audit

Registration

Date of Registration: (2) ☐☐/☐☐/☐☐

(3) Hospital Code HospCode ☐☐

(1) Study No StudyNO ☐☐☐☐☐

Name: (5) (4) _____

Address: _____

_____ Post code: _____

Hospital No: (6) ☐☐☐☐☐☐☐☐☐☐ HospNumber

Sex: 1. Male 2. Female

Address Label

Date of Birth: __/__/__

Date of Admission: __/__/__

GP Name: _____

Address: _____

_____ Post code: _____

Consultant: _____

On completion of the above details, please return to:

Mrs H Wotherspoon
Research Nurse
University Department of Surgery
Royal Infirmary
10 Alexandra Parade
Glasgow G31 2ER

For Official Use Only

Refer to Hospital No. for Data Entry

(7) SEX ☐

(8) DOB ☐/☐/☐

(9) DATEADMISS ☐/☐/☐

(10) CODE-CONSULTANT ☐

Colorectal Cancer Audit

In-Patient Record

Name: _____

Hospital No: ☐☐☐☐☐☐☐☐☐☐

Date of Admission: ☐☐/☐☐/☐☐

Hospital Code ☐☐ ✓

Study No ☐☐☐☐ ✓

Date of Diagnosis ___/___/___

A **Presentation**

1. Elective
2. Emergency (Obstruction)
3. Emergency (Perforation)
4. Emergency (Other)

B **Family History of Colorectal Cancer**

1. No 2. Yes 3. Unknown

If yes, relation to patient _____

C **Site**

1. Caecum
2. Ascend
3. Hepatic
4. Transverse
5. Splenic
6. Descend
7. Sigmoid
8. Rectosigmoid above peritoneal reflection
9. Rectosigmoid straddling peritoneal reflection
10. Rectosigmoid below peritoneal reflection
11. Rectosigmoid - not specified
12. Rectum
13. Multiple invasive tumours

If 13, (specify sites) _____

D **Pre-Op Assessment**

Health Status 1. Fit 2. Unfit

Assessment of Tumour Spread

Liver USS	1. pos	2. susp	3. neg	4. not done
Liver CT	1. pos	2. susp	3. neg	4. not done
Chest X-ray	1. pos	2. susp	3. neg	4. not done
Chest CT	1. pos	2. susp	3. neg	4. not done

Any other evidence of tumour spread 1. No 2. Yes

If yes, (specify) _____

If yes, (method of detection) _____

[redacted] DateDiag (11)

[redacted] ✓ PRESENT (12)

[redacted] (13)

[redacted] (14)

[redacted] SITE ✓ (15)

[redacted] (16)

[redacted] (17)

[redacted] (18)

[redacted] (19)

[redacted] (20)

[redacted] (21)

[redacted] (22)

[redacted] (23)

[redacted] (24)

Colorectal Cancer Audit

In-Patient Record

D **Cont'd.**

CEA Level: _____ (if available)

For Official Use Only

25

E **Abdominal Surgery during initial admission**

ABDOSURG

1. **No** (go to section I) 2. **Yes** 26

1. Emergency (<24 hours post admission)
2. Delayed Emergency (> 24 hours post admission) 27
3. Elective

DATEOP_1

First Op Date of operation ___/___/___ 28

Type of operation _____ 29

Surgeon grade: _____ 30

Assistant grade: _____ 31

DATEOP_2

Second Op (if applicable) Date of operation ___/___/___ 32

Type of operation _____ 33

Surgeon grade: _____ 34

Assistant grade: _____ 35

Further operations during this admission? 1. **No** 2. **Yes** 36

F **Operative Findings**

Fixity 1. No 2. Yes FIXITY ✓ 37

Loco-regional nodes 1. No 2. Yes LRNODES ✓ 38

Liver metastasis 1. No 2. Yes LIVERMETS ✓ 39

If yes, method of detection _____ LIVER_DETEC 40

(eg visual /palpation, biopsy, intra op. us)

41 ✓ 42

Other distant spread 1. No 2. Yes (specify) _____ DIST_SPR

Was tumour resected? 1. No 2. Yes RESECTED ✓ 43

If tumour resected, was procedure 1. **Curative** 2. **Palliative** PROCEDURE ✓ 44

For palliative resection, evidence of
1. **Residual local disease**
2. **Distant spread**
3. **Both** EVIDENCE ✓ 45

Dukes D 1. No 2. Yes 46

Colorectal Cancer Audit
In-Patient Record

G Post-Op Complications

First Op:

Death 1. No 2. Yes If yes, date ____/____/____

Clinical Leak 1. No 2. Anast. Leak / fistula 3. Abscess 4. Both

Reoperation 1. No 2. Yes (should be noted in Section E)

If yes, reason _____

Second Op:
(if applicable)

Death 1. No 2. Yes If yes, date ____/____/____

Clinical Leak 1. No 2. Anast. Leak / fistula 3. Abscess 4. Both

Reoperation 1. No 2. Yes (should be noted in Section E)

If yes, reason _____

H Adjuvant Therapy

Pre-op Radiotherapy 1. No 2. Yes

Post-op Radiotherapy 1. No 2. Yes 3. Planned

Post-op Chemotherapy 1. No 2. Yes 3. Planned

Other 1. No 2. Yes 3. Planned

If yes, specify _____

Is the patient in Axis Study? 1. No 2. Yes

I No Abdominal Surgery

Was any treatment given? 1. No 2. Yes
(e.g. polypectomy, fulguration, laser)

If yes (specify) _____

Date commenced ____/____/____

OP1_DEATH (47)
DATE_DEATH1 (48)
OP1_LEAK (49)
OP1_REOP (50)
(51)
OP2_DEATH (52)
DATE_DEATH2 (53)
OP2_LEAK (54)
OP2_REOP (55)
(56)
(57)
(58)
(59)
(60)
(61)
(62)
(63)
(64)
(65)

Colorectal Cancer Audit

In-Patient Record

J **Date of Discharge** __/__/__

K **Date of Death** __/__/__ Not applicable ☐ ✓ (67)

(complete the following if known)

Was death colorectal cancer related? 1. No 2. Yes

If No, was colorectal cancer present at time of death? 1. No 2. Yes

Specify cause of death: _____

Consultant's signature: _____

Date: _____

DATEDISCH [redacted] (66)

DATEDIED [redacted] (68)

CRCDEATH [redacted] (69)

CRCPRESENT [redacted] (70)

[redacted] (71)

On completion of the above details, please return to:

Mrs H Wotherspoon
Research Nurse
University Department of Surgery
Royal Infirmary
10 Alexandra Parade
Glasgow G31 2ER

Colorectal Cancer Audit

Pathology

Name: _____

Hospital No: ☐☐☐☐☐☐☐☐☐☐

Date of Operation: ☐☐/☐☐/☐☐

Hospital Code ☐☐ ✓

Study No ☐☐☐☐☐ ✓

---

(72) + (80)
Pathology No: _____

For Official Use Only

- (73) + (81)

● **Tumour type:** 1. Adenocarcinoma 2. Other

- (74) + (82)

Maximum size of tumour: ☐ mm

- (75) + (83)

Tumour at resection margins: 1. No 2. Yes

DUKES CLASS +B

Dukes classification: 1. A 2. B 3. C 4. D

- (76) + (84)

DIFFERENT +B

Differentiation: 1. Well 2. Mod 3. Poor

- (77) + (85)

● **Polyps in resected specimen:** 1. No 2. Yes

- (78) + (86)

For AP resections, distance from tumour to anus: ☐ mm

- (79) + (87)

Pathologists signature: _____

Date: _____

On completion of the above details, please return to:

Mrs H Wotherspoon
Research Nurse
University Department of Surgery
Royal Infirmary
10 Alexandra Parade
Glasgow G31 2ER

Colorectal Cancer Audit

Follow-Up

Name: _____

Hospital Code ☐☐ ✓

Study No ☐☐☐☐☐ ✓

Hospital No: ☐☐☐☐☐☐☐☐☐☐

Date of Birth: ☐☐/☐☐/☐☐ Date of Registration: ☐☐/☐☐/☐☐

For Official Use Only

A **Is the patient alive?** 1. No 2. Yes ☐ (88)

If yes, date last seen ___/___/___ ☐☐/☐☐/☐☐ (89)

B **Has the patient undergone any abdominal surgery since last visit?** 1. No 2. Yes ☐ (90)

Date of operation ___/___/___ ☐☐/☐☐/☐☐ (91)

Type of operation _____ ☐ (92)

Date of operation ___/___/___ ☐☐/☐☐/☐☐ (93)

Type of operation _____ ☐ (94)

Post-Operative Death? 1. No 2. Yes ☐ (95)

C **Recurrence**

Local recurrence 1. No 2. Yes 3. Suspected ☐ (96)

Date ___/___/___ ☐☐/☐☐/☐☐ (97)

Distant recurrence 1. No 2. Yes 3. Suspected ☐ (98)

Site _____ ☐ (99)

Date ___/___/___ ☐☐/☐☐/☐☐ (100)

Colorectal Cancer Audit

Follow-Up

(continued)

For Official Use Only

FUD_DATEDEATH (102)

D **Date of Death** __/__/__ Not applicable ☑ (101)

(complete the following if known)

Was death colorectal cancer related? 1. No 2. Yes

FUD_CCREL ✓ (103)

If No, was colorectal cancer present at time of death? 1. No 2. Yes

FUD_CCPRES ✓ (104)

Specify cause of death: _____

(105)

Consultant's signature: _____

Date: _____

On completion of the above details, please return to:

Mrs H Wotherspoon
Research Nurse
University Department of Surgery
Royal Infirmary
10 Alexandra Parade
Glasgow G31 2ER

Appendix 4

A sample UKCSR form referred to in Chapter 5

CONGENITAL – 1

OPERATIONS FOR CONGENITAL HEART DISEASE

.B. Each operation must be entered once only in the category of what is considered to be the major procedure.

1980

	Over 1 Year				Under 1 Year			
	Closed		Open		Closed			
	No.	D.	No.	D.	No.	D.	No.	D
EXTRACARDIAC LESIONS								
Persistent ductus arteriosus	5	0						
A-P window								
Coarctation of aorta – simple	3	0						
– complicated (plus other major defect)								
Interrupted Aortic Arch								
Vascular Ring								
CONGENITAL VALVE SURGERY								
Aortic (exclude adult calcific disease)								
– valvotomy								
– subvalvar stenosis								
– supravalvar stenosis								
– replacement (homograft								
(xenograft								
(prosthesis								
Pulmonary – valvotomy and/or infundibulectomy. (With or without ASD).			2	0				
– replacement (homograft								
(xenograft								
(prosthesis								
Mitral – valvotomy/repair								
– replacement (prosthesis								
(xenograft								
(other								
Other (Please specify)								
Supra-aortic stenosis + pulm. valvotomy			1	0				
Inspection aortic valve			1	0				
TOTAL CONGENITAL PAGE 1	8	0	4	0				

CONGENITAL

	Over 1 Year				Under 1 Year			
	Closed		Open		Closed			
	No.	D.	No.	D.	No.	D.	No.	
Atrial septal defect (Secundum or Sinus Venosus)			19	0				
Partial A-V Canal (Primum ASD)								
Common (single) atrium								
Complete A-V Canal								
(palliative procedure								
(corrective procedure								
Ventricular septal defect (with or without ASD)								
(palliative procedure								
(corrective procedure								
Double outlet RV								
(palliative procedure								
(corrective procedure								
Truncus arteriosus								
(palliative procedure								
(corrective procedure								
GA with intact ventricular septum								
(palliative procedure								
(corrective procedure								
TGA with VSD								
(palliative procedure								
(corrective procedure								
TGA with SD and PAS								
(palliative procedure								
(corrective procedure								
TGA with other significant anomaly								
(palliative procedure								
(corrective procedure								
Corrected TGA and VSD								
(palliative procedure								
(corrective procedure			1	1				
Corrected TGA and other significant anomaly								
(palliative procedure								
(corrective procedure								
Single ventricle								
(palliative procedure	1	0						
(septation procedure								
(Fontan-type procedure								
CONGENITAL PAGE 2	1	0	20	1				

CONGENITAL - 3 (1)

	Over 1 Year				Under 1 Year			
	Closed		Open		Closed		Open	
	No.	D.	No.	D.	No.	D.	No.	D.
Tetralogy of Fallot								
(palliative procedure								
(corrective procedure			4	0				
VSD and PS								
(palliative procedure								
(corrective procedure			2	0				
Pulmonary atresia with VSD								
(palliative procedure								
(corrective procedure								
Pulmonary atresia with intact septum								
(palliative procedure								
(corrective procedure								
Tricuspid atresia								
(palliative procedure								
(Fontan type procedure								
Ebstein's anomaly								
(palliative procedure								
(corrective procedure								
MISCELLANEOUS								
Total anomalous pulmonary venous return.								
Cor triatriatum								
Anomalous origin coronary artery								
Coronary artery fistula								
Exploration only								
Surgical procedures for other conditions (Please specify)								
VSD + AVR			1	0				
VSD + TVR			1	1				
Closure sinus Valsalva			1	0				
TOTAL CONGENITAL PAGE 3			9	1				
TOTAL CONGENITAL PAGE 2	1	0	20 ✓	1 ✓				
TOTAL CONGENITAL PAGE 1	8	0	4	0				
TOTAL CONGENITAL OPERATIONS	9	0	33	2				

33 2

SURGERY FOR CONGENITAL HEART DISEASE — CONGENTIAL - 1

N.B. Each operation must be entered once only in the category of what is considered to be the major procedure.

	Over 1 Year				Under 1 Year			
	Closed		Open		Closed		Open	
	No.	D.	No.	D.	No.	D.	No.	D.
EXTRACARDIAC LESIONS								
Persistent ductus arteriosus								
A-P window								
Coarctation of aorta - simple								
- complicated (plus other major defect)								
Interrupted Aortic Arch								
Vascular Ring								
CONGENITAL VALVE SURGERY								
Aortic (exclude adult calcific disease)								
- stenosis - valvotomy								
- regurgitation - repair								
- subvalvar stenosis - repair								
- supravalvar stenosis - repair								
- replacement (homograft)								
(xenograft)								
(prosthesis)								
Pulmonary - stenosis - valvotomy								
- regurgitation - repair								
- RVOT obstruction - repair								
- replacement (homograft)								
(xenograft)								
(prosthesis)								
Mitral - stenosis - valvotomy								
- regurgitation - repair								
- replacement (homograft)								
(xenograft)								
(prosthesis)								
Other (please specify)								
TOTAL CONGENITAL PAGE 1								

SURGERY FOR CONGENITAL HEART DISEASE — CONGENITAL - 2

DEFECTS OF PARTITIONING	Over 1 Year				Under 1 Year			
	Closed		Open		Closed		Open	
	No.	D.	No.	D.	No.	D.	No.	D.
Atrial septal defect (Secundum or Sinus Venosus)			6					
Partial A-V Canal (Primum ASD)								
Common (single) atrium								
Complete A-V Canal								
(palliative procedure								
(corrective procedure								
Ventricular septal defect (with or without ASD)								
(palliative procedure								
(corrective procedure								
Double outlet RV								
(palliative procedure								
(corrective procedure								
Double outlet RV and PS								
(palliative procedure								
(corrective procedure								
Truncus arteriosus								
(palliative procedure								
(corrective procedure								
TGA with intact ventricular septum								
(palliative procedure								
(corrective procedure								
TGA with VSD								
(palliative procedure								
(corrective procedure								
TGA with VSD and PS								
(palliative procedure								
(corrective procedure								
TGA with other significant anomaly								
(palliative procedure								
(corrective procedure								
Corrected TGA and VSD								
(palliative procedure								
(corrective procedure								
Corrected TGA and other significant anomaly								
(palliative procedure								
(corrective procedure								
Single ventricle								
(palliative procedure								
(septation procedure								
(Fontan-type procedure								
TOTAL CONGENITAL PAGE 2			6					

SURGERY FOR CONGENITAL HEART DISEASE — CONGENTIAL - 3

	Over 1 Year				Under 1 Year			
	Closed		Open		Closed		Open	
	No.	D.	No.	D.	No.	D.	No.	D.
Tetralogy of Fallot								
(palliative procedure			1	1				
(corrective procedure								
VSD and PS (Valve or Sub-valve)								
(palliative procedure								
(corrective procedure								
Pulmonary atresia with VSD								
(palliative procedure								
(corrective procedure								
Pulmonary atresia with intact septum								
(palliative procedure								
(corrective procedure								
Tricuspid atresia								
(palliative procedure								
(corrective procedure								
Ebstein's anomaly								
(palliative procedure								
(corrective procedure								
MISCELLANEOUS								
Total anomalous pulmonary venous return								
Cor Triatriatum								
Anomalous origin coronary artery								
Coronary artery fistula								
Exploration only								
Surgical procedures for other conditions (please specify)								
Re-do VSD			1					
TOTAL CONGENITAL PAGE 3			2	1				
TOTAL CONGENITAL PAGE 2		0	6	0	0	0	0	0
TOTAL CONGENITAL PAGE 1					0	0	0	0
TOTAL CONGENITTAL OPERATIONS			8 9	1				

CONGENITAL SURGERY 2 DEFECTS OF PARTITIONING		UNDER 1 YEAR				1 to 15 years of age inclusive				16 YEARS of age or older			
		Closed		C.P.B.		Closed		C.P.B.		Closed		C.P.B.	
		No.	Died	No.	Died	No.	Died	No.	Died	No.	Died	No.	Died
ASD (Secundum or sinus venosus)													
Partial A-V canal (Primum ASD)													
Common (single) atrium													
Complete A-V canal	Palliative												
	Corrective												
VSD ± ASD	Palliative												
	Corrective												
Double Outlet RV	Palliative												
	Corrective												
DORV + PS	Palliative												
	Corrective												
Truncus arteriosus	Palliative												
	Corrective												
TGA with intact septum	Palliative												
	Corrective												
TGA + VSD	Palliative												
	Corrective												
TGA + VSD + PS	Palliative												
	Corrective												
TGA + other anomaly	Palliative												
	Corrective												
Corrected TGA + VSD	Palliative												
	Corrective												
Corrected TGA + other anomaly	Palliative												
	Corrective												
Single ventricle	Palliative												
	Septation procedure												
	Fontan type procedure												
TOTAL (For partitioning defects)													

CONGENITAL SURGERY 3		UNDER 1 YEAR				1 to 15 years of age inclusive				16 YEARS of age or older			
		Closed		C.P.B.		Closed		C.P.B.		Closed		C.P.B.	
		No.	Died	No.	Died	No.	Died	No.	Died	No.	Died	No.	Died
Fallot's Tetralogy	Palliative												
	Corrective												
VSD ± PS (valve or sub valve)	Palliative												
	Corrective												
Pulmonary atresia with VSD	Palliative												
	Corrective												
Tricuspid atresia	Palliative												
	Corrective												
Truncus arteriosus	Palliative												
	Corrective												
Ebstein's anomaly	Palliative												
	Corrective												
MISCELLANEOUS PROCEDURES FOR CONGENITAL HEART DISEASE													
Total anomalous pulmonary venous drainage													
Cor triatrium													
Anomalous origin of coronary artery													
Coronary artery fistula													
Exploration only													
Other (please specify)													
Total Congenital surgery 3													
Total Congenital surgery 2													
Total Congenital surgery 1													
Total Congenital Operations												1	

References

Aitken *et al,* Training in large bowel cancer surgery : observations from three prospective regional United Kingdom audits, *British Medical Journal,* 1999, **318,** 702-703

Altman DG, *Practical Statistics for Medical Research*, Chapman & Hall, London 1991

Andersen PK, Klein JP, Zhang MJ, Testing for centre effects in multi centre survival studies: a monte carlo comparison of fixed and random effects tests, , *Statistics in Medicine,* 1999, **18,** 1489-1500

Aylin *et al*, Analysis of Hospital Episode Statistics for the Bristol Royal Infirmary Inquiry, www.bristol-inquiry.org.uk/Documents/Health Episode Statistics (Aylin).pdf, 1999

Aylin *et al,* Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1984-1996 : was Bristol an outlier ?, *The Lancet,* 2001, **358**, 181-187

Best, N G, Spiegelhalter, D J, Thomas, A and Brayne, C E G, Bayesian analysis of realistically complex models. *Journal Royal Statistical Society A*, 1999,**159**, 323-342.

Breslow, N E and Day, N E *Statistical Methods on Cancer Research Volume 1: Case-Control Studies*. 1980, International Agency for Cancer Research, Lyon.

Brooks SP, Markov Chain Monte Carlo and its application, *Statistician,* 1998, **47**, 69-100

Bryk AS, Raudenbush SW, *Hierarchical Linear Models,* Sage, London, 1992

Buck N, Devlin HB, Lunn JN, The Report of a Confidential Enquiry into Perioperative Deaths. *Nuffield Provincial Hospitals Trust and the King's Fund*, 1987, London

Cale ARJ, King PM, Macleod DAD, Practical surgical audit: a morbidity profile. *Journal Royal College Surgeons Edinburgh*; 1991, **36**:41-4

Carter DC *et al,* Lothian and Borders large bowel cancer project : immediate outcome after surgery, *British Journal of Surgery,* 1995, **82,** 888-890

Carlin BP, Louis TA, *Bayes and Empirical Bayes Methods for Data Analysis,* Chapman & Hall, London, 1996

Chapuis *et al,* A Multivariate analysis of clinical and pathological variables in prognosis after resection of large bowel cancer, *British Journal of Surgery,* 1985, **72,** 698-702

Clarke DN, Jones PF, Needham CD, Outcome in colorectal carcinoma: seven year study of a population, *British Medical Journal,* 1980; 431-435

Clasby *et al,* Variable management of soft tissue carcinoma : regional audit with implications for specialist care, *British Journal of Surgery,* 1997, **84,** 1692-1696

Clinical Indicators Support Team, *History of the SMR,* www.show.scot.nhs.uk/indicators/Publications/SMR.htm, 2001

Clinical Resource and Audit Group, *Clinical Indicators December 2000,* www.show.scot.nhs.uk/crag/main.htm, 2000

Clinical Standards Board for Scotland, www.show.scot.nhs.uk/crag/topics/cssbs/csbmain.htm, 2001

Consultant Surgeons and Pathologists of the Lothian and Borders Health Boards, Lothian and Borders large bowel cancer project: immediate outcome after surgery. *British Journal of Surgery*;1995, **82**:888-90

Cook *et al,* An audit of hospital mortality after urgent and emergency surgery in the elderly, *Annals Royal College Surgeons England,* 1997, **79**, 361-367

Copeland GP, Jones D, Harris PL, Wilcox A, Comparative Vascular Audit using the POSSUM scoring system. *Annals Royal College Surgeons England*;1993, **75**:175-7 (a)

Copeland GP, Jones D, Walters M, POSSUM: a scoring system for surgical audit. *British Journal of Surgery*;1991, **78**:356-60 (b)

Copeland GP, Sagar P, Brennan J, Roberts G, Ward J, Cornford P, Millar A, Harris C, Risk adjusted analysis of surgeon performance: a 1-year study. *British Journal of Surgery*;1995, **82**:408-411

Cox DR, Oakes D, *Analysis of Survival Data,* Chapman & Hall, London, 1984

Cox DR, Regression models and life tables, *Journal of the Royal Statistical Society (B)*, 1972, **34**, 187-220

Crombie IK, Davies HTO, Missing link in the audit cycle. *Quality in Health Care*;1993, **2**:47-8

Crombie IK, Davies HTO, Abraham, SCS, du V Florey C. *The Audit Handbook.* Wiley, England, 1993.

Deans GT *et al,* Colorectal carcinoma: importance of clinical and pathological factors in survival. *Annals Royal College Surgeons England*; 1994 **76**:59-64

Deans GT, Odling-Smee W, McKelvey STD, Parks GT, Roy DA, Auditing perioperative mortality. *Annals Royal College Surgeons England*;1987, **69**:183-7

Deely JJ, Smith AFM, Quantitative refinements for comparisons of institutional performance, *Journal of the Royal Statistical Society (A),* 1998, **161**, 5-12

Delamothe T (ed), *Outcomes into Clinical Practice,* British Medical Journal Publishing , London 1994

Department of Health, *Hospital In Patients Data Based on Hospital Episode Statistic (HES)*, www.doh.gov.uk/hes/index.htm, 2001 (a)

Department of Health. *Quality and Performance in the NHS : Performance Indicators,* www.doh.gov.uk/nhsperformance indicators/index.htm, 2001 (b)

Dudley HAF, Necessity for Surgical Audit. *British Medical Journal,* 1974, (i):275-7

Dukes CE, The classification of cancer of the rectum, *Journal of Pathology,* 1932, **35**, 323-332

Dukes CE, Cancer of the rectum: an analysis of 1000 cases, *Journal of Pathology,* 1940, **50**, 527-539

Dunn DC, Audit of a surgical firm by microcomputer: 5 years experience. *British Medical Journal*; 1988, **296**:687-91

Dunn DC, Dale RF, Combined computer generated discharge documents and surgical audit. *British Medical Journal*; 1986,**292**:816-8

Dunn DC, Dale RF, Gumpert JRW, Duffy TJ, Combined surgical audit by microcomputer involving units in four health regions. *Annals Royal College Surgeons England,*1992;**74**:47-53

Dunn DC, Fowler S, Comparative audit: an experimental study of 147,882 general surgical admissions during 1990, *British Journal of Surgery,* 1992,**79**:1073-6

Ellis BW, How to set up an audit. *British Medical Journal,* 1989;**298**:1635-7

Ellis BW, Michie HR, Esufali ST, Pyper RJD, Dudley HAF. Development of a microcomputer-based system for surgical audit and patient administration: a review. *J Roy Soc Med*; 1987,80:157-61

Emberton M, Rivett R, Ellis BW,. Comparative Audit: A new method of delivering audit. *Annals Royal College Surgeons England*; 1991,**73**:117-20

Fielding LP, Stewart-Brown S, Blesovsky L, Kearny G,. Anastomotic integrity after operations for large bowel cancer: a multicentre study. *British Medical Journal* ; 1980,**288**:411-4

Gamerman D, *Markov Chain Monte Carlo,* , Chapman & Hall, London, 1997

Gelman A, Carlin JB, Stern HS, Rubin DB, *Bayesian Data Analysis*, Chapman & Hall, London, 1995

Geweke J, Gowrisankaran G and Town RJ, Inferring Hospital Quality from Patient Discharge Records uUsing an Bayesian Selection Mode*l, Working paper* available from www.biz.uiowa.edu/faculty/jgeweke/

Gilks WR, Richardson S, Speigelhalter DJ (eds), *Markov Chain Monte Carlo in Practice,* Chapman & Hall, London, 1996

Gilks, W R, Clayton, D G, Spiegelhalter, D J, Best, N G, McNeil, A J, Sharples, L D and Kirby, A J (1993) Modelling complexity: applications of Gibbs sampling in medicine. *J Roy Statist Soc B*, **55**, 39-52.

Gilks, W R, Thomas, A and Siegelhalter, D J A language and program for complex Bayesian modelling. *The Statistician*, 1994, **43**, 169-78.

Gill *et al*, 1978, The survival of patients with colorectal cancer treated in a single hospital, *British Journal of Surgery,* 1978, **65**,17-20

Gilmore OJA, Griffiths NJ, Connoly JC, Dunlop AW, Hart S, Thomson JPS, Todd IP, Surgical audit: comparison of the workload and result of two hospitals in the same district. *British Medical Journal*; 1980,**281**:1050-2

Goldstein H, *Multilevel Statistical Models,* Arnold, London, 1995

Goldstein H, Spiegelhater D, League tables and their limitations: statistical issues in comparisons of institutional performance, *Journal of the Royal Statistical Society (A),* 1996, **159**, 385-443

Gough MH, Kettlewell MGW, Marks CG, Holmes SJK, Holderness J,. Audit: an annual assessment of the work and performance of a surgical firm in a regional teaching hospital. *British Medical Journal*; 1980,**281**:913-918

Gruer R, Gordon DS, Gunn AA, Ruckley CV, Audit of surgical audit. *Lancet,* 1986 (i):23-6

Gowrisankaran G and Town RJ, Estimating the Quality of Care in Hospitals Using Instrumental Variables, *Journal of Health Economics*; 18, 747-67

Hartley MN, Sagar PM,. The surgeon's 'gut feeling' as a predictor of postoperative outcome. *Annals Royal College Surgeons England*; 1994,**76**:supp277-8

Hastings WK, Monte Carlo sampling methods using Markov chains and their applications, *Biometrica,* 1970, **57**, 97-109

Hayes CE, Statistical aspects of surgical audit, *PhD Thesis,* Glasgow University, 1995.

Holm *et al,* Influence of hospital and surgeon related factors on outcome after treatment for rectal cancer with or without preoperative radiotherapy, *British Journal of Surgery,*1997, **84,** 657-663

Hosmer DW, Lemeshow S, *Applied Logistic Regression,* Wiley, New York, 1989

Houghton A, Variation in outcome of surgical procedures. *British Journal of Surgery*; 1994, **81**:653-60

Jarman *et al,* Explaining differences in English hospital death rates using routinely collected data, *British Medical Journal,*1999, **318,** 1515-1520

Jones DR, Copeland GP, Decossart L, Comparison of POSSUM with APACHE II for prediction of outcome from a surgical high dependency unit. *British Journal of Surgery*; 1992,**79**:1293-6

Jones HJS, de Cossart L, Risk Scoring in surgical patients, *British Journal of Surgery,*1998, **86,** 149-157

Knaus WA,. The science of prediction and its implications for clinicians today. *Theor Surg*; 1988,**3**:93-101

Knaus WA, Draper EA, Wagner DP, Zimmerman JE,. APACHE II: a severity of disease classification system. *Crit Care Med*; 1985,**13**:818-29

Knaus WA, Wagner DP, Draper EA, Zimmerman JE, et al,. The APACHE III Prognostic System. *Chest*; 1991,**100**:1619-36

Knaus WA, Zimmermann JE, Wagner DP, Draper EA, Lawrence DE,. APACHE – acute, physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med*; 1981,**9**:591-7

Kreft I, De Leeuw J, *Introducing Multilevel Modeling*, Sage, London, 1998

Lawrance *et al,* Use of cumulative mortality data in patients with acute myocardial infarction for early detection of variation in clinical practice : observational study, *British Medical Journal,*2001, **323,** 324-327

Lawrence AE & Murray GD, The UK Cardiac Surgical Register : assessment of data quality issues for the Bristol Royal Infirmary Inquiry, www.bristol-inquiry.org.uk/final_report/annex_b/images/Ukcsr_S_R.pdf, 2000

Leyland AH, Examining the relationship between length of stay and readmission rates for selected diagnoses in Scottish hospitals, *IMA Journal of Mathematics Applied in Medicine and Biology*, 1995, **12**, 175-184

Longford NT, *Random Coefficient Models,* Oxford University Press, Oxford, 1993

Mannion R, Goddard M, Impact of published clinical outcome data : case study in NHS hospital trusts, *British Medical Journal,*2001, **323,** 260-263

Marshall EC, Spiegelhalter DS, Reliability of league tables of *in vitro* fertilisation clinics : retrospective analysis of live birth rates, *British Medical Journal,*1998, **316,** 1701-1704

Matasuyama Y, Sakamoto J, Ohashi, Y, A Bayesian hierarchical survival model for the institutional effects in a multi centre cancer clinical trial, *Statistics in Medicine,*1998, **17,** 1893-1908

Mathsoft Inc. *S-PLUS User's Guide and Guide to Statistics,* Seattle, 1997

Matthews HR, Powell DJ, McConkey CC, Effect of surgical experience on the results of resection for oesophagael carcinoma, *British Journal of Surgery,*1986, **73,** 621-623

McArdle CS *et al* Prospective study of colorectalk cancer in the West of Scotland :10 year follow up, *British Journa of Surgery*; 1990,**77**:280-282

McArdle CS, Hole D, Impact of variability among surgeons on postoperative morbidity and mortality and ultimate survival. *British Medical Journal*; 1991,**302**:1501-5

McArdle CS, Wotherspoon H, Hole D, Murray GD, Colorectal Cancer: a continuing problem, *GI Cancer*, 1996; **1**, 171-176

McArdle *et al,* Prospective study of colorectal cancer in the West of Scotland : 10 year follow up, *British Journal of Surgery,*1990, **77,** 280-282

McCarthy MJ, Byrne, GJ, Surgical Audit : the junior doctors' viewpoint, *Journal of the Royal College of Surgeons Edinburgh,* 1997, **42**, 317-318

McClellan M, McNel B and Newhouse J, Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality, *Journal of the American Medical Association*; 272, 859-866

Mella *et al,* Population based audit of colorectal cancer management in two UK heath regions, *British Journal of Surgery,* 1997, **84,** 1731-1736

Metroplolis *et al* (1953) Equations of state calculations by fast computing machines, *Journal of Chemical Physics,* 1953, **21**, 1087-1091

Minitab Inc, *Minitab v 11.0 User's Guide,* New York, 1999

Morgan BJT, *Elements of Simulation*, Chapman & Hall, London 1984

Moses LE, Mosteller F, Institutional differences in postoperative death rates. *Journal American Medical Association*; 1968,**203**:150-2

Murray GD, Hayes C, Fowler S, Dunn DC, Presentation of Comparative Audit Data. *British Journal of Surgery*; 1995, **82**:329-32

Murray GD, Lawrence AE, Pollock J, A report on the UK Cardiac Surgical Register and the South West Congenital Heart Register for the Bristol Inquiry: a statistical analysis and review of key data sources relevant to the Inquiry's remit, , www.bristol-inquiry.org.uk/Documents/UKCSRSWCHR (Murray).pdf, 1999

Murray GD, Murray LS, Barlow P, Teasdale GM, Jennett WB, Assessing the performance and clinical impact of a computerised prognostic system in severe head injury. *Statistics in Medicine*; 1986,**5**:403-10

Nashef SAM *et al,* Risk Stratification for open heart surgery: trial of the Parsonnet system in a British hospital, *British Medical Journal,* 1992, **305,** 1066-1067

National Confidential Inquiry into Perioperative deaths (1999), National Confidential Inquiry into Perioperative deaths

National Institute of Clinical Excellence, www.nice.org.uk/nice-web/, 2001

Neugebauer *et al*, Conventional versus laparoscopic cholecystectomy and the randomised controlled trial, *British Journal of Surgery,* 1991, **78**, 150-154

NHS Executive. *The New NHS Performance Assessment Framework*, www.doh.gov.uk/nhsexec/nhspaf.htm,

Nixon, SJ, Does audit result in change of practice? The Lothian surgical experience *Quality in Health Care*; 1992, **1,** supp:S25-S27

O'Hagan A, *Bayesian Inference,* Arnold, London, 1994

Parry *et al,* Annual league tables of mortality in neonatal intensive care units: longitudinal study, *British Medical Journal,*1998, **316,** 1931-1935

Parsonnet V, Dean D, Bernstein AD, A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease, *Circulation,*1989, **79,** 3-11

Peterkin T, Top surgeon confesses to deadly blunders, *Scotland on Sunday*, 8 April 2001

Phillips *et al,* Local recurrence following 'curative' surgery for large bowel cancer: 1. The whole picture, *British Journal of Surgery,*1984, **71,** 12-16

Playforth *et al,* Pre-operative assessment of fitness score, *British Journal of Surgery,* 1987, **74,** 890-892

Pollock A, Evans M. *Surgical Audit.* Butterworths, London, 1989.

Pollock AV,. Surgical evaluation at the crossroads. *British Journal of Surgery*; 1993, **80**:974-6.

Poloniecki J, Valencia O, Littlejohns P, Cumulatve risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery, *British Medical Journal,*1998, **316,** 1697-1700

Porter *et al,* Surgeon related factors and outcome in rectal cancer, *Annals of Surgery,* 1998, **227**, 157-167

Prytherch DR *et al,* POSSUM and Portsmoth POSSUM for predicting mortality, *British Journal of Surgery,*1998, **85,** 1217-1220

Rice N, Leyland A, Multilevel models : applications to health data, *Journal of Heath Service Research Policy,*1996, **1,** 154-164

Ruckley CV, Mechanisms of audit: discussion paper. *J Roy Soc Med*; 1984,**77**:40-44.

Sagar PM, Hartley MN, Mancey-Jones B, Sedman PC, May J, MacFie J,. Comparative Audit of Colorectal Resection with the POSSUM scoring system. *British Journal of Surgery*; 1994, **81**:1492-4.

Sanderson C, McKee M, Commentary: How robust are rankings ? The implications of confidence intervals, *British Medical Journal,* 1998, **316,** 1705

Schein M,. Acute surgical disease and scoring systems in daily surgical practice. *British Journal of Surgery*; 1988,**75**:731-2.

Secretary of State for Health, *The New NHS*, Stationery Office 1997 (cm3807)

Singh KK *et al,* Audit of colorectal cancer surgery by non specialist surgeons, *British Journal of Surgery,* 1997, **84,** 343-347

Smith R (ed), *Audit in Action,* British Medical Journal Publishing , London 1992

Speigelhalter D, An initial synthesis of statistical sources concerning the nature and outcomes of paediatric cardiac surgical services at Bristol relative to other specialist centres from 1984 to 1995 , www.bristol-inquiry.org.uk/Documents/synthesis 2.pdf, 1999

Spiegelhater D, Evans SJW, Aylin P, Murray G, Overview of statistical evidence presented to the Bristol Royal Infirmary Inquiry concerning the nature and outcomes of paediatric cardiac surgical servces at Bristol relative to other specialist centres from 1984 to 1995 , www.bristol-inquiry.org.uk/Documents/statistical overview report.pdf, 1999 (a)

Spiegelhater D, Thomas A, Best N, *WinBUGS User Manual Version 1.2,* MRC Biostatistics Unit, 1999 (b)

Spiegelhater D, Thomas A, Best N, Gilks W, *BUGS 0.5 Bayesian Inference Using Gibbs Sampling Manual,* MRC Biostatistics Unit, 1995

Spiegelhater D, Thomas A, Best N, Gilks W, *BUGS 0.5 Examples Volume 1,* MRC Biostatistics Unit, 1996 (a)

Spiegelhater D, Thomas A, Best N, Gilks W, *BUGS 0.5 Examples Volume 1l,* MRC Biostatistics Unit, 1996 (b)

Spiegelhater D, Surgical Audit: statistical lessons from Nightingale and Codman, *Journal of the Royal Statistical Society (A),* 1999, **162**, 45-58

Spence RAJ. Colorectal carcinoma: importance of clinical and pathological factors in survival. A*nnals Royal College Surgeons England* ; 1994,**76**:59-64

SPSS Inc, *SPSS v 10 Base and Advanced Statistics,* Chicago, 1998

Steiner SH *et al,* Monitoring surgical performance using risk adjusted cumulative sum charts, *Biostatistics,*2000, **1,** 441-452

The Bristol Royal Infirmary Inquiry, *Learning from Bristol: the report of the public Inquiry into children's heart surgery at the Bristol Royal Infirmary 1984 to 1995*, (cm5207), 2001 (a)

The Bristol Royal Infirmary Inquiry, www.bristol-inquiry.org.uk, (b)

The Consultant Surgeons and Pathologists of the Lothian and Borders Health Boards, Lothian and Borders large bowel cancer project: immediate outcome after surgery, *British Journal of Surgery,* 1995, **82,** 888-890

The Shipman Inquiry, www.the–shipman-inquiry.org.uk.

Tremblay *et al,* A simple classification of the risk in cardiac surgery : the first decade, *Canadian Journal Anesthesia* , 1993, **40,** 103-111

Wijesinghte LD *et al,* Comparison of POSSUM and the Portsmouth predictor equation for predicting death following vascular surgery, *British Journal of Surgery,* 1998, **85,** 209-212

Williams SM, Parry BR, Schlup MMT, Quality control: an application of the cusum, *British Medical Journal,* 1992, **304,** 1359-1361